An epsilon of room: pages from year three of a mathematical blog

Terence Tao

Department of Mathematics, UCLA, Los Angeles, CA 90095

E-mail address: tao@math.ucla.edu

To Garth Gaudry, who set me on the road;
To my family, for their constant support;
And to the readers of my blog, for their feedback and contributions.

Contents

Preface

In February of 2007, I converted my "What's new" web page of research updates into a blog at `terrytao.wordpress.com`. This blog has since grown and evolved to cover a wide variety of mathematical topics, ranging from my own research updates, to lectures and guest posts by other mathematicians, to open problems, to class lecture notes, to expository articles at both basic and advanced levels.

With the encouragement of my blog readers, and also of the AMS, I published many of the mathematical articles from the first two years of the blog as [**Ta2008**] and [**Ta2009**], which will henceforth be referred to as *Structure and Randomness* and *Poincaré's Legacies Vols. I, II* throughout this book. This gave me the opportunity to improve and update these articles to a publishable (and citeable) standard, and also to record some of the substantive feedback I had received on these articles by the readers of the blog.

The current text contain many (though not all) of the posts for the third year (2009) of the blog, focusing primarily on those posts of a mathematical nature which were not contributed primarily by other authors, and which are not published elsewhere.

This year, over half of the material consists of lecture notes from my graduate real analysis courses that I taught at UCLA (Chapter 1), together with some related material in Chapter 2. These notes cover the second part of the graduate real analysis sequence here,

and therefore assume some familiarity with general measure theory (in particular, the construction of Lebesgue measure and the Lebesgue integral, and more generally the material reviewed in Section 1.1), as well as undergraduate real analysis (e.g. various notions of limits and convergence). The notes then cover more advanced topics in measure theory (notably, the Lebesgue-Radon-Nikodym and Riesz representation theorems), as well as a number of topics in functional analysis, such as the theory of Hilbert and Banach spaces, and the study of key function spaces such as the Lebesgue and Sobolev spaces, or spaces of distributions. The general theory of the Fourier transform is also discussed. In addition, a number of auxiliary (but optional) topics, such as Zorn's lemma, are discussed in Chapter 2. In my own course, I covered the material in Chapter 1 only, and also used Folland's text [**Fo2000**] as a secondary source; but I hope that this text may be useful in other graduate real analysis courses, particularly in conjunction with a secondary text (in particular, one that covers the prerequisite material on measure theory).

The rest of this text consists of sundry articles on a variety of mathematical topics, which I have divided (somewhat arbitrarily) into expository articles (Chapter 3) which are introductory articles on topics of relatively broad interest, and more technical articles (Chapter 4) which are narrower in scope, and often related to one of my current research interests. These can be read in any order, although they often reference each other, as well as articles from previous volumes in this series.

A remark on notation

For reasons of space, we will not be able to define every single mathematical term that we use in this book. If a term is italicised for reasons other than emphasis or for definition, then it denotes a standard mathematical object, result, or concept, which can be easily looked up in any number of references. (In the blog version of the book, many of these terms were linked to their Wikipedia pages, or other on-line reference pages.)

I will however mention a few notational conventions that I will use throughout. The cardinality of a finite set E will be denoted

$|E|$. We will use the asymptotic notation $X = O(Y)$, $X \ll Y$, or $Y \gg X$ to denote the estimate $|X| \leq CY$ for some absolute constant $C > 0$. In some cases we will need this constant C to depend on a parameter (e.g. d), in which case we shall indicate this dependence by subscripts, e.g. $X = O_d(Y)$ or $X \ll_d Y$. We also sometimes use $X \sim Y$ as a synonym for $X \ll Y \ll X$.

In many situations there will be a large parameter n that goes off to infinity. When that occurs, we also use the notation $o_{n\to\infty}(X)$ or simply $o(X)$ to denote any quantity bounded in magnitude by $c(n)X$, where $c(n)$ is a function depending only on n that goes to zero as n goes to infinity. If we need $c(n)$ to depend on another parameter, e.g. d, we indicate this by further subscripts, e.g. $o_{n\to\infty;d}(X)$.

We will occasionally use the averaging notation $\mathbf{E}_{x\in X} f(x) := \frac{1}{|X|}\sum_{x\in X} f(x)$ to denote the average value of a function $f : X \to \mathbf{C}$ on a non-empty finite set X.

Acknowledgments

The author is supported by a grant from the MacArthur Foundation, by NSF grant DMS-0649473, and by the NSF Waterman award.

Thanks to Blake Stacey and anonymous commenters for global corrections to the text.

Chapter 1

Real analysis

1.1. A quick review of measure and integration theory

In this section we quickly review the basics of abstract measure theory and integration theory, which was covered in the previous course but will of course be relied upon in the current course. This is only a brief summary of the material; of course, one should consult a real analysis text for the full details of the theory.

1.1.1. Measurable spaces. Ideally, measure theory on a space X should be able to assign a measure (or "volume", or "mass", etc.) to every set in X. Unfortunately, due to paradoxes such as the *Banach-Tarski paradox*, many natural notions of measure (e.g. *Lebesgue measure*) cannot be applied to measure all subsets of X; instead, one must restrict attention to certain measurable subsets of X. This turns out to suffice for most applications; for instance, just about any "non-pathological" subset of Euclidean space that one actually encounters will be Lebesgue measurable (as a general rule of thumb, any set which does not rely on the axiom of choice in its construction will be measurable).

To formalise this abstractly, we use

Definition 1.1.1 (Measurable spaces)**.** A *measurable space* $(X, \mathcal{X})$ is a set X, together with a collection $\mathcal{X}$ of subsets of X which form a σ-algebra, thus $\mathcal{X}$ contains the empty set and X, and is closed under countable intersections, countable unions, and complements. A subset of X is said to be measurable with respect to the measurable space if it lies in $\mathcal{X}$.

A function $f : X \to Y$ from one measurable space $(X, \mathcal{X})$ to another $(Y, \mathcal{Y})$ is said to be *measurable* if $f^{-1}(E) \in \mathcal{X}$ for all $E \in \mathcal{Y}$.

Remark 1.1.2. The class of measurable spaces forms a *category*, with the measurable functions being the *morphisms*. The symbol σ stands for "countable union"; cf. σ-compact, σ-finite, F_σ set.

Remark 1.1.3. The notion of a measurable space $(X, \mathcal{X})$ (and of a measurable function) is superficially similar to that of a *topological space* $(X, \mathcal{F})$ (and of a *continuous function*); the topology $\mathcal{F}$ contains

$\emptyset$ and X just as the σ-algebra $\mathcal{X}$ does, but is now closed under arbitrary unions and finite intersections, rather than countable unions, countable intersections, and complements. The two categories are linked to each other by the Borel algebra construction, see Example 1.1.5 below.

Example 1.1.4. We say that one σ-algebra $\mathcal{X}$ on a set X is *coarser* than another $\mathcal{X}'$ (or that $\mathcal{X}'$ is finer than $\mathcal{X}$) if $\mathcal{X} \subset \mathcal{X}'$ (or equivalently, if the identity map from $(X, \mathcal{X}')$ to $(X, \mathcal{X})$ is measurable); thus every set which is measurable in the coarse space is also measurable in the fine space. The coarsest σ-algebra on a set X is the trivial σ-algebra $\{\emptyset,\mathrm{X}\}$, while the finest is the discrete σ-algebra $2^X := \{E : E \subset \mathrm{X}\}$.

Example 1.1.5. The intersection $\bigwedge_{\alpha\in A} \mathcal{X}_\alpha := \bigcap_{\alpha\in A} \mathcal{X}_\alpha$ of an arbitrary family $(\mathcal{X}_\alpha)_{\alpha\in A}$ of σ-algebras on X is another σ-algebra on X. Because of this, given any collection $\mathcal{F}$ of sets on X we can define the σ-algebra $\mathcal{B}[\mathcal{F}]$ *generated by* $\mathcal{F}$, defined to be the intersection of all the σ-algebras containing $\mathcal{F}$, or equivalently the coarsest algebra for which all sets in $\mathcal{F}$ are measurable. (This intersection is non-vacuous, since it will always involve the discrete σ-algebra 2^X.) In particular, the open sets $\mathcal{F}$ of a topological space $(X, \mathcal{F})$ generate a σ-algebra, known as the *Borel σ-algebra* of that space.

We can also define the *join* $\bigvee_{\alpha\in A} \mathcal{X}_\alpha$ of any family $(\mathcal{X}_\alpha)_{\alpha\in A}$ of σ-algebras on X by the formula

$$\bigvee_{\alpha\in A} \mathcal{X}_\alpha := \mathcal{B}[\bigcup_{\alpha\in A} \mathcal{X}_\alpha]. \tag{1.1}$$

For instance, the *Lebesgue σ-algebra* $\mathcal{L}$ of Lebesgue measurable sets on a Euclidean space $\mathbf{R}^n$ is the join of the Borel σ-algebra $\mathcal{B}$ and of the algebra of null sets and their complements (also called co-null sets).

Exercise 1.1.1. A function $f : X \to Y$ from one topological space to another is said to be *Borel measurable* if it is measurable once X and Y are equipped with their respective Borel σ-algebras. Show that every continuous function is Borel measurable. (The converse statement, of course, is very far from being true; for instance, the pointwise limit of a sequence of measurable functions, if it exists, is also measurable,

whereas the analogous claim for continuous functions is completely false.)

Remark 1.1.6. A function $f : \mathbf{R}^n \to \mathbf{C}$ is said to be *Lebesgue measurable* if it is measurable from $\mathbf{R}^n$ (with the Lebesgue σ-algebra) to $\mathbf{C}$ (with the Borel σ-algebra), or equivalently if $f^{-1}(B)$ is Lebesgue measurable for every open ball B in $\mathbf{C}$. Note the asymmetry between Lebesgue and Borel here; in particular, the composition of two Lebesgue measurable functions need not be Lebesgue measurable.

Example 1.1.7. Given a function $f : X \to Y$ from a set X to a measurable space $(Y, \mathcal{Y})$, we can define the *pullback* $f^{-1}(\mathcal{Y})$ of $\mathcal{Y}$ to be the σ-algebra $f^{-1}(\mathcal{Y}) := \{f^{-1}(E) : E \in \mathcal{Y}\}$; this is the coarsest structure on X that makes f measurable. For instance, the pullback of the Borel σ-algebra from $[0,1]$ to $[0,1]^2$ under the map $(x,y) \mapsto x$ consists of all sets of the form $E \times [0,1]$, where $E \subset [0,1]$ is Borel-measurable.

More generally, given a family $(f_\alpha : X \to Y_\alpha)_{\alpha \in A}$ of functions into measurable spaces $(Y_\alpha, \mathcal{Y}_\alpha)$, we can form the σ-algebra $\bigvee_{\alpha \in A} f_\alpha^{-1}(\mathcal{Y}_\alpha)$ generated by the f_α; this is the coarsest structure on X that makes all the f_α simultaneously measurable.

Remark 1.1.8. In probability theory and information theory, the functions $f_\alpha : X \to Y_\alpha$ in Example 1.1.7 can be interpreted as *observables*, and the σ-algebra generated by these observables thus captures mathematically the concept of observable information. For instance, given a time parameter t, one might define the σ-algebra $\mathcal{F}_{\leq t}$ generated by all observables for some random process (e.g. *Brownian motion*) that can be made at time t or earlier; this endows the underlying event space X with an uncountable increasing family of σ-algebras.

Example 1.1.9. If E is a subset of a measurable space $(Y, \mathcal{Y})$, the pullback of $\mathcal{Y}$ under the inclusion map $\iota : E \to Y$ is called the *restriction* of $\mathcal{Y}$ to E and is denoted $\mathcal{Y} \restriction_E$. Thus, for instance, we can restrict the Borel and Lebesgue σ-algebras on a Euclidean space $\mathbf{R}^n$ to any subset of such a space.

Exercise 1.1.2. Let M be an n-dimensional manifold, and let $(\pi_\alpha : U_\alpha \to V_\alpha)$ be an atlas of coordinate charts for M, where U_α is an

open cover of M and V_α are open subsets of $\mathbf{R}^n$. Show that the Borel σ-algebra on M is the unique σ-algebra whose restriction to each U_α is the pullback via π_α of the restriction of the Borel σ-algebra of $\mathbf{R}^n$ to V_α.

Example 1.1.10. A function $f : X \to A$ into some index set A will partition X into level sets $f^{-1}(\{\alpha\})$ for $\alpha \in A$; conversely, every partition $X = \bigcup_{\alpha\in A} E_\alpha$ of X arises from at least one function f in this manner (one can just take f to be the map from points in X to the partition cell that that point lies in). Given such an f, we call the σ-algebra $f^{-1}(2^A)$ the σ-algebra *generated by* the partition; a set is measurable with respect to this structure if and only if it is the union of some sub-collection $\bigcup_{\alpha\in B} E_\alpha$ of cells of the partition.

Exercise 1.1.3. Show that a σ-algebra on a finite set X necessarily arises from a partition $X = \bigcup_{\alpha\in A} E_\alpha$ as in Example 1.1.10, and furthermore the partition is unique (up to relabeling). Thus in the finitary world, σ-algebras are essentially the same concept as partitions.

Example 1.1.11. Let $(X_\alpha, \mathcal{X}_\alpha)_{\alpha\in A}$ be a family of measurable spaces, then the Cartesian product $\prod_{\alpha\in A} X_\alpha$ has canonical projection maps $\pi_\beta : \prod_{\alpha\in A} X_\alpha \to X_\beta$ for each $\beta \in A$. The product σ-algebra $\prod_{\alpha\in A} \mathcal{X}_\alpha$ is defined as the σ-algebra on $\prod_{\alpha\in A} X_\alpha$ generated by the π_α as in Example 1.1.7.

Exercise 1.1.4. Let $(X_\alpha)_{\alpha\in A}$ be an at most countable family of second countable topological spaces. Show that the Borel σ-algebra of the product space (with the product topology) is equal to the product of the Borel σ-algebras of the factor spaces. In particular, the Borel σ-algebra on $\mathbf{R}^n$ is the product of n copies of the Borel σ-algebra on $\mathbf{R}$. (The claim can fail when the countability hypotheses are dropped, though in most applications in analysis, these hypotheses are satisfied.) We caution however that the Lebesgue σ-algebra on $\mathbf{R}^n$ is not the product of n copies of the one-dimensional Lebesgue σ-algebra, as it contains some additional null sets; however, it is the completion of that product.

Exercise 1.1.5. Let $(X, \mathcal{X})$ and $(Y, \mathcal{Y})$ be measurable spaces. Show that if E is measurable with respect to $\mathcal{X} \times \mathcal{Y}$, then for every $x \in X$,

the set $\{y \in Y : (x,y) \in E\}$ is measurable in $\mathcal{Y}$, and similarly for every $y \in Y$, the set $\{x \in X : (x,y) \in E\}$ is measurable in $\mathcal{X}$. Thus, sections of Borel-measurable sets are again Borel-measurable. (The same is not true for Lebesgue-measurable sets.)

1.1.2. Measure spaces. Now we endow measurable spaces with a measure, turning them into measure spaces.

Definition 1.1.12 (Measures)**.** A (non-negative) *measure* μ on a measurable space $(X, \mathcal{X})$ is a function $\mu : \mathcal{X} \to [0, +\infty]$ such that $\mu(\emptyset) = 0$, and such that we have the countable additivity property $\mu(\bigcup_{n=1}^{\infty} E_n) = \sum_{n=1}^{\infty} \mu(E_n)$ whenever $E_1, E_2, \ldots$ are disjoint measurable sets. We refer to the triplet $(X, \mathcal{X}, \mu)$ as a *measure space*.

A measure space $(X, \mathcal{X}, \mu)$ is *finite* if $\mu(X) < \infty$; it is a *probability space* if $\mu(X) = 1$ (and then we call μ a *probability measure*). It is *σ-finite* if X can be covered by countably many sets of finite measure.

A measurable set E is a *null set* if $\mu(E) = 0$. A property on points x in X is said to hold for *almost every* $x \in X$ (or *almost surely*, for probability spaces) if it holds outside of a null set. We abbreviate almost every and almost surely as a.e. and a.s. respectively. The complement of a null set is said to be a *co-null set* or to have *full measure*.

Example 1.1.13 (Dirac measures)**.** Given any measurable space $(X, \mathcal{X})$ and a point $x \in X$, we can define the *Dirac measure* (or *Dirac mass*) δ_x to be the measure such that $\delta_x(E) = 1$ when $x \in E$ and $\delta_x(E) = 0$ otherwise. This is a probability measure.

Example 1.1.14 (Counting measure)**.** Given any measurable space $(X, \mathcal{X})$, we define *counting measure* $\#$ by defining $\#(E)$ to be the cardinality $|E|$ of E when E is finite, or $+\infty$ otherwise. This measure is finite when X is finite, and σ-finite when X is at most countable. If X is also finite, we can define *normalised counting measure* $\frac{1}{|E|}\#$; this is a probability measure, also known as *uniform probability measure* on X (especially if we give X the discrete σ-algebra).

Example 1.1.15. Any finite non-negative linear combination of measures is again a measure; any finite convex combination of probability measures is again a probability measure.

Example 1.1.16. If $f : X \to Y$ is a measurable map from one measurable space $(X, \mathcal{X})$ to another $(Y, \mathcal{Y})$, and μ is a measure on $\mathcal{X}$, we can define the push-forward $f_*\mu : \mathcal{Y} \to [0, +\infty]$ by the formula $f_*\mu(E) := \mu(f^{-1}(E))$; this is a measure on $(Y, \mathcal{Y})$. Thus, for instance, $f_*\delta_x = \delta_{f(x)}$ for all $x \in X$.

We record some basic properties of measures of sets:

Exercise 1.1.6. Let $(X, \mathcal{X}, \mu)$ be a measure space. Show the following statements:

(i) (Monotonicity) If $E \subset F$ are measurable sets, then $\mu(E) \leq \mu(F)$. (In particular, any measurable subset of a null set is again a null set.)

(ii) (Countable subadditivity) If $E_1, E_2, \ldots$ are a countable sequence of measurable sets, then $\mu(\bigcup_{n=1}^\infty E_n) \leq \sum_{n=1}^\infty \mu(E_n)$. (Of course, one also has subadditivity for finite sequences.) In particular, any countable union of null sets is again a null set.

(iii) (Monotone convergence for sets) If $E_1 \subset E_2 \subset \ldots$ are measurable, then $\mu(\bigcup_{n=1}^\infty E_n) = \lim_{n\to\infty} \mu(E_n)$.

(iv) (Dominated convergence for sets) If $E_1 \supset E_2 \supset \ldots$ are measurable, and $\mu(E_1)$ is finite, then $\mu(\bigcap_{n=1}^\infty E_n) = \lim_{n\to\infty} \mu(E_n)$. Show that the claim can fail if $\mu(E_1)$ is infinite.

Exercise 1.1.7. A measure space is said to be *complete* if every subset of a null set is measurable (and is thus again a null set). Show that every measure space $(X, \mathcal{X}, \mu)$ has a unique minimal complete refinement $(X, \overline{\mathcal{X}}, \mu)$, known as the completion of $(X, \mathcal{X}, \mu)$, and that a set is measurable in $\overline{\mathcal{X}}$ if and only if it is equal almost everywhere to a measurable set in $\mathcal{X}$. (The completion of the Borel σ-algebra with respect to Lebesgue measure is known as the *Lebesgue σ-algebra*.)

A powerful way to construct measures on σ-algebras $\mathcal{X}$ is to first construct them on a smaller Boolean algebra $\mathcal{A}$ that generates $\mathcal{X}$, and then extend them via the following result:

Theorem 1.1.17 (Carathéodory's extension theorem, special case)**.** *Let $(X, \mathcal{X})$ be a measurable space, and let $\mathcal{A}$ be a Boolean algebra*

(i.e. closed under finite unions, intersections, and complements) that generates $\mathcal{X}$. Let $\mu : \mathcal{A} \to [0, +\infty]$ be a function such that

(i) $\mu(\emptyset) = 0$;

(ii) *If $A_1, A_2, \ldots \in \mathcal{A}$ are disjoint and $\bigcup_{n=1}^{\infty} A_n \in \mathcal{A}$, then $\mu(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mu(A_n)$.*

Then μ can be extended to a measure $\mu : \mathcal{X} \to [0, +\infty]$ on $\mathcal{X}$, which we shall also call μ.

Remark 1.1.18. The conditions (i), (ii) in the above theorem are clearly necessary if μ has any hope to be extended to a measure on $\mathcal{X}$. Thus this theorem gives a necessary and sufficient condition for a function on a Boolean algebra to be extended to a measure. The extension can easily be shown to be unique when X is σ-finite.

Proof. (Sketch) Define the outer measure $\mu_*(E)$ of any set $E \subset X$ as the infimum of $\sum_{n=1}^{\infty} \mu(A_n)$, where $(A_n)_{n=1}^{\infty}$ ranges over all coverings of E by elements in $\mathcal{A}$. It is not hard to see that if μ_* agrees with μ on $\mathcal{A}$, so it will suffice to show that it is a measure on $\mathcal{X}$.

It is easy to check that μ_* is monotone and countably subadditive (as in parts (i), (ii) of Exercise 1.1.6) on all of 2^X, and assigns zero to $\emptyset$; thus it is an outer measure in the abstract sense. But we need to show countable additivity on $\mathcal{X}$. The key is to first show the related property

$$\mu_*(A) = \mu_*(A \cap E) + \mu_*(A \backslash E) \tag{1.2}$$

for all $A \subset X$ and $E \in \mathcal{X}$. This can first be shown for $E \in \mathcal{A}$, and then one observes that the class of E that obey (1.2) for all A is a σ-algebra; we leave this as a (moderately lengthy) exercise.

The identity (1.2) already shows that μ_* is finitely additive on $\mathcal{X}$; combining this with countable subadditivity and monotonicity, we conclude that μ_* is countably additive, as required. □

Exercise 1.1.8. Let the notation and hypotheses be as in Theorem 1.1.17. Show that given any $\varepsilon > 0$ and any set $E \in \mathcal{X}$ of finite measure, there exists a set $F \in \mathcal{A}$ which differs from E by a set of measure at most ε. If X is σ-finite, show that the hypothesis that E has finite measure can be removed. (*Hint*: first reduce to the

case when X is finite, then show that the class of all E obeying this property is a σ-algebra.) Thus sets in the σ-algebra $\mathcal{X}$ "almost" lie in the algebra $\mathcal{A}$; this is an example of *Littlewood's first principle*. The same statements of course apply for the completion $\overline{\mathcal{X}}$ of $\mathcal{X}$.

One can use Theorem 1.1.17 to construct Lebesgue measure on $\mathbf{R}$ and on $\mathbf{R}^n$ (taking $\mathcal{A}$ to be, say, the algebra generated by half-open intervals or boxes), although the verification of hypothesis (ii) of Theorem 1.1.17 turns out to be somewhat delicate, even in the one-dimensional case. But one can at least get the higher-dimensional Lebesgue measure from the one-dimensional one by the product measure construction:

Exercise 1.1.9. Let $(X_1, \mathcal{X}_1, \mu_1), \ldots, (X_n, \mathcal{X}_n, \mu_n)$ be a finite collection of measure spaces, and let $(\prod_{i=1}^n X_i, \prod_{i=1}^n \mathcal{X}_i)$ be the product measurable space . Show that there exists a unique measure μ on this space such that $\mu(\prod_{i=1}^n A_i) = \prod_{i=1}^n \mu(A_i)$ for all $A_i \in \mathcal{X}_i$. The measure μ is referred to as the product measure of the $\mu_1, \ldots, \mu_n$ and is denoted $\prod_{i=1}^n \mu_i$.

Exercise 1.1.10. Let E be a Lebesgue measurable subset of $\mathbf{R}^n$. and let m be Lebesgue measure. Establish the inner regularity property

$$m(E) = \sup\{\mu(K) : K \subset E, \text{ compact}\} \tag{1.3}$$

and the outer regularity property

$$m(E) = \inf\{\mu(U) : E \subset U, \text{open}\}. \tag{1.4}$$

Combined with the fact that m is locally finite, this implies that m is a Radon measure.

1.1.3. Integration. Now we define integration on a measure space $(X, \mathcal{X}, \mu)$.

Definition 1.1.19 (Integration)**.** Let $(X, \mathcal{X}, \mu)$ be a measure space.

(i) If $f : X \to [0, +\infty]$ is a non-negative simple function (i.e. a measurable function that only takes on finitely many values $a_1, \ldots, a_n$), we define the *integral* $\int_X f \, d\mu$ of f to be $\int_X f \, d\mu = \sum_{i=1}^n a_i \mu(f^{-1}(\{a_i\}))$ (with the convention that $\infty \cdot 0 = 0$). In particular, if $f = 1_A$ is the indicator function of a measurable set A, then $\int_X 1_A \, d\mu = \mu(A)$.

(ii) If $f : X \to [0, +\infty]$ is a non-negative measurable function, we define the *integral* $\int_X f\ d\mu$ to be the supremum of $\int_X g\ d\mu$, where g ranges over all simple functions bounded between 0 and f.

(iii) If $f : X \to [-\infty, +\infty]$ is a measurable function, whose positive and negative parts $f_+ := \max(f, 0)$, $f_- := \max(-f, 0)$ have finite integral, we say that f is *absolutely integrable* and define $\int_X f\ d\mu := \int_X f_+\ d\mu - \int_X f_-\ d\mu$.

(iv) If $f : X \to \mathbf{C}$ is a measurable function with real and imaginary parts absolutely integrable, we say that f is *absolutely integrable* and define $\int_X f\ d\mu := \int_X \mathrm{Re}\, f\ d\mu + i \int_X \mathrm{Im}\, f\ d\mu$.

We will sometimes show the variable of integration, e.g. writing $\int_X f(x)\ d\mu(x)$ for $\int_X f\ d\mu$, for sake of clarity.

The following results are standard, and the proofs are omitted:

Theorem 1.1.20 (Standard facts about integration)**.** *Let $(X, \mathcal{X}, \mu)$ be a measure space.*

- *All the above integration notions are compatible with each other; for instance, if f is both non-negative and absolutely integrable, then the definitions (ii) and (iii) (and (iv)) agree.*
- *The functional $f \mapsto \int_X f\ d\mu$ is linear over $\mathbf{R}^+$ for simple functions or non-negative functions, is linear over $\mathbf{R}$ for real-valued absolutely integrable functions, and linear over $\mathbf{C}$ for complex-valued absolutely integrable functions. In particular, the set of (real or complex) absolutely integrable functions on $(X, \mathcal{X}, \mu)$ is a (real or complex) vector space.*
- *A complex-valued measurable function $f : X \to \mathbf{C}$ is absolutely integrable if and only if $\int_X |f|\ d\mu < \infty$, in which case we have the triangle inequality $|\int_X f\ d\mu| \leq \int_X |f|\ d\mu$. Of course, the same claim holds for real-valued measurable functions.*
- *If $f : X \to [0, +\infty]$ is non-negative, then $\int_X f\ d\mu \geq 0$, with equality holding if and only if $f = 0$ a.e..*
- *If one modifies an absolutely integrable function on a set of measure zero, then the new function is also absolutely*

integrable, and has the same integral as the original function. Similarly, two non-negative functions that agree a.e. have the same integral. (Because of this, we can meaningfully integrate functions that are only defined almost everywhere.)

- *If $f : X \to \mathbf{C}$ is absolutely integrable, then f is finite a.e., and vanishes outside of a σ-finite set.*
- *If $f : X \to \mathbf{C}$ is absolutely integrable, and $\varepsilon > 0$ then there exists a complex-valued simple function $g : X \to \mathbf{C}$ such that $\int_X |f - g| \, d\mu \leq \varepsilon$. (This is a manifestation of* Littlewood's second principle*.)*
- *(Change of variables formula) If $\phi : X \to Y$ is a measurable map to another measurable space $(Y, \mathcal{Y})$, and $g : Y \to \mathbf{C}$, then we have $\int_X g \circ \phi \, d\mu = \int_Y g \, d\phi_* \mu$, in the sense that whenever one of the integrals is well defined, then the other is also, and equals the first.*
- *It is also important to note that the Lebesgue integral on $\mathbf{R}^n$ extends the more classical Riemann integral. As a consequence, many properties of the Riemann integral (e.g. change of variables formula with respect to smooth diffeomorphisms) are inherited by the Lebesgue integral, thanks to various limiting arguments.*

We now recall the fundamental convergence theorems relating limits and integration: the first three are for non-negative functions, the last three are for absolutely integrable functions. They are ultimately derived from their namesakes in Exercise 1.1.5 and an approximation argument by simple functions, and the proofs are again omitted. (They are also closely related to each other, and are in fact largely equivalent.)

Theorem 1.1.21 (Convergence theorems)**.** *Let $(X, \mathcal{X}, \mu)$ be a measure space.*

- *(Monotone convergence for sequences) If $0 \leq f_1 \leq f_2 \leq \ldots$ are measurable, then $\int_X \lim_{n\to\infty} f_n \, d\mu = \lim_{n\to\infty} \int_X f_n \, d\mu$.*
- *(Monotone convergence for series) If $f_n : X \to [0, +\infty]$ are measurable, then $\int_X \sum_{n=1}^\infty f_n \, d\mu = \sum_{n=1}^\infty \int_X f_n \, d\mu$.*

- *(Fatou's lemma) If $f_n : X \to [0,+\infty]$ are measurable, then $\int_X \liminf_{n\to\infty} f_n \, d\mu \leq \liminf_{n\to\infty} \int_X f_n \, d\mu$.*
- *(Dominated convergence for sequences) If $f_n : X \to \mathbf{C}$ are measurable functions converging pointwise a.e. to a limit f, and $|f_n| \leq g$ a.e. for some absolutely integrable $g : X \to [0,+\infty]$, then $\int_X \lim_{n\to\infty} f_n \, d\mu = \lim_{n\to\infty} \int_X f_n \, d\mu$.*
- *(Dominated convergence for series) If $f_n : X \to \mathbf{C}$ are measurable functions with $\sum_n \int_X |f_n| \, d\mu < \infty$, then $\sum_n f_n(x)$ is absolutely convergent for a.e. x and $\int_X \sum_{n=1}^\infty f_n \, d\mu = \sum_{n=1}^\infty \int_X f_n \, d\mu$.*
- *(Egorov's theorem) If $f_n : X \to \mathbf{C}$ are measurable functions converging pointwise a.e. to a limit f on a subset A of X of finite measure, and $\varepsilon > 0$, then there exists a set of measure at most ε, outside of which f_n converges uniformly to f in A. (This is a manifestation of* Littlewood's third principle*.)*

Remark 1.1.22. As a rule of thumb, if one does not have exact or approximate monotonicity or domination (where "approximate" means "up to an error e whose L^1 norm $\int_X |e| \, d\mu$ goes to zero"), then one should not expect the integral of a limit to equal the limit of the integral in general; there is just too much room for oscillation.

Exercise 1.1.11. Let $f : X \to \mathbf{C}$ be an absolutely integrable function on a measure space $(X, \mathcal{X}, \mu)$. Show that f is uniformly integrable, in the sense that for every $\varepsilon > 0$ there exists $\delta > 0$ such that $\int_E |f| \, d\mu \leq \varepsilon$ whenever E is a measurable set of measure at most δ. (The property of uniform integrability becomes more interesting, of course when applied to a family of functions, rather than to a single function.)

With regard to product measures and integration, the fundamental theorem in this subject is

Theorem 1.1.23 (Fubini-Tonelli theorem)**.** *Let $(X, \mathcal{X}, \mu)$ and $(Y, \mathcal{Y}, \nu)$ be σ-finite measure spaces, with product space $(X \times Y, \mathcal{X} \times \mathcal{Y}, \mu \times \nu)$.*

- *(Tonelli theorem) If $f : X \times Y \to [0,+\infty]$ is measurable, then $\int_{X\times Y} f \, d\mu\times\nu = \int_X (\int_Y f(x,y) \, d\nu(y)) \, d\mu(x) = \int_Y (\int_X f(x,y) \, d\mu(x)) d\nu(y)$.*

- *(Fubini theorem) If $f : X \times Y \to \mathbf{C}$ is absolutely integrable, then we also have $\int_{X\times Y} f\, d\mu\times\nu = \int_X (\int_Y f(x,y)\, d\nu(y))\, d\mu(x) = \int_Y (\int_X f(x,y)\, d\mu(x)) d\nu(y)$, with the inner integrals being absolutely integrable a.e. and the outer integrals all being absolutely integrable.*

If $(X, \mathcal{X}, \mu)$ and $(Y, \mathcal{Y}, \nu)$ are complete measure spaces, then the same claims hold with the product σ-algebra $\mathcal{X} \times \mathcal{Y}$ replaced by its completion.

Remark 1.1.24. The theorem fails for non-σ-finite spaces, but virtually every measure space actually encountered in "hard analysis" applications will be σ-finite. (One should be cautious, however, with any space constructed using *ultrafilters* or the *first uncountable ordinal*.) It is also important that f obey some measurability in the product space; there exist non-measurable f for which the iterated integrals exist (and may or may not be equal to each other, depending on the properties of f and even on which axioms of set theory one chooses), but the product integral (of course) does not.

Notes. This lecture first appeared at `terrytao.wordpress.com/2009/01/01`. Thanks to Andy, PDEBeginner, Phil, Sune Kristian Jacobsen, wangtwo, and an anonymous commenter for corrections.

Several commenters noted *Solovay's theorem*, which asserts that there exist models of set theory without the axiom of choice in which all sets are measurable. This led to some discussion of the extent in which one could formalise the claim that any set which could be defined without the axiom of choice was necessarily measurable, but the discussion was inconclusive.

1.2. Signed measures and the Radon-Nikodym-Lebesgue theorem

In this section, $X = (X, \mathcal{X})$ is a fixed measurable space. We shall often omit the σ-algebra $\mathcal{X}$, and simply refer to elements of $\mathcal{X}$ as measurable sets. Unless otherwise indicated, all subsets of X appearing below are restricted to be measurable, and all functions on X appearing below are also restricted to be measurable.

We let $\mathcal{M}_+(X)$ denote the space of measures on X, i.e. functions $\mu : \mathcal{X} \to [0, +\infty]$ which are countably additive and send $\emptyset$ to 0. For reasons that will be clearer later, we shall refer to such measures as *unsigned measures*. In this section we investigate the structure of this space, together with the closely related spaces of signed measures and finite measures.

Suppose that we have already constructed one unsigned measure $m \in \mathcal{M}_+(X)$ on X (e.g. think of X as the real line with the Borel σ-algebra, and let m be Lebesgue measure). Then we can obtain many further unsigned measures on X by multiplying m by a function $f : X \to [0, +\infty]$, to obtain a new unsigned measure m_f, defined by the formula

$$m_f(E) := \int_X 1_E f \, d\mu \tag{1.5}$$

If $f = 1_A$ is an indicator function, we write $m \lfloor_A$ for m_{1_A}, and refer to this measure as the *restriction* of m to A.

Exercise 1.2.1. Show (using the monotone convergence theorem, Theorem 1.1.21) that m_f is indeed a unsigned measure, and for any $g : X \to [0, +\infty]$, we have $\int_X g \, dm_f = \int_X gf \, dm$. We will express this relationship symbolically as

$$dm_f = f dm. \tag{1.6}$$

Exercise 1.2.2. Let m be σ-finite. Given two functions $f, g : X \to [0, +\infty]$, show that $m_f = m_g$ if and only if $f(x) = g(x)$ for m-almost every x. (*Hint*: as usual, first do the case when m is finite. The key point is that if f and g are not equal m-almost everywhere, then either $f > g$ on a set of positive measure, or $f < g$ on a set of positive measure.) Give an example to show that this uniqueness statement can fail if m is not σ-finite. (*Hint*: the space X can be very simple.)

In view of Exercises 1.2.1 and 1.2.2, let us temporarily call a measure μ *differentiable with respect to* m if $d\mu = f dm$ (i.e. $\mu = m_f$) for some $f : X \to [0, +\infty]$, and call f the *Radon-Nikodym derivative* of μ with respect to m, writing

$$f = \frac{d\mu}{dm}; \tag{1.7}$$

by Exercise 1.2.2, we see if m is σ-finite that this derivative is defined up to m-almost everywhere equivalence.

Exercise 1.2.3 (Relationship between Radon-Nikodym derivative and classical derivative)**.** Let m be Lebesgue measure on $[0, +\infty)$, and let μ be an unsigned measure that is differentiable with respect to m. If μ has a continuous Radon-Nikodym derivative $\frac{d\mu}{dm}$, show that the function $x \mapsto \mu([0, x])$ is differentiable, and $\frac{d}{dx}\mu([0, x]) = \frac{d\mu}{dm}(x)$ for all x.

Exercise 1.2.4. Let X be at most countable. Show that every measure on X is differentiable with respect to counting measure $\#$.

If every measure was differentiable with respect to m (as is the case in Exercise 1.2.4), then we would have completely described the space of measures of X in terms of the non-negative functions of X (modulo m-almost everywhere equivalence). Unfortunately, not every measure is differentiable with respect to every other: for instance, if x is a point in X, then the only measures that are differentiable with respect to the Dirac measure δ_x are the scalar multiples of that measure. We will explore the precise obstruction that prevents all measures from being differentiable, culminating in the Radon-Nikodym-Lebesgue theorem that gives a satisfactory understanding of the situation in the σ-finite case (which is the case of interest for most applications).

In order to establish this theorem, it will be important to first study some other basic operations on measures, notably the ability to subtract one measure from another. This will necessitate the study of *signed measures*, to which we now turn.

1.2.1. Signed measures. We have seen that if we fix a reference measure m, then non-negative functions $f : X \to [0, +\infty]$ (modulo m-almost everywhere equivalence) can be identified with unsigned measures $m_f : \mathcal{X} \to [0, +\infty]$. This motivates various operations on measures that are analogous to operations on functions (indeed, one could view measures as a kind of "generalised function" with respect to a fixed reference measure m). For instance, we can define the sum

of two unsigned measures $\mu, \nu : \mathcal{X} \to [0, +\infty]$ as

$$(\mu + \nu)(E) := \mu(E) + \nu(E) \tag{1.8}$$

and non-negative scalar multiples $c\mu$ for $c > 0$ by

$$(c\mu)(E) := c(\mu(E)). \tag{1.9}$$

We can also say that one measure μ is *less than* another ν if

$$\mu(E) \leq \nu(E) \text{ for all } E \in \mathcal{X}. \tag{1.10}$$

These operations are all consistent with their functional counterparts, e.g. $m_{f+g} = m_f + m_g$, etc.

Next, we would like to define the difference $\mu - \nu$ of two unsigned measures. The obvious thing to do is to define

$$(\mu - \nu)(E) := \mu(E) - \nu(E) \tag{1.11}$$

but we have a problem if $\mu(E)$ and $\nu(E)$ are both infinite: $\infty - \infty$ is undefined! To fix this problem, we will only define the difference of two unsigned measures μ, ν if at least one of them is a finite measure. Observe that in that case, $\mu - \nu$ takes values in $(-\infty, +\infty]$ or $[-\infty, +\infty)$, but not both.

Of course, we no longer expect $\mu - \nu$ to be monotone. However, it is still finitely additive, and even countably additive in the sense that the sum $\sum_{n=1}^{\infty} (\mu - \nu)(E_n)$ converges to $(\mu - \nu)(\bigcup_{n=1}^{\infty} E_n)$ whenever $E_1, E_2, \ldots$ are disjoint sets, and furthermore that the sum is absolutely convergent when $(\mu - \nu)(\bigcup_{n=1}^{\infty} E_n)$ is finite. This motivates

Definition 1.2.1 (Signed measure)**.** A *signed measure* is a map $\mu : \mathcal{X} \to [-\infty, +\infty]$ such that

(i) $\mu(\emptyset) = 0$;

(ii) μ can take either the value $+\infty$ or $-\infty$, but not both;

(iii) If $E_1, E_2, \ldots \subset X$ are disjoint, then $\sum_{n=1}^{\infty} \mu(E_n)$ converges to $\mu(\bigcup_{n=1}^{\infty} E_n)$, with the former sum being absolutely convergent[1] if the latter expression is finite.

[1] Actually, the absolute convergence is automatic from the *Riemann rearrangement theorem*. Another consequence of (iii) is that any subset of a finite measure set is again finite measure, and the finite union of finite measure sets again has finite measure.

Thus every unsigned measure is a signed measure, and the difference of two unsigned measures is a signed measure if at least one of the unsigned measures is finite; we will see shortly that the converse statement is also true, i.e. every signed measure is the difference of two unsigned measures (with one of the unsigned measures being finite). Another example of a signed measure are the measures m_f defined by (1.5), where $f : X \to [-\infty, +\infty]$ is now signed rather than unsigned, but with the assumption that at least one of the signed parts $f_+ := \max(f, 0)$, $f_- := \max(-f, 0)$ of f is absolutely integrable.

We also observe that a signed measure μ is unsigned if and only if $\mu \geq 0$ (where we use (1.10) to define order on measures).

Given a function $f : X \to [-\infty, +\infty]$, we can partition X into one set $X_+ := \{x : f(x) \geq 0\}$ on which f is non-negative, and another set $X_- := \{x : f(x) < 0\}$ on which f is negative; thus $f \restriction_{X_+} \geq 0$ and $f \restriction_{X_-} \leq 0$. It turns out that the same is true for signed measures:

Theorem 1.2.2 (Hahn decomposition theorem)**.** *Let μ be a signed measure. Then one can find a partition $X = X_+ \cup X_-$ such that $\mu \restriction_{X_+} \geq 0$ and $\mu \restriction_{X_-} \leq 0$.*

Proof. By replacing μ with $-\mu$ if necessary, we may assume that μ avoids the value $+\infty$.

Call a set E *totally positive* if $\mu \restriction_E \geq 0$, and *totally negative* if $\mu \restriction_E \leq 0$. The idea is to pick X_+ to be the totally positive set of maximal measure - a kind of "greedy algorithm", if you will. More precisely, define m_+ to be the supremum of $\mu(E)$, where E ranges over all totally positive sets. (The supremum is non-vacuous, since the empty set is totally positive.) We claim that the supremum is actually attained. Indeed, we can always find a maximising sequence $E_1, E_2, \ldots$ of totally positive sets with $\mu(E_n) \to m_+$. It is not hard to see that the union $X_+ := \bigcup_{n=1}^{\infty} E_n$ is also totally positive, and $\mu(X_+) = m_+$ as required. Since μ avoids $+\infty$, we see in particular that m_+ is finite.

Set $X_- := X \backslash X_+$. We claim that X_- is totally negative. We do this as follows. Suppose for contradiction that X_- is not totally negative, then there exists a set E_1 in X_- of strictly positive measure. If E_1 is totally positive, then $X_+ \cup E_1$ is a totally positive set having

measure strictly greater than m_+, a contradiction. Thus E_1 must contain a subset E_2 of strictly larger measure. Let us pick E_2 so that $\mu(E_2) \geq \mu(E_1) + 1/n_1$, where n_1 is the smallest integer for which such an E_2 exists. If E_2 is totally positive, then we are again done, so we can find a subset E_3 with $\mu(E_3) \geq \mu(E_2) + 1/n_2$, where n_2 is the smallest integer for whch such a E_3 exists. Continuing in this fashion, we either stop and get a contradiction, or obtain a nested sequence of sets $E_1 \supset E_2 \supset \ldots$ in X_- of increasing positive measure (with $\mu(E_{j+1}) \geq \mu(E_j) + 1/n_j$). The intersection $E := \bigcap_j E_j$ then also has positive measure, hence finite, which implies that the n_j go to infinity; it is then not difficult to see that E itself cannot contain any subsets of strictly larger measure, and so E is a totally positive set of positive measure in X_-, and we again obtain a contradiction. □

Remark 1.2.3. A somewhat simpler proof of the Hahn decomposition theorem is available if we assume μ to be finite positive variation (which means that $\mu(E)$ is bounded above as E varies). For each positive n, let E_n be a set whose measure $\mu(E_n)$ is within 2^{-n} of $\sup\{\mu(E) : E \in \mathcal{X}\}$. One can easily show that any subset of $E_n \backslash E_{n-1}$ has measure $O(2^{-n})$, and in particular that $E_n \backslash \bigcup_{n'=n_0}^{n-1} E_{n-1}$ has measure $O(2^{-n})$ for any $n_0 \leq n$. This allows one to control the unions $\bigcup_{n=n_0}^{\infty} E_n$, and thence the lim sup X_+ of the E_n, which one can then show to have the required properties. One can in fact show that any signed measure that avoids $+\infty$ must have finite positive variation, but this turns out to require a certain amount of work.

Let us say that a set E is *null* for a signed measure μ if $\mu \lfloor_E = 0$. (This implies that $\mu(E) = 0$, but the converse is not true, since a set E of signed measure zero could contain subsets of non-zero measure.) It is easy to see that the sets X_-, X_+ given by the Hahn decomposition theorem are unique modulo null sets.

Let us say that a signed measure μ is *supported* on E if the complement of E is null (or equivalently, if $\mu \lfloor_E = \mu$. If two signed measures μ, ν can be supported on disjoint sets, we say that they are mutually singular (or that μ is singular with respect to ν) and write $\mu \perp \nu$. If we write $\mu_+ := \mu \lfloor_{X_+}$ and $\mu_- := -\mu \lfloor_{X_-}$, we thus soon establish

Exercise 1.2.5 (Jordan decomposition theorem)**.** Every signed measure μ an be uniquely decomposed as $\mu = \mu_+ - \mu_-$, where μ_+, μ_- are mutually singular unsigned measures. (The only claim not already established is the uniqueness.) We refer to μ_+, μ_- as the *positive and negative parts* (or *positive and negative variations*) of μ.

This is of course analogous to the decomposition $f = f_+ - f_-$ of a function into positive and negative parts. Inspired by this, we define the *absolute value* (or *total variation*) $|\mu|$ of a signed measure to be $|\mu| := \mu_+ + \mu_-$.

Exercise 1.2.6. Show that $|\mu|$ is the minimal unsigned measure such that $-|\mu| \leq \mu \leq |\mu|$. Furthermore, $|\mu|(E)$ is equal to the maximum value of $\sum_{n=1}^{\infty} |\mu(E_n)|$, where $(E_n)_{n=1}^{\infty}$ ranges over the partitions of E. (This may help explain the terminology "total variation".)

Exercise 1.2.7. Show that $\mu(E)$ is finite for every E if and only if $|\mu|$ is a finite unsigned measure, if and only if μ_+, μ_- are finite unsigned measures. If any of these properties hold, we call μ a *finite measure*. (In a similar spirit, we call a signed measure μ *σ-finite* if $|\mu|$ is σ-finite.)

The space of finite measures on X is clearly a real vector space, and is denoted $\mathcal{M}(X)$.

1.2.2. The Lebesgue-Radon-Nikodym theorem. Let m be a reference unsigned measure. We saw at the beginning of this section that the map $f \mapsto m_f$ is an embedding of the space $L^+(X, dm)$ of non-negative functions (modulo m-almost everywhere equivalence) into the space $\mathcal{M}^+(X)$ of unsigned measures. The same map is also an embedding of the space $L^1(X, dm)$ of absolutely integrable functions (again modulo m-almost everywhere equivalence) into the space $\mathcal{M}(X)$ of finite measures. (To verify this, one first makes the easy observation that the Jordan decomposition of a measure m_f given by an absolutely integrable function f is simply $m_f = m_{f_+} - m_{f_-}$.)

In the converse direction, one can ask if every finite measure μ in $\mathcal{M}(X)$ can be expressed as m_f for some absolutely integrable f. Unfortunately, there are some obstructions to this. Firstly, from (1.5) we see that if $\mu = m_f$, then any set that has measure zero with respect

to m, must also have measure zero with respect to μ. In particular, this implies that a non-trivial measure that is singular with respect to m cannot be expressed in the form m_f.

In the σ-finite case, this turns out to be the only obstruction:

Theorem 1.2.4 (Lebesgue-Radon-Nikodym theorem)**.** *Let m be an unsigned σ-finite measure, and let μ be a signed σ-finite measure. Then there exists a unique decomposition $\mu = m_f + \mu_s$, where $f \in L^1(X, dm)$ and $\mu_s \perp m$. If μ is unsigned, then f and μ_s are also.*

Proof. We prove this only for the case when μ, ν are finite rather than σ-finite, and leave the general case as an exercise. The uniqueness follows from Exercise 1.2.2 and the previous observation that m_f cannot be mutually singular with m for any non-zero f, so it suffices to prove existence. By the Jordan decomposition theorem, we may assume that μ is unsigned as well. (In this case, we expect f and μ_s to be unsigned also.)

The idea is to select f "greedily". More precisely, let M be the supremum of the quantity $\int_X f\ dm$, where f ranges over all non-negative functions such that $m_f \leq \mu$. Since μ is finite, M is finite. We claim that the supremum is actually attained for some f. Indeed, if we let f_n be a maximising sequence, thus $m_{f_n} \leq \mu$ and $\int_X f_n\ dm \to M$, one easily checks that the function $f = \sup_n f_n$ attains the supremum.

The measure $\mu_s := \mu - m_f$ is a non-negative finite measure by construction. To finish the theorem, it suffices to show that $\mu_s \perp m$.

It will suffice to show that $(\mu_s - \varepsilon m)_+ \perp m$ for all ε, as the claim then easily follows by letting ε be a countable sequence going to zero. But if $(\mu_s - \varepsilon m)_+$ were not singular with respect to m, we see from the Hahn decomposition theorem that there is a set E with $m(E) > 0$ such that $(\mu_s - \varepsilon m) \restriction_E \geq 0$, and thus $\mu_s \geq \varepsilon m \restriction_E$. But then one could add $\varepsilon 1_E$ to f, contradicting the construction of f. □

Exercise 1.2.8. Complete the proof of Theorem 1.2.4 for the σ-finite case.

We have the following corollary:

Corollary 1.2.5 (Radon-Nikodym theorem)**.** *Let m be an unsigned σ-finite measure, and let μ be a signed σ-finite measure. Then the following are equivalent.*

(i) *$\mu = m_f$ for some $f \in L^1(X, dm)$.*

(ii) *$\mu(E) = 0$ whenever $m(E) = 0$.*

(iii) *For every $\varepsilon > 0$, there exists $\delta > 0$ such that $\mu(E) < \varepsilon$ whenever $m(E) \leq \delta$.*

When any of these statements occur, we say that μ is absolutely continuous *with respect to m, and write $\mu \ll m$. As in the start of this section, we call f the* Radon-Nikodym derivative *of μ with respect to m, and write $f = \frac{d\mu}{dm}$.*

Proof. The implication of (iii) from (i) is Exercise 1.1.11. The implication of (ii) from (iii) is trivial. To deduce (i) from (ii), apply Theorem 1.2.2 to μ and observe that μ_s is supported on a set of m-measure zero E by hypothesis. Since E is null for m, it is null for m_f and μ also, and so μ_s is trivial, giving (i). □

Corollary 1.2.6 (Lebesgue decomposition theorem)**.** *Let m be an unsigned σ-finite measure, and let μ be a signed σ-finite measure. Then there is a unique decomposition $\mu = \mu_{ac} + \mu_s$, where $\mu_{ac} \ll m$ and $\mu_s \perp m$. (We refer to μ_{ac} and μ_s as the* absolutely continuous *and* singular *components of μ with respect to m.) If μ is unsigned, then μ_{ac} and μ_s are also.*

Exercise 1.2.9. If every point in X is measurable, we call a signed measure μ *continuous* if $\mu(\{x\}) = 0$ for all x. Let the hypotheses be as in Corollary 1.2.6, but suppose also that every point is measurable and m is continuous. Show that there is a unique decomposition $\mu = \mu_{ac} + \mu_{sc} + \mu_{pp}$, where $\mu_{ac} \ll m$, μ_{pp} is supported on an at most countable set, and μ_{sc} is both singular with respect to m and continuous. Furthermore, if μ is unsigned, then $\mu_{ac}, \mu_{sc}, \mu_{pp}$ are also. We call μ_{sc} and μ_{pp} the *singular continuous* and *pure point* components of μ respectively.

Example 1.2.7. A *Cantor measure* is singular continuous with respect to Lebesgue measure, while *Dirac measures* are pure point.

Lebesgue measure on a line is singular continuous with respect to Lebesgue measure on a plane containing that line.

Remark 1.2.8. Suppose one is decomposing a measure μ on a Euclidean space $\mathbf{R}^d$ with respect to Lebesgue measure m on that space. Very roughly speaking, a measure is pure point if it is supported on a 0-dimensional subset of $\mathbf{R}^d$, it is absolutely continuous if its support is spread out on a full dimensional subset, and is singular continuous if it is supported on some set of dimension intermediate between 0 and d. For instance, if μ is the sum of a Dirac mass at $(0,0) \in \mathbf{R}^2$, one-dimensional Lebesgue measure on the x-axis, and two-dimensional Lebesgue measure on $\mathbf{R}^2$, then these are the pure point, singular continuous, and absolutely continuous components of μ respectively. This heuristic is not completely accurate (in part because I have left the definition of "dimension" vague) but is not a bad rule of thumb for a first approximation. We will study analytic concepts of dimension in more detail in Section 1.15.

To motivate the terminology "continuous" and "singular continuous", we recall two definitions on an interval $I \subset \mathbf{R}$, and make a third:

- A function $f : I \to \mathbf{R}$ is *continuous* if for every $x \in I$ and every $\varepsilon > 0$, there exists $\delta > 0$ such that $|f(y) - f(x)| \leq \varepsilon$ whenever $y \in I$ is such that $|y - x| \leq \delta$.
- A function $f : I \to \mathbf{R}$ is *uniformly continuous* if for every $\varepsilon > 0$, there exists $\delta > 0$ such that $|f(y) - f(x)| \leq \varepsilon$ whenever $[x, y] \subset I$ has length at most δ.
- A function $f : I \to \mathbf{R}$ is *absolutely continuous* if for every $\varepsilon > 0$, there exists $\delta > 0$ such that $\sum_{i=1}^{n} |f(y_i) - f(x_i)| \leq \varepsilon$ whenever $[x_1, y_1], \ldots, [x_n, y_n]$ are disjoint intervals in I of total length at most δ.

Clearly, absolute continuity implies uniform continuity, which in turn implies continuity. The significance of absolute continuity is that it is the largest class of functions for which the fundamental theorem of calculus holds (using the classical derivative, and the Lebesgue integral), as can be seen in any introductory graduate real analysis course.

Exercise 1.2.10. Let m be Lebesgue measure on the interval $[0, +\infty]$, and let μ be a finite unsigned measure.

Show that μ is a continuous measure if and only if the function $x \mapsto \mu([0, x])$ is continuous. Show that μ is an absolutely continuous measure with respect to m if and only if the function $x \mapsto \mu([0, x])$ is absolutely continuous.

1.2.3. A finitary analogue of the Lebesgue decomposition (optional). At first glance, the above theory is only non-trivial when the underlying set X is infinite. For instance, if X is finite, and m is the uniform distribution on X, then every other measure on X will be absolutely continuous with respect to m, making the Lebesgue decomposition trivial. Nevertheless, there is a non-trivial version of the above theory that can be applied to finite sets (cf. Section 1.3 of *Structure and Randomness*). The cleanest formulation is to apply it to a sequence of (increasingly large) sets, rather than to a single set:

Theorem 1.2.9 (Finitary analogue of the Lebesgue-Radon-Nikodym theorem)**.** *Let X_n be a sequence of finite sets (and with the discrete σ-algebra), and for each n, let m_n be the uniform distribution on X_n, and let μ_n be another probability measure on X_n. Then, after passing to a subsequence, one has a decomposition*

$$\mu_n = \mu_{n,ac} + \mu_{n,sc} + \mu_{n,pp} \tag{1.12}$$

where

(i) *(Uniform absolute continuity) For every $\varepsilon > 0$, there exists $\delta > 0$ (independent of n) such that $\mu_{n,ac}(E) \leq \varepsilon$ whenever $m_n(E) \leq \delta$, for all n and all $E \subset X_n$.*

(ii) *(Asymptotic singular continuity) $\mu_{n,sc}$ is supported on a set of m_n-measure $o(1)$, and we have $\mu_{n,sc}(\{x\}) = o(1)$ uniformly for all $x \in X_n$, where $o(1)$ denotes an error that goes to zero as $n \to \infty$.*

(iii) *(Uniform pure point) For every $\varepsilon > 0$ there exists $N > 0$ (independent of n) such that for each n, there exists a set $E_n \subset X_n$ of cardinality at most N such that $\mu_{n,pp}(X_n \backslash E_n) \leq \varepsilon$.*

Proof. Using the Radon-Nikodym theorem (or just working by hand, since everything is finite), we can write $d\mu_n = f_n\ dm_n$ for some $f_n : X_n \to [0, +\infty)$ with average value 1.

For each positive integer k, the sequence $\mu_n(\{f_n \geq k\})$ is bounded between 0 and 1, so by the *Bolzano-Weierstrass theorem*, it has a convergent subsequence. Applying the usual diagonalisation argument (as in the proof of the *Arzelá-Ascoli theorem*, Theorem 1.8.23), we may thus assume (after passing to a subsequence, and relabeling) that $\mu_n(\{f_n \geq k\})$ converges for positive k to some limit c_k.

Clearly, the c_k are decreasing and range between 0 and 1, and so converge as $k \to \infty$ to some limit $0 < c < 1$.

Since $\lim_{k\to\infty} \lim_{n\to\infty} \mu_n(\{f_n \geq k\}) = c$, we can find a sequence k_n going to infinity such that $\mu_n(\{f_n \geq k_n\}) \to c$ as $n \to \infty$. We now set $\mu_{n,ac}$ to be the restriction of μ_n to the set $\{f_n < k_n\}$. We claim the absolute continuity property (i). Indeed, for any $\varepsilon > 0$, we can find a k such that $c_k \geq c - \varepsilon/10$. For n sufficiently large, we thus have

$$\mu_n(\{f_n \geq k\}) \geq c - \varepsilon/5 \tag{1.13}$$

and

$$\mu_n(\{f_n \geq k_n\}) \leq c + \varepsilon/5 \tag{1.14}$$

and hence

$$\mu_{n,ac}(\{f_n \geq k\}) \leq 2\varepsilon/5. \tag{1.15}$$

If we take $\delta < \varepsilon/5k$, we thus see (for n sufficiently large) that (i) holds. (For the remaining n, one simply shrinks δ as much as is necessary.)

Write $\mu_{n,s} := \mu_n - \mu_{n,ac}$, thus $\mu_{n,s}$ is supported on a set of size $|X_n|/K_n = o(|X_n|)$ by *Markov's inequality*. It remains to extract out the pure point components. This we do by a similar procedure as above. Indeed, by arguing as before we may assume (after passing to a subsequence as necessary) that the quantities $\mu_n\{x : \mu_n(\{x\}) \geq 1/j\}$ converge to a limit d_j for each positive integer j, that the d_j themselves converge to a limit d, and that there exists a sequence $j_n \to \infty$ such that $\mu_n\{x : \mu_n(\{x\}) \geq 1/j_n\}$ converges to d. If one sets μ_{sc} and μ_{pp} to be the restrictions of μ_s to the sets $\{x : \mu_n(\{x\}) <$

$1/j_n\}$ and $\{x : \mu_n(\{x\}) \geq 1/j_n\}$ respectively, one can verify the remaining claims by arguments similar to those already given. □

Exercise 1.2.11. Generalise Theorem 1.2.9 to the setting where the X_n can be infinite and non-discrete (but we still require every point to be measurable), the m_n are arbitrary probability measures, and the μ_n are arbitrary finite measures of uniformly bounded total variation.

Remark 1.2.10. This result is still not fully "finitary" because it deals with a sequence of finite structures, rather than with a single finite structure. It appears in fact to be quite difficult (and perhaps even impossible) to make a fully finitary version of the Lebesgue decomposition (in the same way that the finite convergence principle in Section 1.3 of *Structure and Randomness*vwas a fully finitary analogue of the infinite convergence principle), though one can certainly form some weaker finitary statements that capture a portion of the strength of this theorem. For instance, one very cheap thing to do, given two probability measures μ, m, is to introduce a threshold parameter k, and partition $\mu = \mu_{\leq k} + \mu_{>k}$, where $\mu_{\leq k} \leq km$, and $\mu_{>k}$ is supported on a set of m-measure at most $1/k$; such a decomposition is automatic from Theorem 1.2.4 and Markov's inequality, and has meaningful content even when the underlying space X is finite, but this type of decomposition is not as powerful as the full Lebesgue decompositions (mainly because the size of the support for $\mu_{>k}$ is relatively large compared to the threshold k). Using the finite convergence principle, one can do a bit better, writing $\mu = \mu_{\leq k} + \mu_{k<\cdot\leq F(k)} + \mu_{\geq F(k)}$ for any function F and any $\varepsilon > 0$, where $k = O_{F,\varepsilon}(1)$, $\mu_{\leq k} \leq km$, $\mu_{\geq F(k)}$ is supported on a set of m-measure at most $1/F(k)$, and $\mu_{k<\cdot\leq F(k)}$ has total mass at most ε, but this is still fails to capture the full strength of the infinitary decomposition, because ε needs to be fixed in advance. I have not been able to find a fully finitary statement that is equivalent to, say, Theorem 1.2.9; I suspect that if it does exist, it will have quite a messy formulation.

Notes. This lecture first appeared at `terrytao.wordpress.com/2009/01/04`. The material here is largely based on Folland's text [**Fo2000**], except for the last section. Thanks to Ke, Max Baroi, Xiaochuan Liu, and several anonymous commenters for corrections.

1.3. L^p spaces

Now that we have reviewed the foundations of measure theory, let us now put it to work to set up the basic theory of one of the fundamental families of function spaces in analysis, namely the *L^p spaces* (also known as *Lebesgue spaces*). These spaces serve as important model examples for the general theory of topological and normed vector spaces, which we will discuss a little bit in this lecture and then in much greater detail in later lectures.

Just as scalar quantities live in the space of real or complex numbers, and vector quantities live in vector spaces, functions $f : X \to \mathbf{C}$ (or other objects closely related to functions, such as measures) live in function spaces. Like other spaces in mathematics (e.g. vector spaces, metric spaces, topological spaces, etc.) a function space V is not just mere sets of objects (in this case, the objects are functions), but they also come with various important structures that allow one to do some useful operations inside these spaces, and from one space to another. For example, function spaces tend to have several (though usually not all) of the following types of structures, which are usually related to each other by various compatibility conditions:

- **Vector space structure.** One can often add two functions f, g in a function space V, and expect to get another function $f + g$ in that space V; similarly, one can multiply a function f in V by a scalar c and get another function cf in V. Usually, these operations obey the axioms of a vector space, though it is important to caution that the dimension of a function space is typically infinite. (In some cases, the space of scalars is a more complicated ring than the real or complex field, in which case we need the notion of a module rather than a vector space, but we will not use this more general notion in this course.) Virtually all of the function spaces we shall encounter in this course will be vector spaces. Because the field of scalars is real or complex, vector spaces also come with the notion of convexity, which turns out to be crucial in many aspects of analysis. As a consequence (and in marked contrast to algebra or number theory), much of the theory in real analysis does not seem

to extend to other fields of scalars (in particular, real analysis fails spectacularly in the finite characteristic setting). Algebra structure. Sometimes (though not always), we also wish to multiply two functions f, g in V and get another function fg in V; when combined with the vector space structure and assuming some compatibility conditions (e.g. the distributive law), this makes V an algebra. This multiplication operation is often just pointwise multiplication, but there are other important multiplication operations on function spaces too, such as[2] *convolution*.

- **Norm structure.** We often want to distinguish "large" functions in V from "small" ones, especially in analysis, in which "small" terms in an expression are routinely discarded or deemed to be acceptable errors. One way to do this is to assign a magnitude or norm $\|f\|_V$ to each function that measures its size. Unlike the situation with scalars, where there is basically a single notion of magnitude, functions have a wide variety of useful notions of size, each measuring a different aspect (or combination of aspects) of the function, such as height, width, oscillation, regularity, decay, and so forth. Typically, each such norm gives rise to a separate function space (although sometimes it is useful to consider a single function space with multiple norms on it). We usually require the norm to be compatible with the vector space structure (and algebra structure, if present), for instance by demanding that the *triangle inequality* hold.
- **Metric structure.** We also want to tell whether two functions f, g in a function space V are "near together" or "far apart". A typical way to do this is to impose a metric $d : V \times V \to \mathbf{R}^+$ on the space V. If both a norm $\|\|_V$ and a vector space structure are available, there is an obvious way to do this: define the distance between two functions f, g in

[2]One sometimes sees other algebraic structures than multiplication appear in function spaces, such as *commutators* and *derivations*, but again we will not encounter those in this course. Another common algebraic operation for function spaces is conjugation or adjoint, leading to the notion of a **-algebra*.

V to be[3] $d(f,g) := \|f-g\|_V$. It is often important to know if the vector space is complete[4] with respect to the given metric; this allows one to take limits of Cauchy sequences, and (with a norm and vector space structure) sum absolutely convergent series, as well as use some useful results from point set topology such as the *Baire category theorem*, see Section 1.7. All of these operations are of course vital in analysis.

- **Topological structure.** It is often important to know when a sequence (or, occasionally, *nets*) of functions f_n in V "converges" in some sense to a limit f (which, hopefully, is still in V); there are often many distinct modes of convergence (e.g. pointwise convergence, uniform convergence, etc.) that one wishes to carefully distinguish from each other. Also, in order to apply various powerful topological theorems (or to justify various formal operations involving limits, suprema, etc.), it is important to know when certain subsets of V enjoy key topological properties (most notably compactness and connectedness), and to know which operations on V are continuous. For all of this, one needs a topology on V. If one already has a metric, then one of course has a topology generated by the open balls of that metric; but there are many important topologies on function spaces in analysis that do not arise from metrics. We also often require the topology to be compatible with the other structures on the function space; for instance, we usually require the vector space operations of addition and scalar multiplication to be continuous. In some cases, the topology on V extends to some natural superspace W of more general functions that contain V; in such cases, it is often

[3] This will be the only type of metric on function spaces encountered in this course. But there are some nonlinear function spaces of importance in nonlinear analysis (e.g. spaces of maps from one manifold to another) which have no vector space structure or norm, but still have a metric.

[4] Compactness would be an even better property than completeness to have, but function spaces unfortunately tend be non-compact in various rather nasty ways, although there are useful partial substitutes for compactness that are available, see e.g. Section 1.6 of *Poincaré's Legacies, Vol. I.*

important to know whether V is closed in W, so that limits of sequences in V stay in V.

- **Functional structures.** Since numbers are easier to understand and deal with than functions, it is not surprising that we often study functions f in a function space V by first applying some functional $\lambda : V \to \mathbf{C}$ to V to identify some key numerical quantity $\lambda(f)$ associated to f. Norms $f \mapsto \|f\|_V$ are of course one important example of a functional; integration $f \mapsto \int_X f\, d\mu$ provides another; and evaluation $f \mapsto f(x)$ at a point x provides a third important class. (Note, though, that while evaluation is the fundamental feature of a function in set theory, it is often a quite minor operation in analysis; indeed, in many function spaces, evaluation is not even defined at all, for instance because the functions in the space are only defined almost everywhere!) An inner product $\langle , \rangle$ on V (see below) also provides a large family $f \mapsto \langle f, g\rangle$ of useful functionals. It is of particular interest to study functionals that are compatible with the vector space structure (i.e. are linear) and with the topological structure (i.e. are continuous); this will give rise to the important notion of duality on function spaces.

- **Inner product structure.** One often would like to pair a function f in a function space V with another object g (which is often, though not always, another function in the same function space V) and obtain a number $\langle f, g\rangle$, that typically measures the amount of "interaction" or "correlation" between f and g. Typical examples include inner products arising from integration, such as $\langle f, g\rangle := \int_X f\overline{g}\, d\mu$; integration itself can also be viewed as a pairing, $\langle f, \mu\rangle := \int_X f\, d\mu$. Of course, we usually require such inner products to be compatible with the other structures present on the space (e.g., to be compatible with the vector space structure, we usually require the inner product to be *bilinear* or *sesquilinear*). Inner products, when available, are incredibly useful in understanding the metric and norm geometry of a space, due

to such fundamental facts as the *Cauchy-Schwarz inequality* and the *parallelogram law*. They also give rise to the important notion of *orthogonality* between functions.

- **Group actions.** We often expect our function spaces to enjoy various symmetries; we might wish to rotate, reflect, translate, modulate, or dilate our functions and expect to preserve most of the structure of the space when doing so. In modern mathematics, symmetries are usually encoded by *group actions* (or actions of other group-like objects, such as semigroups or groupoids; one also often upgrades groups to more structured objects such as Lie groups). As usual, we typically require the group action to preserve the other structures present on the space, e.g. one often restricts attention to group actions that are linear (to preserve the vector space structure), continuous (to preserve topological structure), unitary (to preserve inner product structure), isometric (to preserve metric structure), and so forth. Besides giving us useful symmetries to spend, the presence of such group actions allows one to apply the powerful techniques of representation theory, Fourier analysis, and ergodic theory. However, as this is a foundational real analysis class, we will not discuss these important topics much here (and in fact will not deal with group actions much at all).

- **Order structure.** In some cases, we want to utilise the notion of a function f being "non-negative", or "dominating" another function g. One might also want to take the "max" or "supremum" of two or more functions in a function space V, or split a function into "positive" and "negative" components. Such order structures interact with the other structures on a space in many useful ways (e.g. via the *Stone-Weierstrass theorem*, Theorem 1.10.24). Much like convexity, order structure is specific to the real line and is another reason why much of real analysis breaks down over other fields. (The complex plane is of course an extension

of the real line and so is able to exploit the order structure of that line, usually by treating the real and imaginary components separately.)

There are of course many ways to combine various flavours of these structures together, and there are entire subfields of mathematics that are devoted to studying particularly common and useful categories of such combinations (e.g. topological vector spaces, normed vector spaces, Banach spaces, Banach algebras, von Neumann algebras, C^* algebras, Frechet spaces, Hilbert spaces, group algebras, etc.). The study of these sorts of spaces is known collectively as functional analysis. We will study some (but certainly not all) of these combinations in an abstract and general setting later in this course, but to begin with we will focus on the L^p spaces, which are very good model examples for many of the above general classes of spaces, and also of importance in many applications of analysis (such as probability or PDE).

1.3.1. L^p **spaces.** In this section, $(X, \mathcal{X}, \mu)$ will be a fixed measure space; notions such as "measurable", "measure", "almost everywhere", etc. will always be with respect to this space, unless otherwise specified. Similarly, unless otherwise specified, all subsets of X mentioned are restricted to be measurable, as are all scalar functions on X.

For sake of concreteness, we shall select the field of scalars to be the complex numbers $\mathbf{C}$. The theory of real Lebesgue spaces is virtually identical to that of complex Lebesgue spaces, and the former can largely be deduced from the latter as a special case.

We already have the notion of an absolutely integrable function on X, which is a function $f : X \to \mathbf{C}$ such that $\int_X |f|\ d\mu$ is finite. More generally, given any[5] exponent $0 < p < \infty$, we can define a p^{th}-power integrable function to be a function $f : X \to \mathbf{C}$ such that $\int_X |f|^p\ d\mu$ is finite.

[5]Besides $p = 1$, the case of most interest is the case of square-integrable functions, when $p = 2$. We will also extend this notion later to $p = \infty$, which is also an important special case.

Remark 1.3.1. One can also extend these notions to functions that take values in the extended complex plane $\mathbf{C} \cup \{\infty\}$, but one easily observes that p^{th} power integrable functions must be finite almost everywhere, and so there is essentially no increase in generality afforded by extending the range in this manner.

Following the "Lebesgue philosophy" that one should ignore whatever is going on on a set of measure zero, let us declare two measurable functions to be equivalent if they agree almost everywhere. This is easily checked to be an equivalence relation, which does not affect the property of being p^{th}-power integrable. Thus, we can define the Lebesgue space $L^p(X, \mathcal{X}, \mu)$ to be the space of p^{th}-power integrable functions, quotiented out by this equivalence relation. Thus, strictly speaking, a typical element of $L^p(X, \mathcal{X}, \mu)$ is not actually a specific function f, but is instead an equivalence class $[f]$, consisting of all functions equivalent to a single function f. However, we shall abuse notation and speak loosely of a function f "belonging" to $L^p(X, \mathcal{X}, \mu)$, where it is understood that f is only defined up to equivalence, or more imprecisely is "defined almost everywhere". For the purposes of integration, this equivalence is quite harmless, but this convention does mean that we can no longer evaluate a function f in $L^p(X, \mathcal{X}, \mu)$ at a single point x if that point x has zero measure. It takes a little bit of getting used to the idea of a function that cannot actually be evaluated at any specific point, but with some practice you will find that it will not cause[6] any significant conceptual difficulty.

Exercise 1.3.1. If $(X, \mathcal{X}, \mu)$ is a measure space, and $\overline{\mathcal{X}}$ is the completion of $\mathcal{X}$, show that the spaces $L^p(X, \mathcal{X}, \mu)$ and $L^p(X, \overline{\mathcal{X}}, \mu)$ are isomorphic using the obvious candidate for the isomorphism. Because of this, when dealing with L^p spaces, we will usually not be too concerned with whether the underlying measure space is complete.

Remark 1.3.2. Depending on which of the three structures $X, \mathcal{X}, \mu$ of the measure space one wishes to emphasise, the space $L^p(X, \mathcal{X}, \mu)$ is often abbreviated $L^p(X)$, $L^p(\mathcal{X})$, $L^p(X, \mu)$, or even just L^p. Since

[6] One could also take a more abstract view, dispensing with the set X altogether and defining the Lebesgue space $L^p(\mathcal{X}, \mu)$ on abstract measure spaces $(\mathcal{X}, \mu)$, but we will not do so here. Another way to think about elements of L^p is that they are functions which are "unreliable" on an unknown set of measure zero, but remain "reliable" almost everywhere.

for this discussion the measure space $(X, \mathcal{X}, \mu)$ will be fixed, we shall usually use the L^p abbreviation in this section. When the space X is discrete (i.e. $\mathcal{X} = 2^X$) and μ is counting measure, then $L^p(X, \mathcal{X}, \mu)$ is usually abbreviated $\ell^p(X)$ or just ℓ^p (and the almost everywhere equivalence relation trivialises and can thus be completely ignored).

At present, the Lebesgue spaces L^p are just sets. We now begin to place several of the structures mentioned in the introduction to upgrade these sets to richer spaces.

We begin with vector space structure. Fix $0 < p < \infty$, and let $f, g \in L^p$ be two p^{th}-power integrable functions. From the crude pointwise (or more precisely, "pointwise almost everywhere") inequality

$$\begin{aligned} |f(x) + g(x)|^p &\leq (2 \max(|f(x)|, |g(x)|))^p \\ &= 2^p \max(|f(x)|^p, |g(x)|^p) \\ &\leq 2^p (|f(x)|^p + |g(x)|^p) \end{aligned} \tag{1.16}$$

we see that the sum of two p^{th}-power integrable functions is also p^{th}-power integrable. It is also easy to see that any scalar multiple of a p^{th}-power integrable function is also p^{th}-power integrable. These operations respect almost everywhere equivalence, and so L^p becomes a (complex) vector space.

Next, we set up the norm structure. If $f \in L^p$, we define the L^p *norm* $\|f\|_{L^p}$ of f to be the number

$$\|f\|_{L^p} := (\int_X |f|^p \, d\mu)^{1/p}; \tag{1.17}$$

this is a finite non-negative number by definition of L^p; in particular, we have the identity

$$\|f^r\|_{L^p} = \|f\|_{L^{pr}}^r \tag{1.18}$$

for all $0 < p, r < \infty$.

The L^p norm has the following three basic properties:

Lemma 1.3.3. *Let $0 < p < \infty$ and $f, g \in L^p$.*

(i) *(Non-degeneracy) $\|f\|_{L^p} = 0$ if and only if $f = 0$.*

(ii) *(Homogeneity)* $\|cf\|_{L^p} = |c|\|f\|_{L^p}$ *for all complex numbers* c.

(iii) *((Quasi-)triangle inequality) We have* $\|f+g\|_{L^p} \leq C(\|f\|_{L^p} + \|g\|_{L^p})$ *for some constant* C *depending on* p. *If* $p \geq 1$, *then we can take* $C = 1$ *(this fact is also known as* Minkowski's inequality*).*

Proof. The claims (i), (ii) are obvious. (Note how important it is that we equate functions that vanish almost everywhere in order to get (i).) The quasi-triangle inequality follows from a variant of the estimates in (1.16) and is left as an exercise. For the triangle inequality, we have to be more efficient than the crude estimate (1.16). By the non-degeneracy property we may take $\|f\|_{L^p}$ and $\|g\|_{L^p}$ to be non-zero. Using the homogeneity, we can normalise $\|f\|_{L^p} + \|g\|_{L^p}$ to equal 1, thus (by homogeneity again) we can write $f = (1-\theta)F$ and $g = \theta G$ for some $0 < \theta < 1$ and $F, G \in L^p$ with $\|F\|_{L^p} = \|G\|_{L^p} = 1$. Our task is now to show that

$$\int_X |(1-\theta)F(x) + \theta G(x)|^p \, d\mu \leq 1. \tag{1.19}$$

But observe that for $1 \leq p < \infty$, the function $x \mapsto |x|^p$ is convex on $\mathbf{C}$, and in particular that

$$|(1-\theta)F(x) + \theta G(x)|^p \leq (1-\theta)|F(x)|^p + \theta|G(x)|^p. \tag{1.20}$$

(If one wishes, one can use the complex triangle inequality to first reduce to the case when F, G are non-negative, in which case one only needs convexity on $[0, +\infty)$ rather than all of $\mathbf{C}$.) The claim (1.19) then follows from (1.20) and the normalisations of F, G. □

Exercise 1.3.2. Let $0 < p \leq 1$ and $f, g \in L^p$.

(i) Establish the variant $\|f+g\|_{L^p}^p \leq \|f\|_{L^p}^p + \|g\|_{L^p}^p$ of the triangle inequality.

(ii) If furthermore f and g are non-negative (almost everywhere), establish also the reverse triangle inequality $\|f+g\|_{L^p} \geq \|f\|_{L^p} + \|g\|_{L^p}$.

(iii) Show that the best constant C in the quasi-triangle inequality is $2^{\frac{1}{p}-1}$. In particular, the triangle inequality is false for $p < 1$.

(iv) Now suppose instead that $1 < p < \infty$ or $0 < p < 1$. If $f, g \in L^p$ are such that $\|f+g\|_{L^p} = \|f\|_{L^p} + \|g\|_{L^p}$, show that one of the functions f, g is a non-negative scalar multiple of the other (up to equivalence, of course). What happens when $p = 1$?

A vector space V with a function $\|\| : V \to [0, +\infty)$ obeying the non-degeneracy, homogeneity, and (quasi-)triangle inequality is known as a (quasi-)normed vector space, and the function $f \mapsto \|f\|$ is then known as a (quasi-)norm; thus L^p is a normed vector space for $1 \leq p < \infty$ but only a quasi-normed vector space for $0 < p < 1$. A function $\|\| : V \to [0, +\infty)$ obeying the homogeneity and triangle inequality, but not necessarily the non-degeneracy property, is known as a seminorm; thus for instance the L^p norms for $1 \leq p < \infty$ would have been seminorms if we did not equate functions that agreed almost everywhere. (Conversely, given a seminormed vector space $(V, \|\|)$, one can convert it into a normed vector space by quotienting out the subspace $\{f \in V : \|f\| = 0\}$; we leave the details as an exercise for the reader.)

Exercise 1.3.3. Let $\|\| : V \to [0, +\infty)$ be a function on a vector space which obeys the non-degeneracy and homogeneity properties. Show that $\|\|$ is a norm if and only if the closed unit ball $\{x : \|x\| \leq 1\}$ is convex; show that the same equivalence also holds for the open unit ball. This fact emphasises the geometric nature of the triangle inequality.

Exercise 1.3.4. If $f \in L^p$ for some $0 < p < \infty$, show that the support $\{x \in X : f(x) \neq 0\}$ of f (which is defined only up to sets of measure zero) is a σ-finite set. (Because of this, we can often reduce from the non-σ-finite case to the σ-finite case in many, though not all, questions concerning L^p spaces.)

We now are able to define L^p norms and spaces in the limit $p = \infty$. We say that a function $f : X \to \mathbf{C}$ is *essentially bounded* if there exists an M such that $|f(x)| \leq M$ for almost every $x \in X$, and define $\|f\|_{L^\infty}$ to be the least M that serves as such a bound. We let L^∞ denote the space of essentially bounded functions, quotiented out by equivalence, and given the norm $\|\cdot\|_{L^\infty}$. It is not hard to see that this is also a

normed vector space. Observe that a sequence $f_n \in L^\infty$ converges to a limit $f \in L^\infty$ if and only if f_n converges essentially uniformly to f, i.e. it converges uniformly to f outside of a set of measure zero. (Compare with Egorov's theorem (Theorem 1.1.21), which equates pointwise convergence with uniform convergence outside of a set of arbitrarily small measure.)

Now we explain why we call this norm the L^∞ norm:

Example 1.3.4. Let f be a (generalised) step function, thus $f = A1_E$ for some amplitude $A > 0$ and some set E; let us assume that E has positive finite measure. Then $\|f\|_{L^p} = A\mu(E)^{1/p}$ for all $0 < p < \infty$, and also $\|f\|_{L^\infty} = A$. Thus in this case, at least, the L^∞ norm is the limit of the L^p norms. This example illustrates also that the L^p norms behave like combinations of the "height" A of a function, and the "width" $\mu(E)$ of such a function, though of course the concepts of height and width are not formally defined for functions that are not step functions.

Exercise 1.3.5.

- If $f \in L^\infty \cap L^{p_0}$ for some $0 < p_0 < \infty$, show that $\|f\|_{L^p} \to \|f\|_{L^\infty}$ as $p \to \infty$. (*Hint*: use the monotone convergence theorem, Theorem 1.1.21.)
- If $f \notin L^\infty$, show that $\|f\|_{L^p} \to \infty$ as $p \to \infty$.

Once one has a vector space structure and a (quasi-)norm structure, we immediately get a (quasi-)metric structure:

Exercise 1.3.6. Let $(V, \|\|)$ be a normed vector space. Show that the function $d : V \times V \to [0, +\infty)$ defined by $d(f, g) := \|f - g\|$ is a metric on V which is translation-invariant (thus $d(f + h, g + h) = d(f, g)$ for all $f, g \in V$) and homogeneous (thus $d(cf, cg) = |c|d(f, g)$ for all $f, g \in V$ and scalars c). Conversely, show that every translation-invariant homogeneous metric on V arises from precisely one norm in this manner. Establish a similar claim relating quasi-norms with quasi-metrics (which are defined as metrics, but with the triangle inequality replaced by a quasi-triangle inequality), or between seminorms and semimetrics (which are defined as metrics, but where distinct points are allowed to have a zero separation; these are also known as *pseudometrics*).

The (quasi-)metric structure in turn generates a topological structure in the usual manner using the (quasi-)metric balls as a base for the topology. In particular, a sequence of functions $f_n \in L^p$ converges to a limit $f \in L^p$ if $\|f_n - f\|_{L^p} \to 0$ as $n \to \infty$. We refer to this type of convergence as convergence in L^p norm, or strong convergence in L^p (we will discuss other modes of convergence in later lectures). As is usual in (quasi-)metric spaces (or more generally for Hausdorff spaces), the limit, if it exists, is unique. (This is however not the case for topological structures induced by seminorms or semimetrics, though we can solve this problem by quotienting out the degenerate elements as discussed earlier.)

Recall that any series $\sum_{n=1}^\infty a_n$ of scalars is convergent if it is absolutely convergent (i.e. if $\sum_{n=1}^\infty |a_n| < \infty$. This fact turns out to be closely related to the fact that the field of scalars $\mathbf{C}$ is complete. This can be seen from the following result:

Exercise 1.3.7. Let $(V, \|\|)$ be a normed vector space (and hence also a metric space and a topological space). Show that the following are equivalent:

- V is a complete metric space (i.e. every Cauchy sequence converges).
- Every sequence $f_n \in V$ which is absolutely convergent (i.e. $\sum_{n=1}^\infty \|f_n\| < \infty$), is also conditionally convergent (i.e. $\sum_{n=1}^N f_n$ converges to a limit as $N \to \infty$.

Remark 1.3.5. The situation is more complicated for complete quasi-normed vector spaces; not every absolutely convergent series is conditionally convergent. On the other hand, if $\|f_n\|$ decays faster than a sufficiently large negative power of n, one recovers conditional convergence; see [**Ta**].

Remark 1.3.6. Let X be a topological space, and let $BC(X)$ be the space of bounded continuous functions on X; this is a vector space. We can place the uniform norm $\|f\|_u := \sup_{x \in X} |f(x)|$ on this space; this makes $BC(X)$ into a normed vector space. It is not hard to verify that this space is complete, and so every absolutely convergent series in $BC(X)$ is conditionally convergent. This fact is better known as the *Weierstrass M-test*.

A space obeying the properties in Exercise 1.3.5 (i.e. a complete normed vector space) is known as a *Banach space*. We will study Banach spaces in more detail later in this course. For now, we give one of the fundamental examples of Banach spaces.

Proposition 1.3.7. *L^p is a Banach space for every $1 \leq p \leq \infty$.*

Proof. By Exercise 1.3.7, it suffices to show that any series $\sum_{n=1}^\infty f_n$ of functions in L^p which is absolutely convergent, is also conditionally convergent. This is easy in the case $p = \infty$ and is left as an exercise. In the case $1 \leq p < \infty$, we write $M := \sum_{n=1}^\infty \|f_n\|_{L^p}$, which is a finite quantity by hypothesis. By the triangle inequality, we have $\| \sum_{n=1}^N |f_n| \|_{L^p} \leq M$ for all N. By monotone convergence (Theorem 1.1.21), we conclude $\| \sum_{n=1}^\infty |f_n| \|_{L^p} \leq M$. In particular, $\sum_{n=1}^\infty f_n(x)$ is absolutely convergent for almost every x. Write the limit of this series as $F(x)$. By dominated convergence (Theorem 1.1.21), we see that $\sum_{n=1}^N f_n(x)$ converges in L^p norm to F, and we are done. □

An important fact is that functions in L^p can be approximated by simple functions:

Proposition 1.3.8. *If $0 < p < \infty$, then the space of simple functions with finite measure support is a dense subspace of L^p.*

Remark 1.3.9. The concept of a non-trivial dense subspace is one which only comes up in infinite dimensions, and is hard to visualise directly. Very roughly speaking, the infinite number of degrees of freedom in an infinite dimensional space gives a subspace an infinite number of "opportunities" to come as close as one desires to any given point in that space, which is what allows such spaces to be dense.

Proof. The only non-trivial thing to show is the density. An application of the monotone convergence theorem (Theorem 1.1.21) shows that the space of bounded L^p functions are dense in L^p. Another application of monotone convergence (and Exercise 1.3.4) then shows that the space bounded L^p functions of finite measure support are dense in the space of bounded L^p functions. Finally, by discretising the range of bounded L^p functions, we see that the space of simple functions with finite measure support is dense in the space of bounded L^p functions with finite support. □

Remark 1.3.10. Since not every function in L^p is a simple function with finite measure support, we thus see that the space of simple functions with finite measure support with the L^p norm is an example of a normed vector space which is not complete.

Exercise 1.3.8. Show that the space of simple functions (not necessarily with finite measure support) is a dense subspace of L^∞. Is the same true if one reinstates the finite measure support restriction?

Exercise 1.3.9. Suppose that μ is σ-finite and $\mathcal{X}$ is separable (i.e. countably generated). Show that L^p is separable (i.e. has a countable dense subset) for all $1 \leq p < \infty$. Give a counterexample that shows that L^∞ need not be separable. (*Hint*: try using counting measure.)

Next, we turn to algebra properties of L^p spaces. The key fact here is

Proposition 1.3.11 (Hölder's inequality)**.** *Let $f \in L^p$ and $g \in L^q$ for some $0 < p, q \leq \infty$. Then $fg \in L^r$ and $\|fg\|_{L^r} \leq \|f\|_{L^p}\|g\|_{L^q}$, where the exponent r is defined by the formula $\frac{1}{r} = \frac{1}{p} + \frac{1}{q}$.*

Proof. This will be a variant of the proof of the triangle inequality in Lemma 1.3.3, again relying ultimately on convexity. The claim is easy when $p = \infty$ or $q = \infty$ and is left as an exercise for the reader in this case, so we assume $p, q < \infty$. Raising f and g to the power r using (1.17) we may assume $r = 1$, which makes $1 < p, q < \infty$ dual exponents in the sense that $\frac{1}{p} + \frac{1}{q} = 1$. The claim is obvious if either $\|f\|_{L^p}$ or $\|g\|_{L^q}$ are zero, so we may assume they are non-zero; by homogeneity we may then normalise $\|f\|_{L^p} = \|g\|_{L^q} = 1$. Our task is now to show that

$$\int_X |fg| \, d\mu \leq 1. \tag{1.21}$$

Here, we use the convexity of the exponential function $t \mapsto e^t$ on $[0, +\infty)$, which implies the convexity of the function $t \mapsto |f(x)|^{p(1-t)}|g(x)|^{qt}$ for $t \in [0, 1]$ for any x. In particular we have

$$|f(x)g(x)| \leq \frac{1}{p}|f(x)|^p + \frac{1}{q}|g(x)|^q \tag{1.22}$$

and the claim (1.21) follows from the normalisations on p, q, f, g. $\square$

Remark 1.3.12. For a different proof of this inequality (based on the *tensor power trick*), see Section 1.9 of *Structure and Randomness*.

Remark 1.3.13. One can also use Hölder's inequality to prove the triangle inequality for L^p, $1 \leq p < \infty$ (i.e. *Minkowski's inequality*). From the complex triangle inequality $|f+g| \leq |f| + |g|$, it suffices to check the case when f, g are non-negative. In this case we have the identity

$$\|f+g\|_{L^p}^p = \|f|f+g|^{p-1}\|_{L^1} + \|g|f+g|^{p-1}\|_{L^1} \tag{1.23}$$

while Hölder's inequality gives $\|f|f+g|^{p-1}\|_{L^1} \leq \|f\|_{L^p}\|f+g\|_{L^p}^{p-1}$ and $\|g|f+g|^{p-1}\|_{L^1} \leq \|g\|_{L^p}\|f+g\|_{L^p}^{p-1}$. The claim then follows from some algebra (and checking the degenerate cases separately, e.g. when $\|f+g\|_{L^p} = 0$).

Remark 1.3.14. The proofs of Hölder's inequality and Minkowski's inequality both relied on convexity of various functions in $\mathbf{C}$ or $[0, +\infty)$. One way to emphasise this is to deduce both inequalities from *Jensen's inequality*, which is an inequality which manifestly exploits this convexity. We will not take this approach here, but see for instance [**LiLo2000**] for a discussion.

Example 1.3.15. It is instructive to test Hölder's inequality (and also Exercises 1.3.10-1.3.14 below) in the special case when f, g are generalised step functions, say $f = A1_E$ and $g = B1_F$ with A, B non-zero. The inequality then simplifies to

$$\mu(E \cap F)^{1/r} \leq \mu(E)^{1/p}\mu(F)^{1/q} \tag{1.24}$$

which can be easily deduced from the hypothesis $\frac{1}{p} + \frac{1}{q} = \frac{1}{r}$ and the trivial inequalities $\mu(E \cap F) \leq \mu(E)$ and $\mu(E \cap F) \leq \mu(F)$. One then easily sees (when p, q are finite) that equality in (1.24) only holds if $\mu(E \cap F) = \mu(E) = \mu(F)$, or in other words if E and F agree almost everywhere. Note the above computations also explain why the condition $\frac{1}{p} + \frac{1}{q} = \frac{1}{r}$ is necessary.

Exercise 1.3.10. Let $0 < p, q < \infty$, and let $f \in L^p, g \in L^q$ be such that Hölder's inequality is obeyed with equality. Show that of the functions f^p, g^q, one of them is a scalar multiple of the other (up to equivalence, of course). What happens if p or q is infinite?

An important corollary of Hölder's inequality is the *Cauchy-Schwarz inequality*

$$|\int_X f(x)\overline{g(x)}\,d\mu| \leq \|f\|_{L^2}\|g\|_{L^2} \tag{1.25}$$

which can of course be proven by many other means.

Exercise 1.3.11. If $f \in L^p$ for some $0 < p \leq \infty$, and is also supported on a set E of finite measure, show that $f \in L^q$ for all $0 < q \leq p$, with $\|f\|_{L^q} \leq \mu(E)^{\frac{1}{q}-\frac{1}{p}}\|f\|_{L^p}$. When does equality occur?

Exercise 1.3.12. If $f \in L^p$ for some $0 < p < \infty$, and every set of positive measure in X has measure at least m, show that $f \in L^q$ for all $p < q \leq \infty$, with $\|f\|_{L^q} \leq m^{\frac{1}{q}-\frac{1}{p}}\|f\|_{L^p}$. When does equality occur? (This result is especially useful for the ℓ^p spaces, in which μ is counting measure and m can be taken to be 1.)

Exercise 1.3.13. If $f \in L^{p_0} \cap L^{p_1}$ for some $0 < p_0 < p_1 \leq \infty$, show that $f \in L^p$ for all $p_0 \leq p \leq p_1$, and that $\|f\|_{L^p} \leq \|f\|_{L^{p_0}}^{1-\theta}\|f\|_{L^{p_1}}^{\theta}$, where $0 < \theta < 1$ is such that $\frac{1}{p} = \frac{1-\theta}{p_0} + \frac{\theta}{p_1}$. Another way of saying this is that the function $\frac{1}{p} \mapsto \log\|f\|_{L^p}$ is convex. When does equality occur? This convexity is a prototypical example of *interpolation*, about which we shall say more in Section 1.11.

Exercise 1.3.14. If $f \in L^{p_0}$ for some $0 < p_0 \leq \infty$, and its support $E := \{x \in X : f(x) \neq 0\}$ has finite measure, show that $f \in L^p$ for all $0 < p < p_0$, and that $\|f\|_{L^p}^p \to \mu(E)$ as $p \to 0$. (Because of this, the measure of the support of f is sometimes known as the L^0 *norm* of f, or more precisely the L^0 norm raised to the power 0.)

1.3.2. Linear functionals on L^p. Given an exponent $1 \leq p \leq \infty$, define the dual exponent $1 \leq p' \leq \infty$ by the formula $\frac{1}{p} + \frac{1}{p'} = 1$ (thus $p' = p/(p-1)$ for $1 < p < \infty$, while 1 and ∞ are duals of each other). From Hölder's inequality, we see that for any $g \in L^{p'}$, the functional $\lambda_g : L^p \to \mathbf{C}$ defined by

$$\lambda_g(f) := \int_X f\overline{g}\,d\mu \tag{1.26}$$

is well-defined on L^p; the functional is also clearly linear. Furthermore, Hölder's inequality also tells us that this functional is continuous.

A deep and important fact about L^p spaces is that, in most cases, the converse is true: the recipe (1.26) is the *only* way to create continuous linear functionals on L^p.

Theorem 1.3.16 (Dual of L^p)**.** *Let $1 \leq p < \infty$, and assume μ is σ-finite. Let $\lambda : L^p \to \mathbf{C}$ be a continuous linear functional. Then there exists a unique $g \in L^{p'}$ such that $\lambda = \lambda_g$.*

This result should be compared with the Radon-Nikodym theorem (Corollary 1.2.5). Both theorems start with an abstract function $\mu : \mathcal{X} \to \mathbf{R}$ or $\lambda : L^p \to \mathbf{C}$, and create a function out of it. Indeed, we shall see shortly that the two theorems are essentially equivalent to each other. We will develop Theorem 1.3.16 further in Section 1.5, once we introduce the notion of a dual space.

To prove Theorem 1.3.16, we first need a simple and useful lemma:

Lemma 1.3.17 (Continuity is equivalent to boundedness for linear operators)**.** *Let $T : X \to Y$ be a linear transformation from one normed vector space $(X, \|\|_X)$ to another $(Y, \|\|_Y)$. Then the following are equivalent:*

(i) *T is continuous.*

(ii) *T is continuous at 0.*

(iii) *There exists a constant C such that $\|Tx\|_Y \leq C\|x\|_X$ for all $x \in X$.*

Proof. It is clear that (i) implies (ii), and that (iii) implies (ii). Next, from linearity we have $Tx = Tx_0 + T(x - x_0)$ for any $x, x_0 \in X$, which (together with the continuity of addition, which follows from the triangle inequality) shows that continuity of T at 0 implies continuity of T at any x_0, so that (ii) implies (i). The only remaining task is to show that (i) implies (iii). By continuity, the inverse image of the unit ball in Y must be an open neighbourhood of 0 in X, thus there exists some radius $r > 0$ such that $\|Tx\|_Y < 1$ whenever $\|x\|_X < r$. The claim then follows (with $C := 1/r$) by homogeneity. (Alternatively, one can deduce (iii) from (ii) by contradiction. If (iii) failed, then there exists a sequence x_n of non-zero elements of X such that $\|Tx_n\|_Y / \|x_n\|_X$ goes to infinity. By homogeneity, we can arrange

matters so that $\|x_n\|_X$ goes to zero, but $\|Tx_n\|_Y$ stays away from zero, thus contradicting continuity at 0.) □

Proof of Theorem 1.3.16. The uniqueness claim is similar to the uniqueness claim in the Radon-Nikodym theorem (Exercise 1.2.2) and is left as an exercise to the reader; the hard part is establishing existence.

Let us first consider the case when μ is finite. The linear functional $\lambda : L^p \to \mathbf{C}$ induces a functional $\nu : \mathcal{X} \to \mathbf{C}$ on sets E by the formula

$$\nu(E) := \lambda(1_E). \tag{1.27}$$

Since λ is linear, ν is finitely additive (and sends the empty set to zero). Also, if $E_1, E_2, \ldots$ are a sequence of disjoint sets, then $1_{\bigcup_{n=1}^N E_n}$ converges in L^p to $1_{\bigcup_{n=1}^\infty E_n}$ as $n \to \infty$ (by the dominated convergence theorem and the finiteness of μ), and thus (by continuity of λ and finite additivity of ν), ν is countably additive as well. Finally, from (1.27) we also see that $\nu(E) = 0$ whenever $\mu(E) = 0$, thus ν is absolutely continuous with respect to μ. Applying the Radon-Nikodym theorem (Corollary 1.2.5) to both the real and imaginary components of ν, we conclude that $\nu = \mu_g$ for some $g \in L^1$; thus by (1.27) we have

$$\lambda(1_E) = \lambda_g(1_E) \tag{1.28}$$

for all measurable E. By linearity, this implies that λ and λ_g agree on simple functions. Taking uniform limits (using Exercise 1.3.8) and using continuity (and the finite measure of μ) we conclude that λ and λ_g agree on all bounded functions. Taking monotone limits (working on the positive and negative supports of the real and imaginary parts of g separately) we conclude that λ and λ_g agree on all functions in L^p, and in particular that $\int_X f\overline{g}\, d\mu$ is absolutely convergent for all $f \in L^p$.

To finish the theorem in this case, we need to establish that g lies in $L^{p'}$. By taking real and imaginary parts we may assume without loss of generality that g is real; by splitting into the regions where g is positive and negative we may assume that g is non-negative.

We already know that $\lambda_g = \lambda$ is a continuous functional from L^p to $\mathbf{C}$. By Lemma 1.3.17, this implies a bound of the form $|\lambda_g(f)| \leq C\|f\|_{L^p}$ for some $C > 0$.

Suppose first that $p > 1$. Heuristically, we would like to test this inequality with $f := g^{p'-1}$, since we formally have $\lambda_g(f) = \|g\|_{L^{p'}}^{p'}$ and $\|f\|_{L^p} = \|g\|_{L^{p'}}^{p'-1}$. (Not coincidentally, this is also the choice that would make Hölder's inequality an equality, see Exercise 1.3.10.) Cancelling the $\|g\|_{L^{p'}}$ factors would then give the desired finiteness of $\|g\|_{L^{p'}}$.

We can't quite make that argument work, because it is circular: it assumes $\|g\|_{L^{p'}}$ is finite in order to show that $\|g\|_{L^{p'}}$ is finite! But this can be easily remedied. We test the inequality with $f_N := \min(g, N)^{p'-1}$ for some large N; this lies in L^p. We have $\lambda_g(f) \geq \|\min(g, N)\|_{L^{p'}}^{p'}$ and $\|f_N\|_{L^p} = \|\min(g, N)\|_{L^{p'}}^{p'-1}$, and hence $\|\min(g, N)\|_{L^{p'}} \leq C$ for all N. Letting N go to infinity and using monotone convergence (Theorem 1.1.21), we obtain the claim.

In the $p = 1$ case, we instead use $f := 1_{g>N}$ as the test functions, to conclude that g is bounded almost everywhere by N; we leave the details to the reader.

This handles the case when μ is finite. When μ is σ-finite, we can write X as the union of an increasing sequence E_n of sets of finite measure. On each such set, the above arguments let us write $\lambda = \lambda_{g_n}$ for some $g_n \in L^{p'}(E_n)$. The uniqueness arguments tell us that the g_n are all compatible with each other, in particular if $n < m$, then g_n and g_m agree on E_n. Thus all the g_n are in fact restrictions of a single function g to E_n. The previous arguments also tell us that the $L^{p'}$ norm of g_n is bounded by the same constant C uniformly in n, so by monotone convergence (Theorem 1.1.21), g has bounded $L^{p'}$ norm also, and we are done. □

Remark 1.3.18. When $1 < p < \infty$, the hypothesis that μ is σ-finite can be dropped, but not when $p = 1$; see e.g. [**Fo2000**, Section 6.2] for further discussion. In these lectures, though, we will be content with working in the σ-finite setting. On the other hand, the claim fails when $p = \infty$ (except when X is finite); we will see this in Section 1.5, when we discuss the Hahn-Banach theorem.

Remark 1.3.19. We have seen how the Lebesgue-Radon-Nikodym theorem can be used to establish Theorem 1.3.16. The converse is also true: Theorem 1.3.16 can be used to deduce the Lebesgue-Radon-Nikodym theorem (a fact essentially observed by von Neumann). For simplicity, let us restrict attention to the unsigned finite case, thus μ and m are unsigned and finite. This implies that the sum $\mu + m$ is also unsigned and finite. We observe that the linear functional $\lambda : f \mapsto \int_X f \, d\mu$ is continuous on $L^1(\mu + m)$, hence by Theorem 1.3.16, there must exist a function $g \in L^\infty(\mu + m)$ such that

$$\int_X f \, d\mu = \int_X f\overline{g} \, d(\mu + m) \tag{1.29}$$

for all $f \in L^1(\mu + m)$. It is easy to see that g must be real and non-negative, and also at most 1 almost everywhere. If E is the set where $m = 1$, we see by setting $f = 1_E$ in (1.29) that E has m-measure zero, and so $\mu \restriction_E$ is singular. Outside of E, we see from (1.29) and some rearrangement that

$$\int_{X \backslash E} (1-g) f \, d\mu = \int_X fg \, dm \tag{1.30}$$

and one then easily verifies that μ agrees with $m_{\frac{g}{1-g}}$ outside of E'. This gives the desired Lebesgue-Radon-Nikodym decomposition $\mu = m_{\frac{g}{1-g}} + \mu \restriction_E$.

Remark 1.3.20. The argument used in Remark 1.3.19 also shows that the Radon-Nikodym theorem implies the Lebesgue-Radon-Nikodym theorem.

Remark 1.3.21. One can given an alternate proof of Theorem 1.3.16, which relies on the geometry (and in particular, the uniform convexity) of L^p spaces rather than on the Radon-Nikodym theorem, and can thus be viewed as giving an independent proof of that theorem; see Exercise 1.4.14.

Notes. This lecture first appeared at `terrytao.wordpress.com/2009/01/09`. Thanks to Xiaochuan Li for corrections.

1.4. Hilbert spaces

In the next few lectures, we will be studying four major classes of function spaces. In decreasing order of generality, these classes are the *topological vector spaces*, the *normed vector spaces*, the *Banach spaces*, and the *Hilbert spaces*. In order to motivate the discussion of the more general classes of spaces, we will first focus on the most special class - that of (real and complex) Hilbert spaces. These spaces can be viewed as generalisations of (real and complex) Euclidean spaces such as $\mathbf{R}^n$ and $\mathbf{C}^n$ to infinite-dimensional settings, and indeed much of one's Euclidean geometry intuition concerning lengths, angles, orthogonality, subspaces, etc. will transfer readily to arbitrary Hilbert spaces; in contrast, this intuition is not always accurate in the more general vector spaces mentioned above. In addition to Euclidean spaces, another fundamental example[7] of Hilbert spaces comes from the Lebesgue spaces $L^2(X, \mathcal{X}, \mu)$ of a measure space $(X, \mathcal{X}, \mu)$.

Hilbert spaces are the natural abstract framework in which to study two important (and closely related) concepts: orthogonality and unitarity, allowing us to generalise familiar concepts and facts from Euclidean geometry such as the Cartesian coordinate system, rotations and reflections, and the Pythagorean theorem to Hilbert spaces. (For instance, the Fourier transform (Section 1.12) is a unitary transformation and can thus be viewed as a kind of generalised rotation.) Furthermore, the *Hodge duality* on Euclidean spaces has a partial analogue for Hilbert spaces, namely the *Riesz representation theorem* for Hilbert spaces, which makes the theory of duality and adjoints for Hilbert spaces especially simple (when compared with the more subtle theory of duality for, say, Banach spaces; see Section 1.5).

These notes are only the most basic introduction to the theory of Hilbert spaces. In particular, the theory of linear transformations

[7] There are of course many other Hilbert spaces of importance in complex analysis, harmonic analysis, and PDE, such as *Hardy spaces* $\mathcal{H}^2$, *Sobolev spaces* $H^s = W^{s,2}$, and the space HS of *Hilbert-Schmidt operators*; see for instance Section 1.14 for a discussion of Sobolev spaces. Complex Hilbert spaces also play a fundamental role in the foundations of quantum mechanics, being the natural space to hold all the possible states of a quantum system (possibly after projectivising the Hilbert space), but we will not discuss this subject here.

between two Hilbert spaces, which is perhaps the most important aspect of the subject, is not covered much at all here.

1.4.1. Inner product spaces. The Euclidean norm

$$|(x_1, \ldots, x_n)| := \sqrt{x_1^2 + \ldots + x_n^2} \tag{1.31}$$

in real Euclidean space $\mathbf{R}^n$ can be expressed in terms of the *dot product* $\cdot : \mathbf{R}^n \times \mathbf{R}^n \to \mathbf{R}$, defined as

$$(x_1, \ldots, x_n) \cdot (y_1, \ldots, y_n) := x_1 y_1 + \ldots + x_n y_n \tag{1.32}$$

by the well-known formula

$$|x| = (x \cdot x)^{1/2}. \tag{1.33}$$

In particular, we have the positivity property

$$x \cdot x \geq 0 \tag{1.34}$$

with equality if and only if $x = 0$. One reason why it is more advantageous to work with the dot product than the norm is that while the norm function is only sublinear, the dot product is *bilinear*, thus

$$(cx+dy)\cdot z = c(x\cdot z)+d(y\cdot z); \quad z\cdot(cx+dy) = c(z\cdot x)+d(z\cdot y) \tag{1.35}$$

for all vectors x, y and scalars c, d, and also symmetric,

$$x \cdot y = y \cdot x. \tag{1.36}$$

These properties make the inner product easier to manipulate algebraically than the norm.

The above discussion was for the real vector space $\mathbf{R}^n$, but one can develop analogous statements for the complex vector space $\mathbf{C}^n$, in which the norm

$$\|(z_1, \ldots, z_n)\| := \sqrt{|z_1|^2 + \ldots + |z_n|^2} \tag{1.37}$$

can be represented in terms of the complex inner product $\langle , \rangle : \mathbf{C}^n \times \mathbf{C}^n \to \mathbf{C}$ defined by the formula

$$(z_1, \ldots, z_n) \cdot (w_1, \ldots, w_n) := z_1 \overline{w_1} + \ldots + z_n \overline{w_n} \tag{1.38}$$

by the analogue of (1.33), namely

$$\|x\| = (\langle x, x \rangle)^{1/2}. \tag{1.39}$$

In particular, as before with (1.34), we have the positivity property

$$\langle x, x\rangle \geq 0, \tag{1.40}$$

with equality if and only if $x = 0$. The bilinearity property (1.35) is modified to the *sesquilinearity* property

$$\langle cx+dy, z\rangle = c\langle x, z\rangle + d\langle y, z\rangle; \quad \langle z, cx+dy\rangle = \overline{c}\langle z, x\rangle + \overline{d}\langle z, y\rangle \tag{1.41}$$

while the symmetry property (1.36) needs to be replaced with

$$\langle x, y\rangle = \overline{\langle y, x\rangle} \tag{1.42}$$

in order to be compatible with sesquilinearity.

We can formalise all these properties axiomatically as follows.

Definition 1.4.1 (Inner product space)**.** A *complex inner product space* $(V, \langle,\rangle)$ is a complex vector space V, together with an inner product $\langle,\rangle : V \times V \to \mathbf{C}$ which is sesquilinear (i.e. (1.41) holds for all $x, y \in V$ and $c, d \in \mathbf{C}$) and symmetric in the sesquilinear sense (i.e. (1.42) holds for all $x, y \in V$), and obeys the positivity property (1.40) for all $x \in V$, with equality if and only if $x = 0$. We will usually abbreviate $(V, \langle,\rangle)$ as V.

A real inner product space is defined similarly, but with all references to $\mathbf{C}$ replaced by $\mathbf{R}$ (and all references to complex conjugation dropped).

Example 1.4.2. $\mathbf{R}^n$ with the standard dot product (1.32) is a real inner product space, and $\mathbf{C}^n$ with the complex inner product (1.38) is a complex inner product space.

Example 1.4.3. If $(X, \mathcal{X}, \mu)$ is a measure space, then the complex L^2 space $L^2(X, \mathcal{X}, \mu) = L^2(X, \mathcal{X}, \mu; \mathbf{C})$ with the complex inner product

$$\langle f, g\rangle := \int_X f\overline{g}\, d\mu \tag{1.43}$$

(which is well defined, by the Cauchy-Schwarz inequality) is easily verified to be a complex inner product space, and similarly for the real L^2 space (with the complex conjugate signs dropped, of course. Note that the finite dimensional examples $\mathbf{R}^n, \mathbf{C}^n$ can be viewed as the special case of the L^2 examples in which X is $\{1, \ldots, n\}$ with the discrete σ-algebra and counting measure.

Example 1.4.4. Any subspace of a (real or complex) inner product space is again a (real or complex) inner product space, simply by restricting the inner product to the subspace.

Example 1.4.5. Also, any real inner product space V can be *complexified* into the complex inner product space $V_{\mathbf{C}}$, defined as the space of formal combinations $x + iy$ of vectors $x, y \in V$ (with the obvious complex vector space structure), and with inner product

$$\langle a + ib, c + id \rangle := \langle a, c \rangle + i\langle b, c \rangle - i\langle a, d \rangle + \langle b, d \rangle. \tag{1.44}$$

Example 1.4.6. Fix a probability space $(X, \mathcal{X}, \mu)$. The space of square-integrable real-valued random variables of mean zero is an inner product space if one uses covariance as the inner product. (What goes wrong if one drops the mean zero assumption?)

Given a (real or complex) inner product space V, we can define the *norm* $\|x\|$ of any vector $x \in V$ by the formula (1.39), which is well defined thanks to the positivity property; in the case of the L^2 spaces, this norm of course corresponds to the usual L^2 norm. We have the following basic facts:

Lemma 1.4.7. *Let V be a real or complex inner product space.*

(i) *(Cauchy-Schwarz inequality) For any $x, y \in V$, we have $|\langle x, y \rangle| \leq \|x\|\|y\|$.*

(ii) *The function $x \mapsto \|x\|$ is a norm on V. (Thus every inner product space is a normed vector space.)*

Proof. We shall just verify the complex case, as the real case is similar (and slightly easier). The positivity property tells us that the quadratic form $\langle ax + by, ax + by \rangle$ is non-negative for all complex numbers a, b. Using sesquilinearity and symmetry, we can expand this form as

$$|a|^2\|x\|^2 + 2\,\mathrm{Re}(a\bar{b}\langle x, y \rangle) + |b|\|y\|^2. \tag{1.45}$$

Optimising in a, b (see also Section 1.10 of *Structure and Randomness*) we obtain the Cauchy-Schwarz inequality. To verify the norm property, the only non-trivial verification is that of the triangle inequality $\|x+y\| \leq \|x\|+\|y\|$. But on expanding $\|x+y\|^2 = \langle x+y, x+y \rangle$

we see that

$$\|x+y\|^2 = \|x\|^2 + 2\,\mathrm{Re}(\langle x, y\rangle) + \|y\|^2 \tag{1.46}$$

and the claim then follows from the Cauchy-Schwarz inequality. □

Observe from the Cauchy-Schwarz inequality that the inner product $\langle , \rangle : H \times H \to \mathbf{C}$ is continuous.

Exercise 1.4.1. Let $T : V \to W$ be a linear map from one (real or complex) inner product space to another. Show that T preserves the inner product structure (i.e. $\langle Tx, Ty\rangle = \langle x, y\rangle$ for all $x, y \in V$) if and only if T is an isometry (i.e. $\|Tx\| = \|x\|$ for all $x \in V$). (*Hint*: in the real case, express $\langle x, y\rangle$ in terms of $\|x+y\|^2$ and $\|x-y\|^2$. In the complex case, use $x+y, x-y, x+iy, x-iy$ instead of $x+y, x-y$.)

Inspired by the above exercise, we say that two inner product spaces are *isomorphic* if there exists an invertible isometry from one space to the other; such invertible isometries are known as *isomorphisms*.

Exercise 1.4.2. Let V be a real or complex inner product space. If $x_1, \ldots, x_n$ are a finite collection of vectors in V, show that the *Gram matrix* $(\langle x_i, x_j\rangle)_{1\leq i,j\leq n}$ is Hermitian and positive semi-definite, and is positive definite if and only if the $x_1, \ldots, x_n$ are linearly independent. Conversely, given a Hermitian positive semi-definite matrix $(a_{ij})_{1\leq i,j\leq n}$ with real (resp. complex) entries, show that there exists a real (resp. complex) inner product space V and vectors $x_1, \ldots, x_n$ such that $\langle x_i, x_j\rangle = a_{ij}$ for all $1 \leq i, j \leq n$.

In analogy with the Euclidean case, we say that two vectors x, y in a (real or complex) vector space are *orthogonal* if $\langle x, y\rangle = 0$. (With this convention, we see in particular that 0 is orthogonal to every vector, and is the only vector with this property.)

Exercise 1.4.3 (Pythagorean theorem)**.** Let V be a real or complex inner product space. If $x_1, \ldots, x_n$ are a finite set of pairwise orthogonal vectors, then $\|x_1 + \ldots + x_n\|^2 = \|x_1\|^2 + \ldots + \|x_n\|^2$. In particular, we see that $\|x_1 + x_2\| \geq \|x_1\|$ whenever x_2 is orthogonal to x_1.

A (possibly infinite) collection $(e_\alpha)_{\alpha \in A}$ of vectors in a (real or complex) inner product space is said to be *orthonormal* if they are pairwise orthogonal and all of unit length.

Exercise 1.4.4. Let $(e_\alpha)_{\alpha \in A}$ be an orthonormal system of vectors in a real or complex inner product space. Show that this system is (algebraically) linearly independent (thus any non-trivial finite linear combination of vectors in this system is non-zero). If x lies in the algebraic span of this system (i.e. it is a finite linear combination of vectors in the system), establish the *inversion formula*

$$x = \sum_{\alpha \in A} \langle x, e_\alpha \rangle e_\alpha \tag{1.47}$$

(with only finitely many of the terms non-zero) and the (finite) *Plancherel formula*

$$\|x\|^2 = \sum_{\alpha \in A} |\langle x, e_\alpha \rangle|^2. \tag{1.48}$$

Exercise 1.4.5 (Gram-Schmidt theorem)**.** Let $e_1, \ldots, e_n$ be a finite orthonormal system in a real or complex inner product space, and let v be a vector not in the span of $e_1, \ldots, e_n$. Show that there exists a vector e_{n+1} with $\mathrm{span}(e_1, \ldots, e_n, e_{n+1}) = \mathrm{span}(e_1, \ldots, e_n, v)$ such that $e_1, \ldots, e_{n+1}$ is an orthonormal system. Conclude that an n-dimensional real or complex inner product space is isomorphic to $\mathbf{R}^n$ or $\mathbf{C}^n$ respectively. Thus, any statement about inner product spaces which only involves a finite-dimensional subspace of that space can be verified just by checking it on Euclidean spaces.

Exercise 1.4.6 (Parallelogram law)**.** For any inner product space V, establish the *parallelogram law*

$$\|x+y\|^2 + \|x-y\|^2 = 2\|x\|^2 + 2\|y\|^2. \tag{1.49}$$

Show that this inequality fails for $L^p(X, \mathcal{X}, \mu)$ for $p \neq 2$ as soon as X contains at least two disjoint sets of non-empty finite measure. On the other hand, establish the *Hanner inequalities*

$$\|f+g\|_p^p + \|f-g\|_p^p \geq (\|f\|_p + \|g\|_p)^p + |\|f\|_p - \|g\|_p|^p \tag{1.50}$$

and

$$(\|f+g\|_p + \|f-g\|_p)^p + |\|f+g\|_p - \|f-g\|_p|^p \leq 2^p(\|f\|_p^p + \|g\|_p^p \tag{1.51}$$

for $1 \leq p \leq 2$, with the inequalities being reversed for $2 \leq p < \infty$. (*Hint*: (1.51) can be deduced from (1.50) by a simple substitution. For (1.50), reduce to the case when f, g are non-negative, and then exploit the inequality

$$\begin{aligned} |x+y|^p + |x-y|^p \geq ((1+r)^{p-1} + (1-r)^{p-1})x^p \\ + ((1+r)^{p-1} - (1-r)^{p-1})r^{1-p}y^p \end{aligned} \tag{1.52}$$

for all non-negative x, y, $0 < r < 1$, and $1 \leq p \leq 2$, with the inequality being reversed for $2 \leq p < \infty$, and with equality being attained when $y < x$ and $r = y/x$.

1.4.2. Hilbert spaces. Thus far, our discussion of inner product spaces has been largely algebraic in nature; this is because we have not been able to take limits inside these spaces and do some actual analysis. This can be rectified by adding an additional axiom:

Definition 1.4.8 (Hilbert spaces)**.** A (real or complex) *Hilbert space* is a (real or complex) inner product space which is complete (or equivalently, an inner product space which is also a Banach space).

Example 1.4.9. From Proposition 1.3.7, (real or complex) $L^2(X, \mathcal{X}, \mu)$ is a Hilbert space for any measure space $(X, \mathcal{X}, \mu)$. In particular, $\mathbf{R}^n$ and $\mathbf{C}^n$ are Hilbert spaces.

Exercise 1.4.7. Show that a subspace of a Hilbert space H will itself be a Hilbert space if and only if it is closed. (In particular, proper dense subspaces of Hilbert spaces are not Hilbert spaces.)

Example 1.4.10. By Example 1.4.9, the space $l^2(\mathbf{Z})$ of doubly infinite square-summable sequences is a Hilbert space. Inside this space, the space $c_c(\mathbf{Z})$ of sequences of finite support is a proper dense subspace (as can be seen for instance by Proposition 1.3.8, though this can also be seen much more directly), and so cannot be a Hilbert space.

Exercise 1.4.8. Let V be an inner product space. Show that there exists a Hilbert space $\overline{V}$ which contains a dense subspace isomorphic to V; we refer to $\overline{V}$ as a *completion* of V. Furthermore, this space is essentially unique in the sense that if $\overline{V}$, $\overline{V}'$ are two such completions, then there exists an isomorphism from $\overline{V}$ to $\overline{V}'$ which is the identity on

V (if one identifies V with the dense subspaces of $\overline{V}$ and $\overline{V'}$. Because of this fact, inner product spaces are sometimes known as *pre-Hilbert spaces*, and can always be identified with dense subspaces of actual Hilbert spaces.

Exercise 1.4.9. Let H, H' be two Hilbert spaces. Define the *direct sum* $H \oplus H'$ of the two spaces to be the vector space $H \times H'$ with inner product $\langle (x,x'),(y,y')\rangle_{H\oplus H'} := \langle x,x'\rangle_H + \langle y,y'\rangle_{H'}$. Show that $H \oplus H'$ is also a Hilbert space.

Example 1.4.11. If H is a complex Hilbert space, one can define the *complex conjugate* $\overline{H}$ of that space to be the set of formal conjugates $\{\overline{x} : x \in H\}$ of vectors in H, with complex vector space structure $\overline{x} + \overline{y} := \overline{x+y}$ and $c\overline{x} := \overline{\overline{c}x}$, and inner product $\langle \overline{x}, \overline{y}\rangle_{\overline{H}} := \langle y, x\rangle_H$. One easily checks that $\overline{H}$ is again a complex Hilbert space. Note the map $x \mapsto \overline{x}$ is not a complex linear isometry; instead, it is a complex *antilinear* isometry.

A key application of the completeness axiom is to be able to define the "nearest point" from a vector to a closed convex body.

Proposition 1.4.12 (Existence of minimisers)**.** *Let H be a Hilbert space, let K be a non-empty closed convex subset of H, and let x be a point in H. Then there exists a unique y in K that minimises the distance $\|y-x\|$ to x. Furthermore, for any other z in K, we have* $\mathrm{Re}\langle z-y, y-x\rangle \geq 0$.

Recall that a subset K of a real or complex vector space is *convex* if $(1-t)v + tw \in K$ whenever $v, w \in K$ and $0 \leq t \leq 1$.

Proof. Observe from the parallelogram law (1.49) that we have the (geometrically obvious) fact that if y and y' are distinct and equidistant from x, then their midpoint $(y+y')/2$ is strictly closer to x than either of y or y'. This (and convexity) ensures that the distance minimiser, if it exists, is unique. Also, if y is the distance minimiser and z is in K, then $(1-\theta)y+\theta z$ is at least as distant from x as y is for any $0 < \theta < 1$, by convexity; squaring this and rearranging, we conclude that

$$2\,\mathrm{Re}\langle z-y, y-x\rangle + \theta\|z-y\|^2 \geq 0. \tag{1.53}$$

Letting $\theta \to 0$ we obtain the final claim in the proposition.

It remains to show existence. Write $D := \inf_{y \in K} \|x - y\|$. It is clear that D is finite and non-negative. If the infimum is attained then we would be done. We cannot conclude immediately that this is the case, but we can certainly find a sequence $y_n \in K$ such that $\|x - y_n\| \to D$. On the other hand, the midpoints $\frac{y_n+y_m}{2}$ lie in K by convexity and so $\|x - \frac{y_n+y_m}{2}\| \geq D$. Using the parallelogram law (1.49) we deduce that $\|y_n - y_m\| \to 0$ as $n, m \to \infty$ and so y_n is a Cauchy sequence; by completeness, it converges to a limit y, which lies in K since K is closed. From the triangle inequality we see that $\|x - y_n\| \to \|x - y\|$, and thus $\|x - y\| = D$, and so y is a distance minimiser. □

Exercise 1.4.10. Show by constructing counterexamples that the existence of the distance minimiser y can fail if either the closure or convexity hypothesis on K is dropped, or if H is merely an inner product space rather than a Hilbert space. (*Hint*: for the last case, let H be the inner product space $C([0,1]) \subset L^2([0,1])$, and let K be the subspace of continuous functions supported on $[0, 1/2]$.) On the other hand, show that existence (but not uniqueness) can be recovered if K is assumed to be compact rather than convex.

Exercise 1.4.11. Using the Hanner inequalities (Exercise 1.4.6), show that Proposition 1.4.12 also holds for the L^p spaces as long as $1 < p < \infty$. (The specific feature of the L^p spaces that is allowing this is known as *uniform convexity*.) Give counterexamples to show that the propsition can fail for L^1 and for L^∞.

Proposition 1.4.12 has some importance in *calculus of variations*, but we will not pursue those applications here.

Since every subspace is necessarily convex, we have a corollary:

Exercise 1.4.12 (Orthogonal projections)**.** Let V be a closed subspace of a Hilbert space H. Then for every $x \in H$ there exists a unique decomposition $x = x_V + x_{V^\perp}$, where $x_V \in V$ and $x_{V^\perp}$ is orthogonal to every element of V. Furthermore, x_V is the closest element of V to x.

Let $\pi_V : H \to V$ be the map $\pi_V : x \mapsto x_V$, where x_V is given by the above exercise; we refer to π_V as the *orthogonal projection* from H onto V. It is not hard to see that π_V is linear, and from the Pythagorean theorem we see that π_V is a contraction (thus $\|\pi_V x\| \leq \|x\|$ for all $x \in V$). In particular, π_V is continuous.

Exercise 1.4.13 (Orthogonal complement)**.** Given a subspace V of a Hilbert space H, define the *orthogonal complement* $V^\perp$ of V to be the set of all vectors in H that are orthogonal to every element of V. Establish the following claims:

- $V^\perp$ is a closed subspace of H, and that $(V^\perp)^\perp$ is the closure of V.
- $V^\perp$ is the trivial subspace $\{0\}$ if and only if V is dense.
- If V is closed, then H is isomorphic to the direct sum of V and $V^\perp$.
- If V, W are two closed subspaces of H, then $(V+W)^\perp = V^\perp \cap W^\perp$ and $(V \cap W)^\perp = \overline{V^\perp + W^\perp}$.

Every vector v in a Hilbert space gives rise to a continuous linear functional $\lambda_v : H \to \mathbf{C}$, defined by the formula $\lambda_v(w) := \langle w, v\rangle$ (the continuity follows from the Cauchy-Schwarz inequality). The *Riesz representation theorem for Hilbert spaces* gives a converse:

Theorem 1.4.13 (Riesz representation theorem for Hilbert spaces)**.** *Let H be a complex Hilbert space, and let $\lambda : H \to \mathbf{C}$ be a continuous linear functional on H. Then there exists a unique v in H such that $\lambda = \lambda_v$. A similar claim holds for real Hilbert spaces (replacing $\mathbf{C}$ by $\mathbf{R}$ throughout).*

Proof. We just show the claim for complex Hilbert spaces, as the claim for real Hilbert spaces is very similar. First, we show uniqueness: if $\lambda_v = \lambda_{v'}$, then $\lambda_{v-v'} = 0$, and in particular $\langle v-v', v-v'\rangle = 0$, and so $v = v'$.

Now we show existence. We may assume that λ is not identically zero, since the claim is obvious otherwise. Observe that the kernel $V := \{x \in H : \lambda(x) = 0\}$ is then a proper subspace of H, which is closed since λ is continuous. By Exercise 1.4.13, the orthogonal

complement $V^\perp$ must contain at least one non-trivial vector w, which we can normalise to have unit magnitude. Since w doesn't lie in V, $\lambda(w)$ is non-zero. Now observe that for any x in H, $x - \frac{\lambda(x)}{\lambda(w)}w$ lies in the kernel of λ, i.e. it lies in V. Taking inner products with w, we conclude that

$$\langle x, w\rangle - \frac{\lambda(x)}{\lambda(w)} = 0 \tag{1.54}$$

and thus

$$\lambda(x) = \langle x, \overline{\lambda(w)}w\rangle \tag{1.55}$$

Thus we have $\lambda = \lambda_{\overline{\lambda(w)}w}$, and the claim follows. □

Remark 1.4.14. This result gives an alternate proof of the $p = 2$ case of Theorem 1.3.16, and by modifying Remark 1.26, can be used to give an alternate proof of the Lebesgue-Radon-Nikodym theorem (this proof is due to von Neumann).

Remark 1.4.15. In the next set of notes, when we define the notion of a dual space, we can reinterpret the Riesz representation theorem as providing a canonical isomorphism $H^* \equiv \overline{H}$.

Exercise 1.4.14. Using Exercise 1.4.11, give an alternate proof of the $1 < p < \infty$ case of Theorem 1.3.16.

One important consequence of the Riesz representation theorem is the existence of adjoints:

Exercise 1.4.15 (Existence of adjoints)**.** Let $T : H \to H'$ be a continuous linear transformation. Show that that there exists a unique continuous linear transformation $T^\dagger : H' \to H$ with the property that $\langle Tx, y\rangle = \langle x, T^\dagger y\rangle$ for all $x \in H$ and $y \in H'$. The transformation $T^\dagger$ is called the (Hilbert space) *adjoint* of T; it is of course compatible with the notion of an adjoint matrix from linear algebra.

Exercise 1.4.16. Let $T : H \to H'$ be a continuous linear transformation.

- Show that $(T^\dagger)^\dagger = T$.
- Show that T is an isometry if and only if $T^\dagger \mathrm{T} = \mathrm{id}_H$.

- Show that T is an isomorphism if and only if $T^\dagger \mathrm{T} = \mathrm{id}_H$ and $TT^\dagger = \mathrm{id}_{H'}$.
- If $S : H' \to H''$ is another continuous linear transformation, show that $(ST)^\dagger = T^\dagger S^\dagger$.

Remark 1.4.16. An isomorphism of complex Hilbert spaces is also known as a *unitary transformation*. (For real Hilbert spaces, the term *orthogonal transformation* is used instead.) Note that unitary and orthogonal $n \times n$ matrices generate unitary and orthogonal transformations on $\mathbf{C}^n$ and $\mathbf{R}^n$ respectively.

Exercise 1.4.17. Show that the projection map $\pi_V : H \to V$ from a Hilbert space to a closed subspace is the adjoint of the inclusion map $\iota_V : V \to H$.

1.4.3. Orthonormal bases. In the section on inner product spaces, we studied finite linear combinations of orthonormal systems. Now that we have completeness, we turn to *infinite* linear combinations.

We begin with countable linear combinations:

Exercise 1.4.18. Suppose that $e_1, e_2, e_3, \ldots$ is a countable orthonormal system in a complex Hilbert space H, and $c_1, c_2, \ldots$ is a sequence of complex numbers. (As usual, similar statements will hold here for real Hilbert spaces and real numbers.)

(i) Show that the series $\sum_{n=1}^\infty c_n e_n$ is conditionally convergent in H if and only if c_n is square-summable.

(ii) If c_n is square-summable, show that $\sum_{n=1}^\infty c_n e_n$ is unconditionally convergent in H, i.e. every permutation of the $c_n e_n$ sums to the same value.

(iii) Show that the map $(c_n)_{n=1}^\infty \mapsto \sum_{n=1}^\infty c_n e_n$ is an isometry from the Hilbert space $\ell^2(\mathbf{N})$ to H. The image V of this isometry is the smallest closed subspace of H that contains $e_1, e_2, \ldots$, and which we shall therefore call the (Hilbert space) *span* of $e_1, e_2, \ldots$.

(iv) Take adjoints of (ii) and conclude that for any $x \in H$, we have $\pi_V(x) = \sum_{n=1}^\infty \langle x, e_n \rangle e_n$ and $\|\pi_V(x)\| = (\sum_{n=1}^\infty |\langle x, e_n \rangle|^2)^{1/2}$. Conclude in particular the *Bessel inequality* $\sum_{n=1}^\infty |\langle x, e_n \rangle|^2 \leq \|x\|^2$.

Remark 1.4.17. Note the contrast here between conditional and unconditional summability (which needs only square-summability of the coefficients c_n) and absolute summability (which requires the stronger condition that the c_n are absolutely summable). In particular there exist non-absolutely summable series that are still unconditionally summable, in contrast to the situation for scalars, in which one has the *Riemann rearrangement theorem*.

Now we can handle arbitrary orthonormal systems $(e_\alpha)_{\alpha \in A}$. If $(c_\alpha)_{\alpha \in A}$ is square-summable, then at most countably many of the c_α are non-zero (by Exercise 1.3.4). Using parts (i), (ii) of Exercise 1.4.18, we can then form the sum $\sum_{\alpha \in A} c_\alpha e_\alpha$ in an unambiguous manner. It is not hard to use Exercise 1.4.18 to then conclude that this gives an isometric embedding of $\ell^2(A)$ into H. The image of this isometry is the smallest closed subspace of H that contains the orthonormal system, which we call the (Hilbert space) *span* of that system. (It is the closure of the algebraic span of the system.)

Exercise 1.4.19. Let $(e_\alpha)_{\alpha \in A}$ be an orthonormal system in H. Show that the following statements are equivalent:

(i) The Hilbert space span of $(e_\alpha)_{\alpha \in A}$ is all of H.

(ii) The algebraic span of $(e_\alpha)_{\alpha \in A}$ (i.e. the finite linear combinations of the e_α) is dense in H.

(iii) One has the *Parseval identity* $\|x\|^2 = \sum_{\alpha \in A} |\langle x, e_\alpha \rangle|^2$ for all $x \in H$.

(iv) One has the *inversion formula* $x = \sum_{\alpha \in A} \langle x, e_\alpha \rangle e_\alpha$ for all $x \in H$ (in particular, the coefficients $\langle x, e_\alpha \rangle$ are square summable).

(v) The only vector that is orthogonal to all the e_α is the zero vector.

(vi) There is an isomorphism from $\ell^2(A)$ to H that maps δ_α to e_α for all $\alpha \in A$ (where δ_α is the *Kronecker delta* at α).

A system $(e_\alpha)_{\alpha \in A}$ obeying any (and hence all) of the properties in Exercise 1.4.19 is known as an *orthonormal basis* of the Hilbert space H. All Hilbert spaces have such a basis:

Proposition 1.4.18. *Every Hilbert space has at least one orthonormal basis.*

Proof. We use the standard Zorn's lemma argument (see Section 2.4). Every Hilbert space has at least one orthonormal system, namely the empty system. We order the orthonormal systems by inclusion, and observe that the union of any totally ordered set of orthonormal systems is again an orthonormal system. By Zorn's lemma, there must exist a maximal orthonormal system $(e_\alpha)_{\alpha \in A}$. There cannot be any unit vector orthogonal to all the elements of this system, since otherwise one could add that vector to the system and contradict orthogonality. Applying Exercise 1.4.19 in the contrapositive, we obtain an orthonormal basis as claimed. □

Exercise 1.4.20. Show that every vector space V has at least one algebraic basis, i.e. a set of basis vectors such that every vector in V can be expressed uniquely as a finite linear combination of basis vectors. (Such bases are also known as *Hamel bases*.)

Corollary 1.4.19. *Every Hilbert space is isomorphic to $\ell^2(A)$ for some set A.*

Exercise 1.4.21. Let A, B be sets. Show that $\ell^2(A)$ and $\ell^2(B)$ are isomorphic iff A and B have the same cardinality. (*Hint*: the case when A or B is finite is easy, so suppose A and B are both infinite. If $\ell^2(A)$ and $\ell^2(B)$ are isomorphic, show that B can be covered by a family of at most countable sets indexed by A, and vice versa. Then apply the *Schröder-Bernstein theorem* (Section 3.13).

We can now classify Hilbert spaces up to isomorphism by a single cardinal, the dimension of that space:

Exercise 1.4.22. Show that all orthonormal bases of a given Hilbert space H have the same cardinality. This cardinality is called the (Hilbert space) *dimension* of the Hilbert space.

Exercise 1.4.23. Show that a Hilbert space is separable (i.e. has a countable dense subset) if and only if its dimension is at most countable. Conclude in particular that up to isomorphism, there is exactly one separable infinite-dimensional Hilbert space.

Exercise 1.4.24. Let H, H' be complex Hilbert spaces. Show that there exists another Hilbert space $H \otimes H'$, together with a map $\otimes : H \times H' \to H \otimes H'$ with the following properties:

(i) The map $\otimes$ is bilinear, thus $(cx+dy)\otimes x' = c(x\otimes x') + d(y\otimes x')$ and $x \otimes (cx'+dy') = c(x\otimes x') + d(x\otimes y')$ for all $x, y \in H, x', y' \in H', c, d \in \mathbf{C}$;

(ii) We have $\langle x\otimes x', y\otimes y'\rangle_{H\otimes H'} = \langle x, y\rangle_H \langle x', y'\rangle_{H'}$ for all $x, y \in H, x', y' \in H'$.

(iii) The (algebraic) span of $\{x\otimes x' : x \in H, x' \in H'\}$ is dense in $H \otimes H'$.

Furthermore, show that $H \otimes H'$ and $\otimes$ are unique up to isomorphism in the sense that if $H\tilde{\otimes}H'$ and $\tilde{\otimes} : H \times H' \to H\tilde{\otimes}H'$ are another pair of objects obeying the above properties, then there exists an isomorphism $\Phi : H \otimes H' \to H\tilde{\otimes}H'$ such that $x\tilde{\otimes}x' = \Phi(x\otimes x')$ for all $x \in H, x' \in H'$. (*Hint*: to prove existence, create orthonormal bases for H and H' and take formal tensor products of these bases.) The space $H \otimes H'$ is called the (Hilbert space) *tensor product* of H and H', and $x \otimes x'$ is the tensor product of x and x'.

Exercise 1.4.25. Let $(X, \mathcal{X}, \mu)$ and $(Y, \mathcal{Y}, \nu)$ be measure spaces. Show that $L^2(X\times Y, \mathcal{X}\times\mathcal{Y}, \mu\times\nu)$ is the tensor product of $L^2(X, \mathcal{X}, \mu)$ and $L^2(Y, \mathcal{Y}, \mu)$, if one defines the tensor product $f\otimes g$ of $f \in L^2(X, \mathcal{X}, \mu)$ and $g \in L^2(Y, \mathcal{Y}, \mu)$ as $f \otimes g(x, y) := f(x)g(y)$.

We do not yet have enough theory in other areas to give the really useful applications of Hilbert space theory yet, but let us just illustrate a simple one, namely the development of *Fourier series* on the unit circle $\mathbf{R}/\mathbf{Z}$. We can give this space the usual Lebesgue measure (identifying the unit circle with $[0, 1)$, if one wishes), giving rise to the complex Hilbert space $L^2(\mathbf{R}/\mathbf{Z})$. On this space we can form the *characters* $e_n(x) := e^{2\pi inx}$ for all integer n; one easily verifies that $(e_n)_{n\in\mathbf{Z}}$ is an orthonormal system. We claim that it is in fact an orthonormal basis. By Exercise 1.4.19, it suffices to show that the algebraic span of the e_n, i.e. the space of trigonometric polynomials, is dense in $L^2(\mathbf{R}/\mathbf{Z})$. But[8] from an explicit computation (e.g. using *Fejér*

[8]One can also use the Stone-Weierstrass theorem here, see Theorem 1.10.24.

kernels) one can show that the indicator function of any interval can be approximated to arbitrary accuracy in L^2 norm by trigonometric polynomials, and is thus in the closure of the trigonometric polynomials. By linearity, the same is then true of an indicator function of a finite union of intervals; since Lebesgue measurable sets in $\mathbf{R}/\mathbf{Z}$ can be approximated to arbitrary accuracy by finite unions of intervals, the same is true for indicators of measurable sets. By linearity, the same is true for simple functions, and by density (Proposition 1.3.8) the same is true for arbitrary L^2 functions, and the claim follows.

The Fourier transform $\hat{f} : \mathbf{Z} \to \mathbf{C}$ of a function $f \in L^2(\mathbf{R}/\mathbf{Z})$ is defined as

$$\hat{f}(n) := \langle f, e_n \rangle = \int_0^1 f(x) e^{-2\pi i n x}\ dx. \tag{1.56}$$

From Exercise 1.4.19, we obtain the *Parseval identity*

$$\sum_{n \in \mathbf{Z}} |\hat{f}(n)|^2 = \int_{\mathbf{R}/\mathbf{Z}} |f(x)|^2\ dx$$

(in particular, $\hat{f} \in \ell^2(\mathbf{Z})$) and the *inversion formula*

$$f = \sum_{n \in \mathbf{Z}} \hat{f}(n) e_n$$

where the right-hand side is unconditionally convergent. Indeed, the Fourier transform $f \mapsto \hat{f}$ is a unitary transformation between $L^2(\mathbf{R}/\mathbf{Z})$ and $\ell^2(\mathbf{Z})$. (These facts are collectively referred to as *Plancherel's theorem* for the unit circle.) We will develop Fourier analysis on other spaces than the unit circle in Section 1.12.

Remark 1.4.20. Of course, much of the theory here generalises the corresponding theory in finite-dimensional linear algebra; we will continue this theme much later in the course when we turn to the spectral theorem. However, not every aspect of finite-dimensional linear algebra will carry over so easily. For instance, it turns out to be quite difficult to take the determinant or trace of a linear transformation from a Hilbert space to itself in general (unless the transformation is particularly well behaved, e.g. of trace class). The *Jordan normal form* also does not translate to the infinite-dimensional setting, leading to the notorious *invariant subspace problem* in the subject. It is also worth cautioning that while the theory of orthonormal bases

in finite-dimensional Euclidean spaces generalises very nicely to the Hilbert space setting, the more general theory of bases in finite dimensions becomes much more subtle in infinite dimensional Hilbert spaces, unless the basis is "almost orthonormal" in some sense (e.g. if it forms a frame).

Notes. This lecture first appeared at `terrytao.wordpress.com/2009/01/17`. Thanks to Américo Tavares, S, and Xiaochuan Liu for corrections.

Uhlrich Groh and Dmitriy raised the interesting open problem of whether any closed subset K of H for which distance minimisers to every point x existed and unique were necessarily convex, thus providing a converse to Proposition 1.4.12. (Sets with this property are known as *Chebyshev sets*.)

1.5. Duality and the Hahn-Banach theorem

When studying a mathematical space X (e.g. a vector space, a topological space, a manifold, a group, an algebraic variety etc.), there are two fundamentally basic ways to try to understand the space:

(i) By looking at subobjects in X, or more generally maps $f : Y \to X$ from some other space Y into X. For instance, a point in a space X can be viewed as a map from pt to X; a curve in a space X could be thought of as a map from $[0,1]$ to X; a group G can be studied via its subgroups K, and so forth.

(ii) By looking at objects on X, or more precisely maps $f : X \to Y$ from X into some other space Y. For instance, one can study a topological space X via the real- or complex-valued continuous functions $f \in C(X)$ on X; one can study a group G via its quotient groups $\pi : G \to G/H$; one can study an algebraic variety V by studying the polynomials on V (and in particular, the ideal of polynomials that vanish identically on V); and so forth.

(There are also more sophisticated ways to study an object via its maps, e.g. by studying extensions, joinings, splittings, universal lifts, etc. The general study of objects via the maps between them is

formalised abstractly in modern mathematics as *category theory*, and is also closely related to *homological algebra*.)

A remarkable phenomenon in many areas of mathematics is that of (contravariant) *duality*: that the maps into and out of one type of mathematical object X can be naturally associated to the maps out of and into a *dual object* X^* (note the reversal of arrows here!). In some cases, the dual object X^* looks quite different from the original object X. (For instance, in *Stone duality*, discussed in Section 2.3, X would be a Boolean algebra (or some other partially ordered set) and X^* would be a compact totally disconnected Hausdorff space (or some other topological space).) In other cases, most notably with Hilbert spaces as discussed in Section 1.4, the dual object X^* is essentially identical to X itself.

In these notes we discuss a third important case of duality, namely duality of normed vector spaces, which is of an intermediate nature to the previous two examples: the dual X^* of a normed vector space turns out to be another normed vector space, but generally one which is not equivalent to X itself (except in the important special case when X is a Hilbert space, as mentioned above). On the other hand, the double dual $(X^*)^*$ turns out to be closely related to X, and in several (but not all) important cases, is essentially identical to X. One of the most important uses of dual spaces in functional analysis is that it allows one to define the *transpose* $T^* : Y^* \to X^*$ of a continuous linear operator $T : X \to Y$.

A fundamental tool in understanding duality of normed vector spaces will be the *Hahn-Banach theorem*, which is an indispensable tool for exploring the dual of a vector space. (Indeed, without this theorem, it is not clear at all that the dual of a non-trivial normed vector space is non-trivial!) Thus, we shall study this theorem in detail in this section concurrently with our discussion of duality.

1.5.1. Duality. In the category of normed vector spaces, the natural notion of a "map" (or morphism) between two such spaces is that of a continuous linear transformation $T : X \to Y$ between two normed vector spaces X, Y. By Lemma 1.3.17, any such linear transformation is bounded, in the sense that there exists a constant C such that

$\|Tx\|_Y \leq C\|x\|_X$ for all $x \in X$. The least such constant C is known as the *operator norm* of T, and is denoted $\|T\|_{\mathrm{op}}$ or simply $\|T\|$.

Two normed vector spaces X, Y are equivalent if there is an invertible continuous linear transformation $T : X \to Y$ from X to Y, thus T is bijective and there exist constants $C, c > 0$ such that $c\|x\|_X \leq \|Tx\|_Y \leq C\|x\|_X$ for all $x \in X$. If one can take $C = c = 1$, then T is an isometry, and X and Y are called isomorphic. When one has two norms $\|\|1, \|\|_2$ on the same vector space X, we say that the norms are equivalent if the identity from $(X, \|\|_1)$ to $(X, \|\|_2)$ is an invertible continuous transformation, i.e. that there exist constants $C, c > 0$ such that $c\|x\|_1 \leq \|x\|_2 \leq C\|x\|_1$ for all $x \in X$.

Exercise 1.5.1. Show that all linear transformations from a finite-dimensional space to a normed vector space are continuous. Conclude that all norms on a finite-dimensional space are equivalent.

Let $B(X \to Y)$ denote the space of all continuous linear transformations from X to Y. (This space is also denoted by many other names, e.g. $\mathcal{L}(X, Y)$, $\mathrm{Hom}(X \to Y)$, etc.) This has the structure of a vector space: the sum $S + T : x \mapsto Sx + Tx$ of two continuous linear transformations is another continuous linear transformation, as is the scalar multiple $cT : x \mapsto cTx$ of a linear transformation.

Exercise 1.5.2. Show that $B(X \to Y)$ with the operator norm is a normed vector space. If Y is complete (i.e. is a Banach space), show that $B(X \to Y)$ is also complete (i.e. is also a Banach space).

Exercise 1.5.3. Let X, Y, Z be Banach spaces. Show that if $T \in B(X \to Y)$ and $S \in B(Y \to Z)$, then the composition $ST : X \to Z$ lies in $B(X \to Z)$ and $\|ST\|_{\mathrm{op}} \leq \|S\|_{\mathrm{op}}\|T\|_{\mathrm{op}}$. (As a consequence of this inequality, we see that $B(X \to X)$ is a *Banach algebra*.)

Now we can define the notion of a dual space.

Definition 1.5.1 (Dual space)**.** Let X be a normed vector space. The (continuous) *dual space* X^* of X is defined to be $X^* := B(\mathrm{X} \to \mathbf{R})$ if X is a real vector space, and $X^* := B(X \to \mathbf{C})$ if X is a complex vector space. Elements of X^* are known as *continuous linear functionals* (or *bounded linear functionals*) on X.

Remark 1.5.2. If one drops the requirement that the linear functionals be continuous, we obtain the algebraic dual space of linear functionals on X. This space does not play a significant role in functional analysis, though.

From Exercise 1.5.2, we see that the dual of any normed vector space is a Banach space, and so duality is arguably a Banach space notion rather than a normed vector space notion. The following exercise reinforces this:

Exercise 1.5.4. We say that a normed vector space X has a *completion* $\overline{X}$ if $\overline{X}$ is a Banach space and X can be identified with a dense subspace of $\overline{X}$ (cf. Exercise 1.4.8).

(i) Show that every normed vector space X has at least one completion $\overline{X}$, and that any two completions $\overline{X}, \overline{X}'$ are isomorphic in the sense that there exists an isomorphism from $\overline{X}$ to $\overline{X}'$ which is the identity on X.

(ii) Show that the dual spaces X^* and $(\overline{X})^*$ are isomorphic to each other.

The next few exercises are designed to give some intuition as to how dual spaces work.

Exercise 1.5.5. Let $\mathbf{R}^n$ be given the Euclidean metric. Show that $(\mathbf{R}^n)^*$ is isomorphic to $\mathbf{R}^n$. Establish the corresponding result for the complex spaces $\mathbf{C}^n$.

Exercise 1.5.6. Let $c_c(\mathbf{N})$ be the vector space of sequences $(a_n)_{n\in\mathbf{N}}$ of real or complex numbers which are compactly supported (i.e. at most finitely many of the a_n are non-zero). We give c_c the uniform norm $\|\|_{\ell^\infty}$.

(i) Show that the dual space $c_c(\mathbf{N})^*$ is isomorphic to $\ell^1(\mathbf{N})$.

(ii) Show that the completion of $c_c(\mathbf{N})$ is isomorphic to $c_0(\mathbf{N})$, the space of sequences on $\mathbf{N}$ that go to zero at infinity (again with the uniform norm); thus, by Exercise 1.5.4, the dual space of $c_0(\mathbf{N})$ is isomorphic to $\ell^1(\mathbf{N})$ also.

(iii) On the other hand, show that the dual of $\ell^1(\mathbf{N})$ is isomorphic to $\ell^\infty(\mathbf{N})$, a space which is strictly larger than $c_c(\mathbf{N})$ or

$c_0(\mathbf{N})$. Thus we see that the double dual of a Banach space can be strictly larger than the space itself.

Exercise 1.5.7. Let H be a real or complex Hilbert space. Using the Riesz representation theorem for Hilbert spaces (Theorem 1.4.13), show that the dual space H^* is isomorphic (as a normed vector space) to the conjugate space $\overline{H}$ (see Example 1.4.11), with an element $g \in \overline{H}$ being identified with the linear functional $f \mapsto \langle f, g \rangle$. Thus we see that Hilbert spaces are essentially self-dual (if we ignore the pesky conjugation sign).

Exercise 1.5.8. Let $(X, \mathcal{X}, \mu)$ be a σ-finite measure space, and let $1 \leq p < \infty$. Using Theorem 1.3.16, show that the dual space of $L^p(X, \mathcal{X}, \mu)$ is isomorphic to $L^{p'}(X, \mathcal{X}, \mu)$, with an element $g \in L^{p'}(X, \mathcal{X}, \mu)$ being identified with the linear functional $f \mapsto \int_X fg \, d\mu$. (The one tricky thing to verify is that the identification is an isometry, but this can be seen by a closer inspection of the proof of Theorem 1.3.16.) For an additional challenge: remove the σ-finite hypothesis when $p > 1$.

One of the key purposes of introducing the notion of a dual space is that it allows one to define the notion of a *transpose*.

Definition 1.5.3 (Transpose)**.** Let $T : X \to Y$ be a continuous linear transformation from one normed vector space X to another Y. The *transpose* $T^* : Y^* \to X^*$ of T is defined to be the map that sends any continuous linear functional $\lambda \in Y^*$ to the linear functional $T^*\lambda := \lambda \circ \mathrm{T} \in X^*$, thus $(T^*\lambda)(x) = \lambda(Tx)$ for all $x \in X$.

Exercise 1.5.9. Show that the transpose T^* of a continuous linear transformation T between normed vector spaces is again a continuous linear transformation with $\|T^*\|_{\mathrm{op}} \leq \|T\|_{\mathrm{op}}$, thus the transpose operation is itself a linear map from $B(X \to Y)$ to $B(Y^* \to X^*)$. (We will improve this result in Theorem 1.5.13 below.)

Exercise 1.5.10. An $n \times m$ matrix A with complex entries can be identified with a linear transformation $L_A : \mathbf{C}^n \to \mathbf{C}^m$. Identifying the dual space of $\mathbf{C}^n$ with itself as in Exercise 1.5.5, show that the transpose $L_A^* : \mathbf{C}^m \to \mathbf{C}^n$ is equal to L_{A^t}, where A^t is the transpose matrix of A.

Exercise 1.5.11. Show that the transpose of a surjective continuous linear transformation between normed vector spaces is injective. Show also that the condition of surjectivity can be relaxed to that of having a dense image.

Remark 1.5.4. Observe that if $T : X \to Y$ and $S : Y \to Z$ are continuous linear transformations between normed vector spaces, then $(ST)^* = T^*S^*$. In the language of category theory, this means that duality $X \mapsto X^*$ of normed vector spaces, and transpose $T \mapsto T^*$ of continuous linear transformations, form a *contravariant functor* from the category of normed vector spaces (or Banach spaces) to itself.

Remark 1.5.5. The transpose $T^* : \overline{H'} \to \overline{H}$ of a continuous linear transformation $T : H \to H'$ between complex Hilbert spaces is closely related to the adjoint $T^\dagger : H' \to H$ of that transformation, as defined in Exercise 1.4.15, by using the obvious (antilinear) identifications between H and $\overline{H}$, and between H' and $\overline{H'}$. This is analogous to the linear algebra fact that the adjoint matrix is the complex conjugate of the transpose matrix. One should note that in the literature, the transpose operator T^* is also (somewhat confusingly) referred to as the adjoint of T. Of course, for real vector spaces, there is no distinction between transpose and adjoint.

1.5.2. The Hahn-Banach theorem. Thus far, we have defined the dual space X^*, but apart from some concrete special cases (Hilbert spaces, L^p spaces, etc.) we have not been able to say much about what X^* consists of - it is not even clear yet that if X is non-trivial (i.e. not just $\{0\}$), that X^* is also non-trivial - for all one knows, there could be no non-trivial continuous linear functionals on X at all! The Hahn-Banach theorem is used to resolve this, by providing a powerful means to construct continuous linear functionals as needed.

Theorem 1.5.6 (Hahn-Banach theorem)**.** *Let X be a normed vector space, and let Y be a subspace of X. Then any continuous linear functional $\lambda \in Y^*$ on Y can be extended to a continuous linear functional $\tilde{\lambda} \in X^*$ on X with the same operator norm; thus $\tilde{\lambda}$ agrees with λ on Y and $\|\tilde{\lambda}\|_{X^*} = \|\lambda\|_{Y^*}$. (Note: the extension $\tilde{\lambda}$ is, in general, not unique.)*

We prove this important theorem in stages. We first handle the codimension one real case:

Proposition 1.5.7. *The Hahn-Banach theorem is true when X, Y are real vector spaces, and X is spanned by Y and an additional vector v.*

Proof. We can assume that v lies outside Y, since the claim is trivial otherwise. We can also normalise $\|\lambda\|_{Y^*} = 1$ (the claim is of course trivial if $\|\lambda\|_{Y^*}$ vanishes). To specify the extension $\tilde{\lambda}$ of λ, it suffices by linearity to specify the value of $\tilde{\lambda}(v)$. In order for the extension $\tilde{\lambda}$ to continue to have operator norm 1, we require that

$$|\tilde{\lambda}(y+tv)| \leq \|y+tv\|_X$$

for all $t \in \mathbf{R}$ and $y \in Y$. This is automatic for $t = 0$, so by homogeneity it suffices to attain this bound for $t = 1$. We rearrange this a bit as

$$\sup_{y' \in Y} \lambda(y') - \|y'+v\|_X \leq \tilde{\lambda}(v) \leq \inf_{y \in Y} \|y+v\|_X - \lambda(y).$$

But as λ has operator norm 1, an application of the triangle inequality shows that the infimum on the right-hand side is at least as large as the supremum on the left-hand side, and so one can choose $\tilde{\lambda}(v)$ obeying the required properties. □

Corollary 1.5.8. *The Hahn-Banach theorem is true when X, Y are real normed vector spaces.*

Proof. This is a standard "Zorn's lemma" argument (see Section 2.4). Fix Y, X, λ. Define a partial extension of λ to be a pair (Y', λ'), where Y' is an intermediate subspace between Y and X, and λ' is an extension of λ with the same operator norm as λ. The set of all partial extensions is partially ordered by declaring $(Y'', \lambda'') \geq (Y', \lambda')$ if Y'' contains Y' and λ'' extends λ'. It is easy to see that every chain of partial extensions has an upper bound; hence, by Zorn's lemma, there must be a maximal partial extension (Y_*, λ_*). If $Y_* = X$, we are done; otherwise, one can find $v \in X \backslash Y_*$. By Proposition 1.5.7, we can then extend λ_* further to the larger space spanned by Y_* and v, a contradiction; and the claim follows. □

Remark 1.5.9. Of course, this proof of the Hahn-Banach theorem relied on the axiom of choice (via Zorn's lemma) and is thus non-constructive. It turns out that this is, to some extent, necessary: it is not possible to prove the Hahn-Banach theorem if one deletes the axiom of choice from the axioms of set theory (although it is possible to deduce the theorem from slightly weaker versions of this axiom, such as the *ultrafilter lemma*).

Finally, we establish the complex case by leveraging the real case.

Proof of Hahn-Banach theorem (complex case). Let $\lambda : Y \to \mathbf{C}$ be a continuous complex-linear functional, which we can normalise to have operator norm 1. Then the real part $\rho := \mathrm{Re}(\lambda) : Y \to \mathbf{R}$ is a continuous real-linear functional on Y (now viewed as a real normed vector space rather than a complex one), which has operator norm at most 1 (in fact, it is equal to 1, though we will not need this). Applying Corollary 1.5.8, we can extend this real-linear functional ρ to a continuous real-linear functional $\tilde{\rho} : X \to \mathbf{R}$ on X (again viewed now just as a real normed vector space) of norm at most 1.

To finish the job, we have to somehow complexify $\tilde{\rho}$ to a complex-linear functional $\tilde{\lambda} : X \to \mathbf{R}$ of norm at most 1 that agrees with λ on Y. It is reasonable to expect that $\mathrm{Re}\,\tilde{\lambda} = \tilde{\rho}$; a bit of playing around with complex linearity then forces

$$\tilde{\lambda}(x) := \tilde{\rho}(x) - i\tilde{\rho}(ix). \tag{1.57}$$

Accordingly, we shall use (1.57) to *define* $\tilde{\lambda}$. It is easy to see that $\tilde{\lambda}$ is a continuous complex-linear functional agreeing with λ on Y. Since $\tilde{\rho}$ has norm at most 1, we have $|\mathrm{Re}\,\tilde{\lambda}(x)| \leq \|x\|_X$ for all $x \in X$. We can amplify this (cf. Section 1.9 of *Structure and Randomness*) by exploiting phase rotation symmetry, thus $|\mathrm{Re}\,\tilde{\lambda}(e^{i\theta}x)| \leq \|x\|_X$ for all $\theta \in \mathbf{R}$. Optimising in θ we see that $\tilde{\rho}$ has norm at most 1, as required. □

Exercise 1.5.12. In the special case when X is a Hilbert space, give an alternate proof of the Hahn-Banach theorem, using the material from Section 1.4, that avoids Zorn's lemma or the axiom of choice.

Now we put this Hahn-Banach theorem to work in the study of duality and transposes.

Exercise 1.5.13. Let $T : X \to Y$ be a continuous linear transformation which is bounded from below (i.e. there exists $c > 0$ such that $\|Tx\| \geq c\|x\|$ for all $x \in X$); note that this ensures that X is equivalent to some subspace of Y. Show that the transpose $T^* : Y^* \to X^*$ is surjective. Give an example to show that the claim fails if T is merely assumed to be injective rather than bounded from below. (*Hint*: consider the map $(a_n)_{n=1}^\infty \to (a_n/n)_{n=1}^\infty$ on some suitable space of sequences.) This should be compared with Exercise 1.5.11.

Exercise 1.5.14. Let x be an element of a normed vector space X. Show that there exists $\lambda \in X^*$ such that $\|\lambda\|_{X^*} = 1$ and $\lambda(x) = \|x\|_X$. Conclude in particular that the dual of a non-trivial normed vector space is again non-trivial.

Given a normed vector space X, we can form its double dual $(X^*)^*$: the space of linear functionals on X^*. There is a very natural map $\iota : X \to (X^*)^*$, defined as

$$\iota(x)(\lambda) := \lambda(x) \tag{1.58}$$

for all $x \in X$ and $\lambda \in X^*$. (This map is closely related to the *Gelfand transform* in the theory of operator algebras; see Section 1.10.4.) It is easy to see that ι is a continuous linear transformation, with operator norm at most 1. But the Hahn-Banach theorem gives a stronger statement:

Theorem 1.5.10. *ι is an isometry.*

Proof. We need to show that $\|\iota(x)\|_{X^{**}} = \|x\|$ for all $x \in X$. The upper bound is clear; the lower bound follows from Exercise 1.5.14. □

Exercise 1.5.15. Let Y be a subspace of a normed vector space X. Define the complement $Y^\perp$ of Y to be the space of all $\lambda \in X^*$ which vanish on Y.

(i) Show that $Y^\perp$ is a closed subspace of X^*, and that $\overline{Y} := \{x \in X : \lambda(x) = 0 \text{ for all } \lambda \in Y^\perp\}$; (Compare with Exercise 1.4.13.) In other words, $\iota(\overline{Y}) = \iota(X) \cap Y^{\perp\perp}$.

(ii) Show that $Y^\perp$ is trivial if and only if Y is dense, and $Y^\perp = X^*$ if and only if Y is trivial.

(iii) Show that $Y^\perp$ is isomorphic to the dual of the quotient space $X/\overline{Y}$ (which has the norm $\|x+\overline{Y}\|_{X/\overline{Y}} := \inf_{y\in\overline{Y}} \|x+y\|_X$).

(iv) Show that Y^* is isomorphic to $X^*/Y^\perp$.

From Theorem 1.5.10, every normed vector space can be identified with a subspace of its double dual (and every Banach space is identified with a closed subspace of its double dual). If ι is surjective, then we have an isomorphism $X \equiv X^{**}$, and we say that X is reflexive in this case; since X^{**} is a Banach space, we conclude that only Banach spaces can be reflexive. From linear algebra we see in particular that any finite-dimensional normed vector space is reflexive; from Exercises 1.5.7, 1.5.8 we see that any Hilbert space and any L^p space with $1 < p < \infty$ on a σ-finite space is also reflexive (and the hypothesis of σ-finiteness can in fact be dropped). On the other hand, from Exercise 1.5.6, we see that the Banach space $c_0(\mathbf{N})$ is not reflexive.

An important fact is that $l^1(\mathbf{N})$ is also not reflexive: the dual of $l^1(\mathbf{N})$ is equivalent to $l^\infty(\mathbf{N})$, but the dual of $l^\infty(\mathbf{N})$ is strictly larger than that of $l^1(\mathbf{N})$. Indeed, consider the subspace $c(\mathbf{N})$ of $l^\infty(\mathbf{N})$ consisting of bounded convergent sequences (equivalently, this is the space spanned by $c_0(\mathbf{N})$ and the constant sequence $(1)_{n\in\mathbf{N}}$). The limit functional $(a_n)_{n=1}^\infty \mapsto \lim_{n\to\infty} a_n$ is a bounded linear functional on $c(\mathbf{N})$, with operator norm 1, and thus by the Hahn-Banach theorem can be extended to a generalised limit functional $\lambda : l^\infty(\mathbf{N}) \to \mathbf{C}$ which is a continuous linear functional of operator norm 1. As such generalised limit functionals annihilate all of $c_0(\mathbf{N})$ but are still non-trivial, they do not correspond to any element of $\ell^1(\mathbf{N}) \equiv c_0(\mathbf{N})^*$.

Exercise 1.5.16. Let $\lambda : l^\infty(\mathbf{N}) \to \mathbf{C}$ be a generalised limit functional (i.e. an extension of the limit functional of $c(\mathbf{N})$ of operator norm 1) which is also an algebra homomorphism, i.e. $\lambda((x_n y_n)_{n=1}^\infty) = \lambda((x_n)_{n=1}^\infty)\lambda((y_n)_{n=1}^\infty)$ for all sequences $(x_n)_{n=1}^\infty, (y_n)_{n=1}^\infty \in \ell^\infty(\mathbf{N})$. Show that there exists a unique non-principal ultrafilter $p \in \beta\mathbf{N}\backslash\mathbf{N}$ (as defined for instance Section 1.5 of *Structure and Randomness*) such that $\lambda((x_n)_{n=1}^\infty) = \lim_{n\to p} x_n$ for all sequences $(x_n)_{n=1}^\infty \in \ell^\infty(\mathbf{N})$. Conversely, show that every non-principal ultrafilter generates a generalised limit functional that is also an algebra homomorphism. (This

exercise may become easier once one is acquainted with the Stone-Čech compactification, see Section 2.5. If the algebra homomorphism property is dropped, one has to consider probability measures on the space of non-principal ultrafilters instead.)

Exercise 1.5.17. Show that any closed subspace of a reflexive space is again reflexive. Also show that a Banach space X is reflexive if and only if its dual is reflexive. Conclude that if $(X, \mathcal{X}, \mu)$ is a measure space which contains a countably infinite sequence of disjoint sets of positive measure, then $L^1(X, \mathcal{X}, \mu)$ and $L^\infty(X, \mathcal{X}, \mu)$ are not reflexive. (*Hint*: Reduce to the σ-finite case. L^∞ will contain an isometric copy of $\ell^\infty(\mathbf{N})$.)

Theorem 1.5.10 gives a classification of sorts for normed vector spaces:

Corollary 1.5.11. *Every normed vector space X is isomorphic to a subspace of $BC(Y)$, the space of bounded continuous functions on some bounded complete metric space Y, with the uniform norm.*

Proof. Take Y to be the unit ball in X^*, then the map ι identifies X with a subspace of $BC(Y)$. □

Remark 1.5.12. If X is separable, it is known that one can take Y to just be the unit interval $[0, 1]$; this is the *Banach-Mazur theorem*, which we will not prove here.

Next, we apply the Hahn-Banach theorem to the transpose operation, improving Exercise 1.5.9:

Theorem 1.5.13. *Let $T : X \to Y$ be a continuous linear transformation between normed vector spaces. Then $\|T^*\|_{\mathrm{op}} = \|T\|_{\mathrm{op}}$; thus the transpose operation is an isometric embedding of $B(X \to Y)$ into $B(Y^* \to X^*)$.*

Proof. By Exercise 1.5.9, it suffices to show that $\|T^*\|_{\mathrm{op}} \geq \|T\|_{\mathrm{op}}$. Accordingly, let α be any number strictly less than $\|T\|_{\mathrm{op}}$, then we can find $x \in X$ such that $\|Tx\|_Y \geq \alpha\|x\|$. By Exercise 1.5.14 we can then find $\lambda \in Y^*$ such that $\|\lambda\|_{Y^*} = 1$ and $\lambda(Tx) = T^*\lambda(x) = \|Tx\|_Y \geq \alpha\|x\|$, and thus $\|T^*\lambda\|_{X^*} \geq \alpha$. This implies that $\|T^*\|_{\mathrm{op}} \geq \alpha$; taking suprema over all α strictly less than $\|T\|_{\mathrm{op}}$ we obtain the claim. □

If we identify X and Y with subspaces of X^{**} and Y^{**} respectively, we thus see that $T^{**} : X^{**} \to Y^{**}$ is an extension of $T : X \to Y$ with the same operator norm. In particular, if X and Y are reflexive, we see that T^{**} can be identified with T itself (exactly as in the finite-dimensional linear algebra setting).

1.5.3. Variants of the Hahn-Banach theorem (optional). The Hahn-Banach theorem has a number of essentially equivalent variants, which also are of interest for the geometry of normed vector spaces.

Exercise 1.5.18 (Generalised Hahn-Banach theorem)**.** Let Y be a subspace of a real or complex vector space X, let $\rho : X \to \mathbf{R}$ be a sublinear functional on X (thus $\rho(cx) = c\rho(x)$ for all non-negative c and all $x \in X$, and $\rho(x+y) \leq \rho(x) + \rho(y)$ for all $x, y \in X$), and let $\lambda : Y \to \mathbf{R}$ be a linear functional on Y such that $\lambda(y) \leq \rho(y)$ for all $y \in Y$. Show that λ can be extended to a linear functional $\tilde{\lambda}$ on X such that $\tilde{\lambda}(x) \leq \rho(x)$ for all $x \in X$. Show that this statement implies the usual Hahn-Banach theorem. (*Hint*: adapt the proof of the Hahn-Banach theorem.)

Call a subset A of a real vector space V *algebraically open* if the sets $\{t : x + tv \in A\}$ are open in $\mathbf{R}$ for all $x, v \in V$; note that every open set in a normed vector space is algebraically open.

Theorem 1.5.14 (Geometric Hahn-Banach theorem)**.** *Let A, B be convex subsets of a real vector space V, with A algebraically open. Then the following are equivalent:*

(i) *A and B are disjoint.*

(ii) *There exists a linear functional $\lambda : V \to \mathbf{R}$ and a constant c such that $\lambda < c$ on A, and $\lambda \geq c$ on B. (Equivalently, there is a hyperplane separating A and B, with A avoiding the hyperplane entirely.)*

If A and B are convex cones (i.e. $tx \in A$ whenever $x \in A$ and $t > 0$, and similarly for B), we may take $c = 0$.

Remark 1.5.15. In finite dimensions, it is not difficult to drop the algebraic openness hypothesis on A as long as one now replaces the condition $\lambda < c$ by $\lambda \leq c$. However in infinite dimensions one cannot

do this. Indeed, if we take $V = c_c(\mathbf{N})$, let A be the set of sequences whose last non-zero element is strictly positive, and $B = -A$ consist of those sequences whose last non-zero element is strictly negative, then one can verify that there is no hyperplane separating A from B.

Proof. Clearly (ii) implies (i); now we show that (i) implies (ii). We first handle the case when A and B are convex cones.

Define a *good pair* to be a pair (A, B) where A and B are disjoint convex cones, with A algebraically open, thus (A, B) is a good pair by hypothesis. We can order $(A, B) \leq (A', B')$ if A' contains A and B' contains B. A standard application of Zorn's lemma (Section 2.4) reveals that any good pair (A, B) is contained in a maximal good pair, and so without loss of generality we may assume that (A, B) is a maximal good pair.

We can of course assume that neither A nor B is empty. We now claim that B is the complement of A. For if not, then there exists $v \in V$ which does not lie in either A or B. By the maximality of (A, B), the convex cone generated by $B \cup \{v\}$ must intersect A at some point, say w. By dilating w if necessary we may assume that w lies on a line segment between v and some point b in B. By using the convexity and disjointness of A and B one can then deduce that for any $a \in A$, the ray $\{a + t(w - b) : t > 0\}$ is disjoint from B. Thus one can enlarge A to the convex cone generated by A and $w - b$, which is still algebraically open and now strictly larger than A (because it contains v), a contradiction. Thus B is the complement of A.

Let us call a line in V *monochromatic* if it is entirely contained in A or entirely contained in B. Note that if a line is not monochromatic, then (because A and B are convex and partition the line, and A is algebraically open), the line splits into an open ray contained in A, and a closed ray contained in B. From this we can conclude that if a line is monochromatic, then all parallel lines must also be monochromatic, because otherwise we look at the ray in the parallel line which contains A and use convexity of both A and B to show that this ray is adjacent to a halfplane contained in B, contradicting algebraic openness. Now let W be the space of all vectors w for which there exists a monochromatic line in the direction w (including

0). Then W is easily seen to be a vector space; since A, B are non-empty, W is a proper subspace of V. On the other hand, if w and w' are not in W, some playing around with the property that A and B are convex sets partitioning V shows that the plane spanned by w and w' contains a monochromatic line, and hence some non-trivial linear combination of w and w' lies in W. Thus V/W is precisely one-dimensional. Since every line with direction in w is monochromatic, A and B also have well-defined quotients A/W and B/W on this one-dimensional subspace, which remain convex (with A/W still algebraically open). But then it is clear that A/W and B/W are an open and closed ray from the origin in V/W respectively. It is then a routine matter to construct a linear functional $\lambda : V \to \mathbf{R}$ (with null space W) such that $A = \{\lambda < 0\}$ and $B = \{\lambda \geq 0\}$, and the claim follows.

To establish the general case when A, B are not convex cones, we lift to one higher dimension and apply the previous result to convex cones $A', B' \in \mathbf{R} \times V$ defined by $A' := \{(t, tx) : t > 0, x \in A\}$, $B' := \{(t, tx) : t > 0, x \in B\}$; we leave the verification that this works as an exercise. □

Exercise 1.5.19. Use the geometric Hahn-Banach theorem to reprove Exercise 1.5.18, thus providing a slightly different proof of the Hahn-Banach theorem. (It is possible to reverse these implications and deduce the geometric Hahn-Banach theorem from the usual Hahn-Banach theorem, but this is somewhat trickier, requiring one to fashion a norm out of the difference $A - B$ of two convex cones.)

Exercise 1.5.20 (Algebraic Hahn-Banach theorem)**.** Let V be a vector space over a field F, let W be a subspace of V, and let $\lambda : W \to F$ be a linear map. Show that there exists a linear map $\tilde{\lambda} : V \to F$ which extends λ.

Notes. This lecture first appeared at `terrytao.wordpress.com/2009/01/26`. Thanks to Eric, Xiaochuan Li, and an anonymous commenter for corrections.

Some further discussion of variants of the Hahn-Banach theorem (in the finite-dimensional setting) can be found in Section 1.16 of *Structure and Randomness*.

1.6. A quick review of point set topology

To progress further in our study of function spaces, we will need to develop the standard theory of metric spaces, and of the closely related theory of topological spaces (i.e. point-set topology). I will be assuming that readers will already have encountered these concepts in an undergraduate topology or real analysis course, but for sake of completeness I will briefly review the basics of both spaces here.

1.6.1. Metric spaces. In many spaces, one wants a notion of when two points in the space are "near" or "far". A particularly quantitative and intuitive way to formalise this notion is via the concept of a metric space.

Definition 1.6.1 (Metric spaces)**.** A *metric space* $X = (X, d)$ is a set X, together with a distance function $d : X \times X \to \mathbf{R}^+$ which obeys the following properties:

- (Non-degeneracy) For any $x, y \in X$, we have $d(x, y) \geq 0$, with equality if and only if $x = y$.
- (Symmetry) For any $x, y \in X$, we have $d(x, y) = d(y, x)$.
- (Triangle inequality) For any $x, y, z \in X$, we have $d(x, z) \leq d(x, y) + d(y, z)$.

Example 1.6.2. Every normed vector space $(X, \|\|)$ is a metric space, with distance function $d(x, y) := \|x - y\|$.

Example 1.6.3. Any subset Y of a metric space $X = (X, d)$ is also a metric space $Y = (Y, d \restriction_{Y \times Y})$, where $d \restriction_{Y \times Y} : Y \times Y \to \mathbf{R}^+$ is the restriction of d to $Y \times Y$. We call the metric space $Y = (Y, d \restriction_{Y \times Y})$ a *subspace* of the metric space $X = (X, d)$.

Example 1.6.4. Given two metric spaces $X = (X, d_X)$ and $Y = (Y, d_Y)$, we can define the *product space* $X \times Y = (X \times Y, d_X \times d_Y)$ to be the Cartesian product $X \times Y$ with the product metric

$$d_X \times d_Y((x, y), (x', y')) := \max(d_X(x, x'), d_Y(y, y')). \tag{1.59}$$

(One can also pick slightly different metrics here, such as $d_X(x, x') + d_Y(y, y')$, but this metric only differs from (1.59) by a factor of two, and so they are equivalent (see Example 1.6.11 below).

Example 1.6.5. Any set X can be turned into a metric space by using the discrete metric $d : X \times X \to \mathbf{R}^+$, defined by setting $d(x,y) = 0$ when $x = y$ and $d(x,y) = 1$ otherwise.

Given a metric space, one can then define various useful topological structures. There are two ways to do so. One is via the machinery of convergent sequences:

Definition 1.6.6 (Topology of a metric space)**.** Let (X,d) be a metric space.

- A sequence x_n of points in X is said to *converge* to a *limit* $x \in X$ if one has $d(x_n,x) \to 0$ as $n \to \infty$. In this case, we say that $x_n \to x$ in the metric d as $n \to \infty$, and that $\lim_{n\to\infty} x_n = x$ in the metric space X. (It is easy to see that any sequence of points in a metric space has at most one limit.)
- A point x is an *adherent point* of a set $E \subset X$ if it is the limit of some sequence in E. (This is slightly different from being a *limit point* of E, which is equivalent to being an adherent point of $E \backslash \{x\}$; every adherent point is either a limit point or an *isolated point* of E.) The set of all adherent points of E is called the *closure* $\overline{E}$ of X. A set E is *closed* if it contains all its adherent points, i.e. if $E = \overline{E}$. A set E is *dense* if every point in X is adherent to E, or equivalently if $\overline{E} = X$.
- Given any x in X and $r > 0$, define the *open ball* $B(x,r)$ centred at x with radius r to be the set of all y in X such that $d(x,y) < r$. Given a set E, we say that x is an interior point of E if there is some open ball centred at x which is contained in E. The set of all interior points is called the *interior* E° of E. A set is *open* if every point is an interior point, i.e. if $E = E^\circ$.

There is however an alternate approach to defining these concepts, which takes the concept of an open set as a primitive, rather than the distance function, and defines other terms in terms of open sets. For instance:

Exercise 1.6.1. Let (X, d) be a metric space.

(i) Show that a sequence x_n of points in X converges to a limit $x \in X$ if and only if every open neighbourhood of x (i.e. an open set containing x) contains x_n for all sufficiently large n.

(ii) Show that a point x is an adherent point of a set E if and only if every open neighbourhood of x intersects E.

(iii) Show that a set E is closed if and only if its complement is open.

(iv) Show that the closure of a set E is the intersection of all the closed sets containing E.

(v) Show that a set E is dense if and only if every non-empty open set intersects E.

(vi) Show that the interior of a set E is the union of all the open sets contained in E, and that x is an interior point of E if and only if some neighbourhood of x is contained in E.

In the next section we will adopt this "open sets first" perspective when defining topological spaces.

On the other hand, there are some other properties of subsets of a metric space which require the metric structure more fully, and cannot be defined purely in terms of open sets (see e.g. Example 1.6.24), although some of these concepts can still be defined using a structure intermediate to metric spaces and topological spaces, such as *uniform space*. For instance:

Definition 1.6.7. Let (X, d) be a metric space.

- A sequence $(x_n)_{n=1}^{\infty}$ of points in X is a *Cauchy sequence* if $d(x_n, x_m) \to 0$ as $n, m \to \infty$ (i.e. for every $\varepsilon > 0$ there exists $N > 0$ such that $d(x_n, x_m) \leq \varepsilon$ for all $n, m \geq N$).
- A space X is *complete* if every Cauchy sequence is convergent.
- A set E in X is *bounded* if it is contained inside a ball.
- A set E is *totally bounded* in X if for every $\varepsilon > 0$, E can be covered by finitely many balls of radius ε.

Exercise 1.6.2. Show that any metric space X can be identified with a dense subspace of a complete metric space $\overline{X}$, known as a metric completion or Cauchy completion of X. (For instance, $\mathbf{R}$ is a metric completion of $\mathbf{Q}$.) (*Hint*: one can define a real number to be an equivalence class of Cauchy sequences of rationals. Once the reals are defined, essentially the same construction works in arbitrary metric spaces.) Furthermore, if $\overline{X}'$ is another metric completion of X, show that there exists an isometry between $\overline{X}$ and $\overline{X}'$ which is the identity on X. Thus, up to isometry, there is a unique metric completion to any metric space.

Exercise 1.6.3. Show that a metric space X is complete if and only if it is closed in every superspace Y of X (i.e. in every metric space Y for which X is a subspace). Thus one can think of completeness as being the property of being "absolutely closed".

Exercise 1.6.4. Show that every totally bounded set is also bounded. Conversely, in a Euclidean space $\mathbf{R}^n$ with the usual metric, show that every bounded set is totally bounded. But give an example of a set in a metric space which is bounded but not totally bounded. (*Hint*: use Example 1.6.5.)

Now we come to an important concept.

Theorem 1.6.8 (Heine-Borel theorem for metric spaces)**.** *Let (X, d) be a metric space. Then the following are equivalent:*

(i) *(Sequential compactness) Every sequence in X has a convergent subsequence.*

(ii) *(Compactness) Every open cover $(V_\alpha)_{\alpha \in A}$ of X (i.e. a collection of open sets V_α whose union contains X) has a finite subcover.*

(iii) *(Finite intersection property) If $(F_\alpha)_{\alpha \in A}$ is a collection of closed subsets of X such that any finite subcollection of sets has non-empty intersection, then the entire collection has non-empty intersection.*

(iv) *X is complete and totally bounded.*

Proof. ((ii) $\implies$ (i)) If there was an infinite sequence x_n with no convergent subsequence, then given any point x in X there must exist

an open ball centred at x which contains x_n for only finitely many n (since otherwise one could easily construct a subsequence of x_n converging to n. By (ii), one can cover X with a finite number of such balls. But then the sequence x_n would be finite, a contradiction.

((i) $\implies$ (iv)) If X was not complete, then there would exist a Cauchy sequence which is not convergent; one easily shows that this sequence cannot have any convergent subsequences either, contradicting (i). If X was not totally bounded, then there exists $\varepsilon > 0$ such that X cannot be covered by any finite collection of balls of radius ε; a standard greedy algorithm argument then gives a sequence x_n such that $d(x_n, x_m) \geq \varepsilon$ for all distinct n, m. This sequence clearly has no convergent subsequence, again a contradiction.

((ii) $\iff$ (iii)) This follows from *de Morgan's laws* and Exercise 1.6.1(iii).

((iv) $\implies$ (iii)) Let $(F_\alpha)_{\alpha \in A}$ be as in (iii). Call a set E in X *rich* if it intersects all of the F_α. Observe that if one could cover X by a finite number of non-rich sets, then (as each non-rich set is disjoint from at least one of the F_α), there would be a finite number of F_α whose intersection is empty, a contradiction. Thus, whenever we cover X by finitely many sets, at least one of them must be rich.

As X is totally bounded, for each $n \geq 1$ we can find a finite set $x_{n,1}, \ldots, x_{n,m_n}$ such that the balls $B(x_{n,1}, 2^{-n}), \ldots, B(x_{n,m_n}, 2^{-n})$ cover X. By the previous discussion, we can then find $1 \leq i_n \leq m_n$ such that $B(x_{n,i_n}, 2^{-n})$ is rich.

Call a ball $B(x_{n,i}, 2^{-n})$ asymptotically rich if it contains infinitely many of the x_{j,i_j}. As these balls cover X, we see that for each n, $B(x_{n,i}, 2^{-n})$ is asymptotically rich for at least one i. Furthermore, since each ball of radius 2^{-n} can be covered by balls of radius 2^{-n-1}, we see that if $B(x_{n,j}, 2^{-n})$ is asymptotically rich, then it must intersect an asymptotically rich ball $B(x_{n+1,j'}, 2^{-n-1})$. Iterating this, we can find a sequence $B(x_{n,j_n}, 2^{-n})$ of asymptotically rich balls, each one of which intersects the next one. This implies that x_{n,j_n} is a Cauchy sequence and hence (as X is assumed complete) converges to a limit x. Observe that there exist arbitrarily small rich balls that are arbitrarily close to x, and thus x is adherent to every F_α; since the F_α are closed, we see that x lies in every F_α, and we are done. $\square$

Remark 1.6.9. The hard implication $(iv) \implies (iii)$ of the Heine-Borel theorem is noticeably more complicated than any of the others. This turns out to be unavoidable; this component of the Heine-Borel theorem turns out to be logically equivalent to *König's lemma* in the sense of *reverse mathematics*, and thus cannot be proven in sufficiently weak systems of logical reasoning.

Any space that obeys one of the four equivalent properties in Theorem 1.6.8 is called a *compact space*; a subset E of a metric space X is said to be *compact* if it is a compact space when viewed as a subspace of X. There are some variants of the notion of compactness which are also of importance for us:

- A space is *σ-compact* if it can be expressed as the countable union of compact sets. (For instance, the real line $\mathbf{R}$ with the usual metric is σ-compact.)
- A space is *locally compact* if every point is contained in the interior of a compact set. (For instance, $\mathbf{R}$ is locally compact.)
- A subset of a space is *precompact* or *relatively compact* if it is contained inside a compact set (or equivalently, if its closure is compact).

Another fundamental notion in the subject is that of a *continuous map*.

Exercise 1.6.5. Let $f : X \to Y$ be a map from one metric space (X, d_X) to another (Y, d_Y). Then the following are equivalent:

- (Metric continuity) For every $x \in X$ and $\varepsilon > 0$ there exists $\delta > 0$ such that $d_Y(f(x), f(x')) \leq \varepsilon$ whenever $d_X(x, x') \leq \delta$.
- (Sequential continuity) For every sequence $x_n \in X$ that converges to a limit $x \in X$, $f(x_n)$ converges to f(x).
- (Topological continuity) The inverse image $f^{-1}(V)$ of every open set V in Y, is an open set in X.
- The inverse image $f^{-1}(F)$ of every closed set F in Y, is a closed set in X.

A function f obeying any one of the properties in Exercise 1.6.5 is known as a *continuous map*.

Exercise 1.6.6. Let X, Y, Z be metric spaces, and let $f : X \to Y$ and $g : X \to Z$ be continuous maps. Show that the combined map $f \oplus g : X \to Y \times Z$ defined by $f \oplus g(x) := (f(x), g(x))$ is continuous if and only if f and g are continuous. Show also that the projection maps $\pi_Y : Y \times Z \to Y$, $\pi_Z : Y \times Z \to Z$ defined by $\pi_Y(y, z) := y$, $\pi_Z(y, z) := z$ are continuous.

Exercise 1.6.7. Show that the image of a compact set under a continuous map is again compact.

1.6.2. Topological spaces. Metric spaces capture many of the notions of convergence and continuity that one commonly uses in real analysis, but there are several such notions (e.g. pointwise convergence, semi-continuity, or weak convergence) in the subject that turn out to not be modeled by metric spaces. A very useful framework to handle these more general modes of convergence and continuity is that of a topological space, which one can think of as an abstract generalisation of a metric space in which the metric and balls are forgotten, and the open sets become the central object[9].

Definition 1.6.10 (Topological space)**.** A *topological space* $X = (X, \mathcal{F})$ is a set X, together with a collection $\mathcal{F}$ of subsets of X, known as open sets, which obey the following axioms:

- $\emptyset$ and X are open.
- The intersection of any finite number of open sets is open.
- The union of any arbitrary number of open sets is open.

The collection $\mathcal{F}$ is called a *topology* on X.

Given two topologies $\mathcal{F}, \mathcal{F}'$ on a space X, we say that $\mathcal{F}$ is a *coarser* (or *weaker*) topology than $\mathcal{F}'$ (or equivalently, that $\mathcal{F}'$ is a finer (or stronger) topology than $\mathcal{F}$), if $\mathcal{F} \subset \mathcal{F}'$ (informally, $\mathcal{F}'$ has more open sets than $\mathcal{F}$).

[9]There are even more abstract notions, such as *pointless topological spaces*, in which the collection of open sets has become an abstract lattice, in the spirit of Section 2.3, but we will not need such notions in this course.

Example 1.6.11. Every metric space (X, d) generates a topology $\mathcal{F}_d$, namely the space of sets which are open with respect to the metric d. Observe that if two metrics d, d' on X are equivalent in the sense that

$$cd(x,y) \leq d'(x,y) \leq Cd(x,y) \tag{1.60}$$

for all x, y in X and some constants $c, C > 0$, then they generate identical topologies.

Example 1.6.12. The finest (or strongest) topology on any set X is the *discrete topology* $2^X = \{E : E \subset X\}$, in which every set is open; this is the topology generated by the discrete metric (Example 1.6.5). The coarsest (or weakest) topology is the *trivial topology* $\{\emptyset, X\}$, in which only the empty set and the full set are open.

Example 1.6.13. Given any collection $\mathcal{A}$ of sets of X, we can define the topology $\mathcal{F}[\mathcal{A}]$ *generated by* $\mathcal{A}$ to be the intersection of all the topologies that contain $\mathcal{A}$; this is easily seen to be the coarsest topology that makes all the sets in $\mathcal{A}$ open. For instance, the topology generated by a metric space is the same as the topology generated by its open balls.

Example 1.6.14. If $(X, \mathcal{F})$ is a topological space, and Y is a subset of X, then we can define the *relative topology* $\mathcal{F} \restriction_Y := \{E \cap Y : E \in \mathcal{F}\}$ to be the collection of all open sets in X, restricted to Y, this makes $(Y, \mathcal{F} \restriction_Y)$ a topological space, known as a subspace of $(X, \mathcal{F})$.

Any notion in metric space theory which can be defined purely in terms of open sets, can now be defined for topological spaces. Thus for instance:

Definition 1.6.15. Let $(X, \mathcal{F})$ be a topological space.

- A sequence x_n of points in X converges to a limit $x \in X$ if and only if every open neighbourhood of x (i.e. an open set containing x) contains x_n for all sufficiently large n. In this case we write $x_n \to x$ in the topological space $(X, \mathcal{F})$, and (if x is unique) we write $x = \lim_{n\to\infty} x_n$.
- A point is a *sequentially adherent point* of a set E if it is the limit of some sequence in E.

- A point x is an *adherent point* of a set E if and only if every open neighbourhood of x intersects E.
- The set of all adherent points of E is called the *closure* of E and is denoted $\overline{E}$.
- A set E is *closed* if and only if its complement is open, or equivalently if it contains all its adherent points.
- A set E is *dense* if and only if every non-empty open set intersects E, or equivalently if its closure is X.
- The *interior* of a set E is the union of all the open sets contained in E, and x is called an *interior point* of E if and only if some neighbourhood of x is contained in E.
- A space X is *sequentially compact* if every sequence has a convergent subsequence.
- A space X is *compact* if every open cover has a finite subcover.
- The concepts of being σ-compact, locally compact, and precompact can be defined as before. (One could also define sequential σ-compactness, etc., but these notions are rarely used.)
- A map $f : X \to Y$ between topological spaces is *sequentially continuous* if whenever x_n converges to a limit x in X, $f(x_n)$ converges to a limit f(x) in X.
- A map $f : X \to Y$ between topological spaces is *continuous* if the inverse image of every open set is open.

Remark 1.6.16. The stronger a topology becomes, the more open and closed sets it will have, but fewer sequences will converge, there are fewer (sequentially) adherent points and (sequentially) compact sets, closures become smaller, and interiors become larger. There will be more (sequentially) continuous functions on this space, but fewer (sequentially) continuous functions into the space. Note also that the identity map from a space X with one topology $\mathcal{F}$ to the same space X with a different topology $\mathcal{F}'$ is continuous precisely when $\mathcal{F}$ is stronger than $\mathcal{F}'$.

Example 1.6.17. In a metric space, these topological notions coincide with their metric counterparts, and sequential compactness and compactness are equivalent, as are sequential continuity and continuity.

Exercise 1.6.8 (Urysohn's subsequence principle)**.** Let x_n be a sequence in a topological space X, and let x be another point in X. Show that the following are equivalent:

- x_n converges to x.
- Every subsequence of x_n converges to x.
- Every subsequence of x_n has a further subsequence that converges to x.

Exercise 1.6.9. Show that every sequentially adherent point is an adherent point, and every continuous function is sequentially continuous.

Remark 1.6.18. The converses to Exercise 1.6.9 are unfortunately not always true in general topological spaces. For instance, if we endow an uncountable set X with the *cocountable topology* (so that a set is open if it is either empty, or its complement is at most countable) then we see that the only convergent sequences are those which are eventually constant. Thus, every subset of X contains its sequentially adherent points, and every function from X to another topological space is sequentially continuous, even though not every set in X is closed and not every function on X is continuous. An example of a set which is sequentially compact but not compact is the first uncountable ordinal with the order topology (Exercise 1.6.10). It is more tricky to give an example of a compact space which is not sequentially compact; this will have to wait until we establish Tychonoff's theorem (Theorem 1.8.14). However one can "fix" this discrepancy between the sequential and non-sequential concepts by replacing sequences with the more general notion of nets, see Section 1.6.3.

Remark 1.6.19. Metric space concepts such as boundedness, completeness, Cauchy sequences, and uniform continuity do not have counterparts for general topological spaces, because they cannot be

defined purely in terms of open sets. (They can however be extended to some other types of spaces, such as uniform spaces or coarse spaces.)

Now we give some important topologies that capture certain modes of convergence or continuity that are difficult or impossible to capture using metric spaces alone.

Example 1.6.20 (Zariski topology)**.** This topology is important in algebraic geometry, though it will not be used in this course. If F is an algebraically closed field, we define the *Zariski topology* on the vector space F^n to be the topology generated by the complements of proper algebraic varieties in F^n; thus a set is Zariski open if it is either empty, or is the complement of a finite union of proper algebraic varieties. A set in F^n is then Zariski dense if it is not contained in any proper subvariety, and the Zariski closure of a set is the smallest algebraic variety that contains that set.

Example 1.6.21 (Order topology)**.** Any totally ordered set $(X, <)$ generates the order topology, defined as the topology generated by the sets $\{x \in X : x > a\}$ and $\{x \in X : x < a\}$ for all $a \in X$. In particular, the extended real line $[-\infty, +\infty]$ can be given the order topology, and the notion of convergence of sequences in this topology to either finite or infinite limits is identical to the notion one is accustomed to in undergraduate real analysis. (On the real line, of course, the order topology corresponds to the usual topology.) Also observe that a function $n \mapsto x_n$ from the extended natural numbers $\mathbf{N} \cup \{+\infty\}$ (with the order topology) into a topological space X is continuous if and only if $x_n \to x_{+\infty}$ as $n \to \infty$, so one can interpret convergence of sequences as a special case of continuity.

Exercise 1.6.10. Let ω be the first uncountable ordinal, endowed with the order topology. Show that ω is sequentially compact (*Hint*: every sequence has a lim sup), but not compact (*Hint*: every point has a countable neighbourhood).

Example 1.6.22 (Half-open topology)**.** The right half-open topology $\mathcal{F}_r$ on the real line $\mathbf{R}$ is the topology generated by the right half-open intervals $[a, b)$ for $-\infty < a < b < \infty$; this is a bit finer than the usual topology on $\mathbf{R}$. Observe that a sequence x_n converges to a limit x in

the right half-open topology if and only if it converges in the ordinary topology $\mathcal{F}$, and also if $x_n \geq x$ for all sufficiently large x. Observe that a map $f : \mathbf{R} \to \mathbf{R}$ is right-continuous iff it is a continuous map from $(\mathbf{R}, \mathcal{F}_r)$ to $(\mathbf{R}, \mathcal{F})$. One can of course model left-continuity via a suitable left half-open topology in a similar fashion.

Example 1.6.23 (Upper topology)**.** The upper topology $\mathcal{F}_u$ on the real line is defined as the topology generated by the sets $(a, +\infty)$ for all $a \in \mathbf{R}$. Observe that (somewhat confusingly), a function $f : \mathbf{R} \to \mathbf{R}$ is lower semi-continuous iff it is continuous from $(\mathbf{R}, \mathcal{F})$ to $(\mathbf{R}, \mathcal{F}_u)$. One can of course model upper semi-continuity via a suitable lower topology in a similar fashion.

Example 1.6.24 (Product topology)**.** Let Y^X be the space of all functions $f : X \to Y$ from a set X to a topological space Y. We define the product topology on Y^X to be the topology generated by the sets $\{f \in Y^X : f(x) \in V\}$ for all $x \in X$ and all open $V \subset Y$. Observe that a sequence of functions $f_n : X \to Y$ converges pointwise to a limit $f : X \to Y$ iff it converges in the product topology. We will study the product topology in more depth in Section 1.8.3.

Example 1.6.25 (Product topology, again)**.** If $(X, \mathcal{F}_X)$ and $(Y, \mathcal{F}_Y)$ are two topological spaces, we can define the product space $(X \times Y, \mathcal{F}_X \times \mathcal{F}_Y)$ to be the Cartesian product $X \times Y$ with the topology generated by the product sets $U \times V$, where U and V are open in X and Y respectively. Observe that two functions $f : Z \to X$, $g : Z \to Y$ from a topological space Z are continuous if and only if their direct sum $f : Z \to X \times Y$ is continuous in the product topology, and also that the projection maps $\pi_X : X \times Y \to X$ and $\pi_Y : X \times Y \to Y$ are continuous (cf. Exercise 1.6.6).

We mention that not every topological space can be generated from a metric (such topological spaces are called metrisable). One important obstruction to this arises from the Hausdorff property:

Definition 1.6.26. A topological space X is said to be a Hausdorff space if for any two distinct points x, y in X, there exist disjoint neighbourhoods V_x, V_y of x and y respectively.

Example 1.6.27. Every metric space is Hausdorff (one can use the open balls $B(x, d(x,y)/2)$ and $B(y, d(x,y)/2)$ as the separating neighbourhoods). On the other hand, the trivial topology (Example 1.6.13) on two or more points is not Hausdorff, and neither is the cocountable topology (Remark 1.6.18) on an uncountable set, or the upper topology (Example 1.6.23) on the real line. Thus, these topologies do not arise from a metric.

Exercise 1.6.11. Show that the half-open topology (Example 1.6.22) is Hausdorff, but does not arise from a metric. (*Hint*: assume for contradiction that the half-open topology did arise from a metric; then show that for every real number x there exists a rational number q and a positive integer n such that the ball of radius $1/n$ centred at q has infimum x.) Thus there are more obstructions to metrisability than just the Hausdorff property; a more complete answer is provided by Urysohn's metrisation theorem (Theorem 2.5.7).

Exercise 1.6.12. Show that in a Hausdorff space, any sequence can have at most one limit. (For a more precise statement, see Exercise 1.6.16 below.)

A *homeomorphism* (or *topological isomorphism*) between two topological spaces is a continuous invertible map $f : X \to Y$ whose inverse $f^{-1} : Y \to X$ is also continuous. Such a map identifies the topology on X with the topology on Y, and so any topological concept of X will be preserved by f to the corresponding topological concept of Y. For instance, X is compact if and only if Y is compact, X is Hausdorff if and only if Y is Hausdorff, x is adherent to E if and only if $f(x)$ is adherent to $f(E)$, and so forth. When there is a homeomorphism between two topological spaces, we say that X and Y are *homeomorphic* (or *topologically isomorphic*).

Example 1.6.28. The tangent function is a homeomorphism between $(-\pi/2, \pi/2)$ and $\mathbf{R}$ (with the usual topologies), and thus preserves all topological structures on these two spaces. Note however that the former space is bounded as a metric space while the latter is not, and the latter is complete while the former is not. Thus metric properties such as boundedness or completeness are not purely topological properties, since they are not preserved by homeomorphisms.

1.6.3. Nets (optional). A sequence $(x_n)_{n=1}^\infty$ in a space X can be viewed as a function from the natural numbers $\mathbf{N}$ to X. We can generalise this concept as follows.

Definition 1.6.29 (Nets)**.** A *net* in a space X is a tuple $(x_\alpha)_{\alpha\in A}$, where $A = (A, <)$ is a *directed set* (i.e. a partially ordered set such that any two elements have at least one upper bound), and $x_\alpha \in X$ for each $\alpha \in A$. We say that a statement $P(\alpha)$ holds for sufficiently large α in a directed set A if there exists $\beta \in A$ such that $P(\alpha)$ holds for all $\alpha \geq \beta$. (Note in particular that if $P(\alpha)$ and $Q(\alpha)$ separately hold for sufficiently large α, then their conjunction $P(\alpha) \wedge Q(\alpha)$ also holds for sufficiently large α.)

A net $(x_\alpha)_{\alpha\in A}$ in a topological space X is said to *converge* to a limit $x \in X$ if for every neighbourhood V of x, we have $x_\alpha \in V$ for all sufficiently large α.

A *subnet* of a net $(x_\alpha)_{\alpha\in A}$ is a tuple of the form $(x_{\phi(\beta)})_{\beta\in B}$, where $(B, <)$ is another directed set, and $\phi : B \to A$ is a monotone map (thus $\phi(\beta') \geq \phi(\beta)$ whenever $\beta' \geq \beta$) which is also has *cofinal image*, which means that for any $\alpha \in A$ there exists $\beta \in B$ with $\phi(\beta) \geq \alpha$ (in particular, if $P(\alpha)$ is true for sufficiently large α, then $P(\phi(\beta))$ is true for sufficiently large β).

Remark 1.6.30. Every sequence is a net, but one can create nets that do not arise from sequences (in particular, one can take A to be uncountable). Note a subtlety in the definition of a subnet - we do not require ϕ to be injective, so B can in fact be larger than A! Thus subnets differ a little bit from subsequences in that they "allow repetitions".

Remark 1.6.31. Given a directed set A, one can endow $A \cup \{+\infty\}$ with the upper topology (cf. Example 1.6.23) generated by the sets $[\alpha, +\infty] := \{\beta \in A \cup \{+\infty\} : \beta \geq \alpha\}$ for $\alpha \in A$, with the convention that $+\infty > \alpha$ for all $\alpha \in A$. The property of being directed is precisely saying that these sets form a base. A net $(x_\alpha)_{\alpha\in A}$ converges to a limit $x_{+\infty}$ if and only if the function $\alpha \mapsto x_\alpha$ is continuous on $A \cup \{+\infty\}$ (cf. Example 1.6.21). Also, if $(x_{\phi(\beta)})_{\beta\in B}$ is a subnet of $(x_\alpha)_{\alpha\in A}$, then ϕ is a continuous map from $B \cup \{+\infty\}$ to $A \cup \{+\infty\}$, if

we adopt the convention that $\phi(+\infty) = +\infty$. In particular, a subnet of a convergent net remains convergent to the same limit.

The point of working with nets instead of sequences is that one no longer needs to worry about the distinction between sequential and non-sequential concepts in topology, as the following exercises show:

Exercise 1.6.13. Let X be a topological space, let E be a subset of X, and let x be an element of X. Show that x is an adherent point of E if and only if there exists a net $(x_\alpha)_{\alpha \in A}$ in E that converges to x. (*Hint*: take A to be the directed set of neighbourhoods of x, ordered by reverse set inclusion.)

Exercise 1.6.14. Let $f : X \to Y$ be a map between two topological spaces. Show that f is continuous if and only if for every net $(x_\alpha)_{\alpha \in A}$ in X that converges to a limit x, the net $(f(x_\alpha))_{\alpha \in A}$ converges in Y to $f(x)$.

Exercise 1.6.15. Let X be a topological space. Show that X is compact if and only if every net has a convergent subnet. (*Hint*: equate both properties of X with the finite intersection property, and review the proof of Theorem 1.6.8.) Similarly, show that a subset E of X is relatively compact if and only if every net in E has a subnet that converges in X. (Note that as not every compact space is sequentially compact, this exercise shows that we cannot enforce injectivity of ϕ in the definition of a subnet.)

Exercise 1.6.16. Show that a space is Hausdorff if and only if every net has at most one limit.

Exercise 1.6.17. In the product space Y^X in Example 1.6.24, show that a net $(f_\alpha)_{\alpha \in A}$ converges in Y^X to $f \in Y^X$ if and only if for every $x \in X$, the net $(f_\alpha(x))_{\alpha \in A}$ converges in Y to $f(\alpha)$.

Notes. This lecture first appeared at `terrytao.wordpress.com/2009/01/30`. Thanks to Franciscus Rebro, johan, Josh Zahl, Xiaochuan Liu, and anonymous commenters for corrections.

An anonymous commenter pointed out that while the real line can be viewed very naturally as the metric completion of the rationals, this cannot quite be used to give a definition of the real numbers,

because the notion of a metric itself requires the real numbers in its definition! However, K. P. Hart noted that Bourbaki resolves this problem by defining the reals as the completion of the rationals as a *uniform space* rather than as a metric space.

1.7. The Baire category theorem and its Banach space consequences

The notion of what it means for a subset E of a space X to be "small" varies from context to context. For instance, in measure theory, when $X = (X, \mathcal{X}, \mu)$ is a measure space, one useful notion of a "small" set is that of a null set: a set E of measure zero (or at least contained in a set of measure zero). By countable additivity, countable unions of null sets are null. Taking contrapositives, we obtain

Lemma 1.7.1 (Pigeonhole principle for measure spaces)**.** *Let $E_1, E_2, \ldots$ be an at most countable sequence of measurable subsets of a measure space X. If $\bigcup_n E_n$ has positive measure, then at least one of the E_n has positive measure.*

Now suppose that X was a Euclidean space $\mathbf{R}^d$ with Lebesgue measure m. The *Lebesgue differentiation theorem* easily implies that having positive measure is equivalent to being "dense" in certain balls:

Proposition 1.7.2. *Let E be a measurable subset of $\mathbf{R}^d$. Then the following are equivalent:*

- *E has positive measure.*
- *For any $\varepsilon > 0$, there exists a ball B such that $m(E \cap B) \geq (1 - \varepsilon)m(B)$.*

Thus one can think of a null set as a set which is "nowhere dense" in some measure-theoretic sense.

It turns out that there are analogues of these results when the measure space $X = (X, \mathcal{X}, \mu)$ is replaced instead by a complete metric space $X = (X, d)$. Here, the appropriate notion of a "small" set is not a null set, but rather that of a *nowhere dense set*: a set E which is not dense in any ball, or equivalently a set whose closure has empty interior. (A good example of a *nowhere dense set* would be a proper

subspace, or smooth submanifold, of $\mathbf{R}^d$, or a Cantor set; on the other hand, the rationals are a dense subset of $\mathbf{R}$ and thus clearly not nowhere dense.) We then have the following important result:

Theorem 1.7.3 (Baire category theorem)**.** *Let $E_1, E_2, \ldots$ be an at most countable sequence of subsets of a complete metric space X. If $\bigcup_n E_n$ contains a ball B, then at least one of the E_n is dense in a sub-ball B' of B (and in particular is not nowhere dense). To put it in the contrapositive: the countable union of nowhere dense sets cannot contain a ball.*

Exercise 1.7.1. Show that the Baire category theorem is equivalent to the claim that in a complete metric space, the countable intersection of open dense sets remain dense.

Exercise 1.7.2. Using the Baire category theorem, show that any non-empty complete metric space without isolated points is uncountable. (In particular, this shows that Baire category theorem can fail for incomplete metric spaces such as the rationals $\mathbf{Q}$.)

To quickly illustrate an application of the Baire category theorem, observe that it implies that one cannot cover a finite-dimensional real or complex vector space $\mathbf{R}^n, \mathbf{C}^n$ by a countable number of proper subspaces. One can of course also establish this fact by using Lebesgue measure on this space. However, the advantage of the Baire category approach is that it also works well in infinite dimensional complete normed vector spaces, i.e. Banach spaces, whereas the measure-theoretic approach runs into significant difficulties in infinite dimensions. This leads to three fundamental equivalences between the qualitative theory of continuous linear operators on Banach spaces (e.g. finiteness, surjectivity, etc.) to the quantitative theory (i.e. estimates):

- The *uniform boundedness principle*, that equates the qualitative boundedness (or convergence) of a family of continuous operators with their quantitative boundedness.
- The *open mapping theorem*, that equates the qualitative solvability of a linear problem $Lu = f$ with the quantitative solvability.

- The *closed graph theorem*, that equates the qualitative regularity of a (weakly continuous) operator T with the quantitative regularity of that operator.

Strictly speaking, these theorems are not used much directly in practice, because one usually works in the reverse direction (i.e. first proving quantitative bounds, and then deriving qualitative corollaries); but the above three theorems help explain why we usually approach qualitative problems in functional analysis via their quantitative counterparts.

Let us first prove the Baire category theorem:

Proof of Baire category theorem. Assume that the Baire category theorem failed; then it would be possible to cover a ball $B(x_0, r_0)$ in a complete metric space by a countable family $E_1, E_2, E_3, \ldots$ of nowhere dense sets.

We now invoke the following easy observation: if E is nowhere dense, then every ball B contains a subball B' which is disjoint from E. Indeed, this follows immediately from the definition of a nowhere dense set.

Invoking this observation, we can find a ball $B(x_1, r_1)$ in $B(x_0, r_0/10)$ (say) which is disjoint from E_1; we may also assume that $r_1 \leq r_0/10$ by shrinking r_1 as necessary. Then, inside $B(x_1, r_1/10)$, we can find a ball $B(x_2, r_2)$ which is also disjoint from E_2, with $r_2 \leq r_1/10$. Continuing this process, we end up with a nested sequence of balls $B(x_n, r_n)$, each of which are disjoint from $E_1, \ldots, E_n$, and such that $B(x_n, r_n) \subset B(x_{n-1}, r_{n-1}/10)$ and $r_n \leq r_{n-1}/10$ for all $n = 1, 2, \ldots$.

From the triangle inequality we have $d(x_n, x_{n-1}) \leq 2r_{n-1}/10 \leq 2 \times 10^{-n} r_0$, and so the sequence x_n is a Cauchy sequence. As X is complete, x_n converges to a limit x. Summing the geometric series, one verifies that $x \in B(x_{n-1}, r_{n-1})$ for all $n = 1, 2, \ldots$, and in particular x is an element of B which avoids all of $E_1, E_2, E_3, \ldots$, a contradiction. □

We can illustrate the analogy between the Baire category theorem and the measure-theoretic analogs by introducing some further definitions. Call a set E *meager* or *of the first category* if it can be

expressed (or covered) by a countable union of nowhere dense sets, and *of the second category* if it is not meager. Thus, the Baire category theorem shows that any subset of a complete metric space with non-empty interior is of the second category, which may help explain the name for the property. Call a set *co-meager* or *residual* if its complement is meager, and call a set *Baire* or *almost open* if it differs from an open set by a meager set (note that a Baire set is unrelated to the Baire σ-algebra). Then we have the following analogy between complete metric space topology, and measure theory:

Complete non-empty metric space X	Measure space X of positive measure
first category (meager)	zero measure (null)
second category	positive measure
residual (co-meager)	full measure (co-null)
Baire	measurable

Nowhere dense sets are meager, and meager sets have empty interior. Contrapositively, sets with non-empty interior are residual, and residual sets are somewhere dense. Taking complements instead of contrapositives, we see that open dense sets are co-meager,and co-meager sets are dense.

While there are certainly many analogies between meager sets and null sets (for instance, both classes are closed under countable unions, or under intersections with arbitrary sets), the two concepts can differ in practice. For instance, in the real line $\mathbf{R}$ with the standard metric and measure space structures, the set

$$\bigcup_{n=1}^{\infty} (q_n - 2^{-n}, q_n + 2^{-n}), \tag{1.61}$$

where $q_1, q_2, \ldots$ is an enumeration of the rationals, is open and dense, but has Lebesgue measure at most 2; thus its complement has infinite measure in $\mathbf{R}$ but is nowhere dense (hence meager). As a variant of this, the set

$$\bigcap_{m=1}^{\infty} \bigcup_{n=1}^{\infty} (q_n - 2^{-n}/m, q_n + 2^{-n}/m), \tag{1.62}$$

is a null set, but is the intersection of countably many open dense sets and is thus co-meager.

Exercise 1.7.3. A real number x is *Diophantine* if for every $\varepsilon > 0$ there exists $c > 0$ such that $|x - \frac{a}{q}| \geq \frac{c_\varepsilon}{|q|^{2+\varepsilon}}$ for every rational number $\frac{a}{q}$. Show that the set of Diophantine real numbers has full measure but is meager.

Remark 1.7.4. If one assumes some additional axioms of set theory (e.g. the *continuum hypothesis*), it is possible to show that the collection of meager subsets of $\mathbf{R}$ and the collection of null subsets of $\mathbf{R}$ (viewed as σ-ideals of the collection of all subsets of $\mathbf{R}$) are isomorphic; this is the *Sierpinski-Erdös theorem*, which we will not prove here. Roughly speaking, this theorem tells us that any "effective" first-order statement which is true about meager sets will also be true about null sets, and conversely.

1.7.1. The uniform boundedness principle. As mentioned in the introduction, the Baire category theorem implies various equivalences between qualitative and quantitative properties of linear transformations between Banach spaces. Note that Lemma 1.16 already gave a prototypical such equivalence between a qualitative property (continuity) and a quantitative one (boundedness).

Theorem 1.7.5 (Uniform boundedness principle)**.** *Let X be a Banach space, let Y be a normed vector space, and let $(T_\alpha)_{\alpha \in A}$ be a family of continuous linear operators $T_\alpha : X \to Y$. Then the following are equivalent:*

- *(Pointwise boundedness) For every $x \in X$, the set $\{T_\alpha x : \alpha \in A\}$ is bounded.*
- *(Uniform boundedness) The operator norms $\{\|T_\alpha\|_{\mathrm{op}} : \alpha \in A\}$ are bounded.*

The uniform boundedness principle is also known as the *Banach-Steinhaus theorem*.

Proof. It is clear that (ii) implies (i); now assume (i) holds and let us obtain (ii).

For each $n = 1, 2, \ldots$, let E_n be the set

$$E_n := \{x \in X : \|T_\alpha x\|_Y \leq n \text{ for all } \alpha \in A\}. \tag{1.63}$$

The hypothesis (i) is nothing more than the assertion that the E_n cover X, and thus by the Baire category theorem must be dense in a ball. Since the T_α are continuous, the E_n are closed, and so one of the E_n contains a ball. Since $E_n - E_n \subset E_{2n}$, we see that one of the E_n contains a ball centred at the origin. Dilating n as necessary, we see that one of the E_n contains the unit ball $B(0,1)$. But then all the $\|T_\alpha\|_{\mathrm{op}}$ are bounded by n, and the claim follows. □

Exercise 1.7.4. Give counterexamples to show that the uniform boundedness principle fails one relaxes the assumptions in any of the following ways:

- X is merely a normed vector space rather than a Banach space (i.e. completeness is dropped).
- The T_α are not assumed to be continuous.
- The T_α are allowed to be nonlinear rather than linear.

Thus completeness, continuity, and linearity are all essential for the uniform boundedness principle to apply.

Remark 1.7.6. It is instructive to establish the uniform boundedness principle more "constructively" without the Baire category theorem (though the proof of the Baire category theorem is still implicitly present), as follows. Suppose that (ii) fails, then $\|T_\alpha\|_{\mathrm{op}}$ is unbounded. We can then find a sequence $\alpha_n \in A$ such that $\|T_{\alpha_{n+1}}\|_{\mathrm{op}} > 100^n \|T_{\alpha_n}\|_{\mathrm{op}}$ (say) for all n. We can then find unit vectors x_n such that $\|T_{\alpha_n} x_n\|_Y \geq \frac{1}{2}\|T_{\alpha_n}\|_{\mathrm{op}}$.

We can then form the absolutely convergent (and hence conditionally convergent, by completeness) sum $x = \sum_{n=1}^\infty \epsilon_n 10^{-n} x_n$ for some choice of signs $\epsilon_n = \pm 1$ recursively as follows: once $\epsilon_1, \ldots, \epsilon_{n-1}$ have been chosen, choose the sign ϵ_n so that

$$\| \sum_{m=1}^n \epsilon_n 10^{-n} T_{\alpha_n} x_n \|_Y \geq \|10^{-n} T_{\alpha_n} x_n\|_Y \geq \frac{1}{2} 10^{-n} \|T_{\alpha_n}\|_{\mathrm{op}}. \tag{1.64}$$

From the triangle inequality we soon conclude that

$$\|T_{\alpha_n} x\|_Y \geq \frac{1}{4} 10^{-n} \|T_{\alpha_n}\|_{\mathrm{op}}. \tag{1.65}$$

But by hypothesis, the right-hand side of (1.65) is unbounded in n, contradicting (i).

A common way to apply the uniform boundedness principle is via the following corollary:

Corollary 1.7.7 (Uniform boundedness principle for norm convergence)**.** *Let X be a Banach space, let Y be a normed vector space, and let $(T_n)_{n=1}^\infty$ be a family of continuous linear operators $T_n : X \to Y$. Then the following are equivalent:*

(i) *(Pointwise convergence) For every $x \in X$, $T_n x$ converges strongly in Y as $n \to \infty$.*

(ii) *(Pointwise convergence to a continuous limit) There exists a continuous linear $T : X \to Y$ such that for every $x \in X$, $T_n x$ converges strongly in Y to Tx as $n \to \infty$.*

(iii) *(Uniform boundedness + dense subclass convergence) The operator norms $\{\|T_n\| : n = 1, 2, \ldots\}$ are bounded, and for a dense set of x in X, $T_n x$ converges strongly in Y as $n \to \infty$.*

Proof. Clearly (ii) implies (i), and as convergent sequences are bounded, we see from Theorem 1.7.3 that (i) implies (iii). The implication of (ii) from (iii) follows by a standard limiting argument and is left as an exercise. □

Remark 1.7.8. The same equivalences hold if one replaces the sequence $(T_n)_{n=1}^\infty$ by a net $(T_\alpha)_{\alpha \in A}$.

Example 1.7.9 (Fourier inversion formula)**.** For any $f \in L^2(\mathbf{R})$ and $N > 0$, define the *Dirichlet summation operator*

$$S_N f(x) := \int_{-N}^{N} \hat{f}(\xi) e^{2\pi i x \xi}\, d\xi \tag{1.66}$$

where $\hat{f}$ is the *Fourier transform* of f, defined on smooth compactly supported functions $f \in C_0^\infty(\mathbf{R})$ by the formula $\hat{f}(\xi) := \int_{-\infty}^{\infty} f(x) e^{-2\pi i x \xi}\, dx$ and then extended to L^2 by the *Plancherel theorem* (see Section 1.12). Using the *Plancherel identity*, we can verify that the operator norms $\|S_N\|_{\mathrm{op}}$ are uniformly bounded (indeed, they are all 1); also, one can check that for $f \in C_0^\infty(\mathbf{R})$, that $S_N f$ converges in L^2 norm to f as $N \to \infty$. As $C_0^\infty(\mathbf{R})$ is known to be dense in $L^2(\mathbf{R})$, this implies that $S_N f$ converges in L^2 norm to f for every $f \in L^2(\mathbf{R})$.

This argument only used the "easy" implication of Corollary 1.7.7, namely the deduction of (ii) from (iii). The "hard" implication using the Baire category theorem was not directly utilised. However, from a *metamathematical* standpoint, that implication is important because it tells us that the above strategy to prove convergence in norm of the Fourier inversion formula on L^2 - i.e. to obtain uniform operator norms on the partial sums, and to establish convergence on a dense subclass of "nice" functions - is in some sense the *only* strategy available to prove such a result.

Remark 1.7.10. There is a partial analogue of Corollary 1.7.7 for the question of pointwise almost everywhere convergence rather than norm convergence, known as *Stein's maximal principle* (discussed for instance in Section 1.9 of *Structure and Randomness*). For instance, it reduces *Carleson's theorem* on the pointwise almost everywhere convergence of Fourier series to the boundedness of a certain maximal function (the Carleson maximal operator) related to Fourier summation, although the latter task is again quite non-trivial. (As in Example 1.7.9, the role of the maximal principle is meta-mathematical rather than direct.)

Remark 1.7.11. Of course, if we omit some of the hypotheses, it is no longer true that pointwise boundedness and uniform boundedness are the same. For instance, if we let $c_0(\mathbf{N})$ be the space of complex sequences with only finitely many non-zero entries and with the uniform topology, and let $\lambda_n : c_0(\mathbf{N}) \to \mathbf{C}$ be the map $(a_m)_{m=1}^\infty \to n a_n$, then the λ_n are pointwise bounded but not uniformly bounded; thus completeness of X is important. Also, even in the one-dimensional case $X = Y = \mathbf{R}$, the uniform boundedness principle can easily be seen to fail if the T_α are non-linear transformations rather than linear ones.

1.7.2. The open mapping theorem. A map $f : X \to Y$ between topological spaces X and Y is said to be *open* if it maps open sets to open sets. This is similar to, but slightly different, from the more familiar property of being continuous, which is equivalent to the *inverse* image of open sets being open. For instance, the map $f : \mathbf{R} \to \mathbf{R}$ defined by $f(x) := x^2$ is continuous but not open; conversely, the

function $g : \mathbf{R}^2 \to \mathbf{R}$ defined by $g(x,y) := \mathrm{sgn}(y) + x$ is discontinuous but open.

We have just seen that it is quite possible for non-linear continuous maps to fail to be open. But for linear maps between Banach spaces, the situation is much better:

Theorem 1.7.12 (Open mapping theorem)**.** *Let $L : X \to Y$ be a continuous linear transformation between two Banach spaces X and Y. Then the following are equivalent:*

(i) *L is surjective.*

(ii) *L is open.*

(iii) *(Qualitative solvability) For every $f \in Y$ there exists a solution $u \in X$ to the equation $Lu = f$.*

(iv) *(Quantitative solvability) There exists a constant $C > 0$ such that for every $f \in Y$ there exists a solution $u \in X$ to the equation $Lu = f$, which obeys the bound $\|u\|_X \leq C\|f\|_Y$.*

(v) *(Quantitative solvability for a dense subclass) There exists a constant $C > 0$ such that for a dense set of f in Y, there exists a solution $u \in X$ to the equation $Lu = f$, which obeys the bound $\|u\|_X \leq C\|f\|_Y$.*

Proof. Clearly (iv) implies (iii), which is equivalent to (i), and it is easy to see from linearity that (ii) and (iv) are equivalent (cf. the proof of Lemma 1.3.17). (iv) trivially implies (v), while to conversely obtain (iv) from (v), observe that if E is any dense subset of the Banach space Y, then any f in Y can be expressed as an absolutely convergent series $f = \sum_n f_n$ of elements in E (since one can iteratively approximate the residual $f - \sum_{n=1}^{N-1} f_n$ to arbitrary accuracy by an element of E for $N = 1, 2, 3, \ldots$), and the claim easily follows. So it suffices to show that (iii) implies (iv).

For each n, let $E_n \subset Y$ be the set of all $f \in Y$ for which there exists a solution to $Lu = f$ with $\|u\|_X \leq n\|f\|_Y$. From the hypothesis (iii), we see that $\bigcup_n E_n = Y$. Since Y is complete, the Baire category theorem implies that there is some E_n which is dense in some ball $B(f_0, r)$ in Y. In other words, the problem $Lu = f$ is approximately quantitatively solvable in the ball $B(f_0, r)$ in the sense that for every

$\varepsilon > 0$ and every $f \in B(f_0, r)$, there exists an approximate solution u with $\|Lu - f\|_Y \leq \varepsilon$ and $\|u\|_X \leq n\|Lu\|_Y$, and thus $\|u\|_X \leq nr + n\varepsilon$.

By subtracting two such approximate solutions, we conclude that for any $f \in B(0, 2r)$ and any $\varepsilon > 0$, there exists $u \in X$ with $\|Lu - f\|_Y \leq 2\varepsilon$ and $\|u\|_X \leq 2nr + 2n\varepsilon$.

Since L is homogeneous, we can rescale and conclude that for any $f \in Y$ and any $\varepsilon > 0$ there exists $u \in X$ with $\|Lu - f\|_Y \leq 2\varepsilon$ and $\|u\|_X \leq 2n\|f\|_Y + 2n\varepsilon$.

In particular, setting $\varepsilon = \frac{1}{4}\|f\|_Y$ (treating the case $f = 0$ separately), we conclude that for any $f \in Y$, we may write $f = Lu + f'$, where $\|f'\|_Y \leq \frac{1}{2}\|f\|_Y$ and $\|u\|_X \leq \frac{5}{2}n\|f\|_Y$.

We can iterate this procedure and then take limits (now using the completeness of X rather than Y) to obtain a solution to $Lu = f$ for every $f \in Y$ with $\|u\|_X \leq 5n\|f\|_Y$, and the claim follows. □

Remark 1.7.13. The open mapping theorem provides metamathematical justification for the *method of a priori estimates* for solving linear equations such as $Lu = f$ for a given datum $f \in Y$ and for an unknown $u \in X$, which is of course a familiar problem in linear PDE. The *a priori* method assumes that f is in some dense class of nice functions (e.g. smooth functions) in which solvability of $Lu = f$ is presumably easy, and then proceeds to obtain the *a priori* estimate $\|u\|_X \leq C\|f\|_Y$ for some constant Y. Theorem 1.7.12 then assures that $Lu = f$ is solvable for all f in Y (with a similar bound). As before, this implication does not directly use the Baire category theorem, but that theorem helps explain why this method is "not wasteful".

A pleasant corollary of the open mapping theorem is that, as with ordinary linear algebra or with arbitrary functions, invertibility is the same thing as bijectivity:

Corollary 1.7.14. *Let $T : X \to Y$ be a continuous linear operator between two Banach spaces X, Y. Then the following are equivalent:*

- *(Qualitative invertibility) T is bijective.*
- *(Quantitative invertibility) T is bijective, and $T^{-1} : Y \to X$ is a continuous (hence bounded) linear transformation.*

Remark 1.7.15. The claim fails without the completeness hypotheses on X and Y. For instance, consider the operator $T : c_c(\mathbf{N}) \to c_c(\mathbf{N})$ defined by $T(a_n)_{n=1}^\infty := (\frac{a_n}{n})_{n=1}^\infty$, where we give $c_c(\mathbf{N})$ the uniform norm. Then T is continuous and bijective, but T^{-1} is unbounded.

Exercise 1.7.5. Show that Corollary 1.7.14 can still fail if we drop the completeness hypothesis on just X, or just Y.

Exercise 1.7.6. Suppose that $L : X \to Y$ is a surjective continuous linear transformation between Banach spaces. By combining the open mapping theorem with the Hahn-Banach theorem, show that the transpose map $L^* : Y^* \to X^*$ is bounded from below, i.e. there exists $c > 0$ such that $\|L^*\lambda\|_{X^*} \geq c\|\lambda\|_{Y^*}$ for all $\lambda \in Y^*$. Conclude that L^* is an isomorphism between Y^* and $L^*(Y^*)$.

Let L be as in Theorem 1.7.12, so that the problem $Lu = f$ is both qualitatively and quantitatively solvable. A standard application of Zorn's lemma (similar to that used to prove the Hahn-Banach theorem) shows that the problem $Lu = f$ is also qualitatively linearly solvable, in the sense that there exists a linear transformation $S : Y \to X$ such that $LSf = f$ for all $f \in Y$ (i.e. S is a right-inverse of L). In view of the open mapping theorem, it is then tempting to conjecture that L must also be quantitatively linearly solvable, in the sense that there exists a continuous linear transformation $S : Y \to X$ such that $LSf = f$ for all $f \in Y$. By Corollary 1.7.14, we see that this conjecture is true when the problem $Lu = f$ is *determined*, i.e. there is exactly one solution u for each datum f. Unfortunately, the conjecture can fail when $Lu = f$ is underdetermined (more than one solution u for each f); we discuss this in Section 1.7.4. On the other hand, the situation is much better for Hilbert spaces:

Exercise 1.7.7. Suppose that $L : H \to H'$ is a surjective continuous linear transformation between Hilbert spaces. Show that there exists a continuous linear transformation $S : H' \to H$ such that $LS = I$. Furthermore, show that we can ensure that the range of S is orthogonal to the kernel of L, and that this condition determines S uniquely.

Remark 1.7.16. In fact, Hilbert spaces are essentially the only type of Banach space for which we have this nice property, due to the Lindenstrauss-Tzafriri solution [**LiTz1971**] of the complemented subspaces problem.

Exercise 1.7.8. Let M and N be closed subspaces of a Banach space X. Show that the following statements are equivalent:

(i) (Qualitative complementation) Every x in X can be expressed in the form $m+n$ for $m \in M, n \in N$ in exactly one way.

(ii) (Quantitative complementation) Every x in X can be expressed in the form $m+n$ for $m \in M, n \in N$ in exactly one way. Furthermore there exists $C > 0$ such that $\|m\|_X, \|n\|_X \leq C\|x\|_X$ for all x.

When either of these two properties hold, we say that M (or N) is a *complemented subspace*, and that N is a *complement* of M (or vice versa).

The property of being complemented is closely related to that of quantitative linear solvability:

Exercise 1.7.9. Let $L : X \to Y$ be a surjective map between Banach spaces. Show that there exists a bounded linear map $S : Y \to X$ such that $LSf = f$ for all $f \in Y$ if and only if the kernel $\{u \in X : Lu = 0\}$ is a complemented subspace of X.

Exercise 1.7.10. Show that any finite-dimensional or finite co-dimensional subspace of a Banach space is complemented.

Remark 1.7.17. The problem of determining whether a given closed subspace of a Banach space is complemented or not is, in general, quite difficult. However, non-complemented subspaces do exist in abundance; some example are given in the apendix, and the Lindenstrauss-Tzafriri theorem [**LiTz1971**] asserts that any Banach space not isomorphic to a Hilbert space contains at least one non-complemented subspace. There is also a remarkable construction of Gowers and Maurey [**Go1993**] of a Banach space such that every subspace, other than those ruled out by Exercise 1.7.10, are uncomplemented.

1.7.3. The closed graph theorem. Recall that a map $T : X \to Y$ between two metric spaces is continuous if and only if, whenever x_n converges to x in X, Tx_n converges to Tx in Y. We can also define the weaker property of being *closed*: an map $T : X \to Y$ is closed if and only if whenever x_n converges to x in X, and Tx_n converges to a limit y in Y, then y is equal to Tx; equivalently, T is closed if its graph $\{(x, Tx) : x \in X\}$ is a closed subset of $X \times Y$. This is weaker than continuity because it has the additional requirement that the sequence Tx_n is already convergent. the name, closed operators are not directly related to open operators.)

Example 1.7.18. Let $T : c_0(\mathbf{N}) \to c_0(\mathbf{N})$ be the transformation $T(a_m)_{m=1}^\infty := (ma_m)_{m=1}^\infty$. This transformation is unbounded and hence discontinuous, but one easily verifies that it is closed.

As Example 1.7.18 shows, being closed is often a weaker property than being continuous. However, the remarkable *closed graph theorem* shows that as long as the domain and range of the operator are both Banach spaces, the two statements are equivalent:

Theorem 1.7.19 (Closed graph theorem)**.** *Let $T : X \to Y$ be a linear transformation between two Banach spaces. Then the following are equivalent:*

(i) *T is continuous.*

(ii) *T is closed.*

(iii) *(Weak continuity) There exists some topology $\mathcal{F}$ on Y, weaker than the norm topology (i.e. containing fewer open sets) but still Hausdorff, for which $T : X \to (Y, \mathcal{F})$ is continuous.*

Proof. It is clear that (i) implies (iii) (just take $\mathcal{F}$ to equal the norm topology). To see why (iii) implies (ii), observe that if $x_n \to x$ in X and $Tx_n \to y$ in norm, then $Tx_n \to y$ in the weaker topology $\mathcal{F}$ as well; but by weak continuity $Tx_n \to Tx$ in $\mathcal{F}$. Since Hausdorff topological spaces have unique limits, we have $Tx = y$ and so T is closed.

Now we show that (ii) implies (i). If T is closed, then the graph $\Gamma := \{(x, Tx) : x \in X\}$ is a closed linear subspace of the Banach space $X \times Y$ and is thus also a Banach space. On the other hand, the

projection map $\pi : (x, Tx) \mapsto x$ from Γ to X is clearly a continuous linear bijection. By Corollary 1.7.14, its inverse $x \mapsto (x, Tx)$ is also continuous, and so T is continuous as desired. □

We can reformulate the closed graph theorem in the following fashion:

Corollary 1.7.20. *Let X, Y be Banach spaces, and suppose we have some continuous inclusion $Y \subset Z$ of Y into a Hausdorff topological vector space Z. Let $T : X \to Z$ be a continuous linear transformation. Then the following are equivalent.*

(i) *(Qualitative regularity) For all $x \in X$, $Tx \in Y$.*

(ii) *(Quantitative regularity) For all $x \in X$, $Tx \in Y$, and furthermore $\|Tx\|_Y \leq C\|x\|_X$ for some $C > 0$ independent of x.*

(iii) *(Quantitative regularity on a dense subclass) For all x in a dense subset of X, $Tx \in Y$, and furthermore $\|Tx\|_Y \leq C\|x\|_X$ for some $C > 0$ independent of x.*

Proof. Clearly (ii) implies (iii) or (i). If we have (iii), then T extends uniquely to a bounded linear map from X to Y, which must agree with the original continuous map from X to Z since limits in the Hausdorff space Z are unique, and so (iii) implies (ii). Finally, if (i) holds, then we can view T as a map from X to Y, which by Theorem 1.7.19 is continuous, and the claim now follows from Lemma 1.3.17. □

In practice, one should think of Z as some sort of "low regularity" space with a weak topology, and Y as a "high regularity" subspace with a stronger topology. Corollary 1.7.20 motivates the *method of a priori estimates* to establish the Y-regularity of some linear transform Tx of an arbitrary element x in a Banach space X, by first establishing the *a priori estimate* $\|Tx\|_Y \leq C\|x\|_X$ for a dense subclass of "nice" elements of X, and then using the above corollary (and some weak continuity of T in a low regularity space) to conclude. The closed graph theorem provides the metamathematical explanation as to why this approach is at least as powerful as any other approach to proving regularity.

Example 1.7.21. Let $1 \leq p \leq 2$, and let p' be the dual exponent of p. To prove that the Fourier transform $\hat{f}$ of a function $f \in L^p(\mathbf{R})$ necessarily lies in $L^{p'}(\mathbf{R})$, it suffices to prove the *Hausdorff-Young inequality*

$$\|\hat{f}\|_{L^{p'}(\mathbf{R})} \leq C_p \|f\|_{L^p(\mathbf{R})} \tag{1.67}$$

for some constant C_p and all f in some suitable dense subclass of $L^p(\mathbf{R})$ (e.g. the space $C_0^\infty(\mathbf{R})$ of smooth functions of compact support), together with the "soft" observation that the Fourier transform is continuous from $L^p(\mathbf{R})$ to the space of tempered distributions, which is a Hausdorff space into which $L^{p'}(\mathbf{R})$ embeds continuously. (We will prove this inequality in (1.103).) One can replace the Hausdorff-Young inequality here by countless other estimates in harmonic analysis to obtain similar qualitative regularity conclusions.

1.7.4. Nonlinear solvability (optional). In this appendix we give an example of a linear equations $Lu = f$ which can only be quantitatively solved in a nonlinear fashion. We will use a number of basic tools which we will only cover later in this course, and so this material is optional reading.

Let $X = \{0,1\}^{\mathbf{N}}$ be the infinite discrete cube with the product topology; by Tychonoff's theorem (Theorem 1.8.14), this is a compact Hausdorff space. The Borel σ-algebra is generated by the cylinder sets

$$E_n := \{(x_m)_{m=1}^\infty \in \{0,1\}^{\mathbf{N}} : x_n = 1\}. \tag{1.68}$$

(From a probabilistic view point, one can think of X as the event space for flipping a countably infinite number of coins, and E_n as the event that the n^{th} coin lands as heads.)

Let $M(X)$ be the space of finite Borel measures on X; this can be verified to be a Banach space. There is a map $L : M(X) \to \ell^\infty(\mathbf{N})$ defined by

$$L(\mu) := (\mu(E_n))_{n=1}^\infty. \tag{1.69}$$

This is a continuous linear transformation. The equation $Lu = f$ is quantitatively solvable for every $f \in \ell^\infty(\mathbf{N})$. Indeed, if f is an indicator function $f = 1_A$, then $f = L\delta_{x_A}$, where $x_A \in \{0,1\}^{\mathbf{Z}}$ is the sequence that equals 1 on A and 0 outside of A, and δ_{x_A} is the Dirac

mass at A. The general case then follows by expressing a bounded sequence as an integral of indicator functions (e.g. if f takes values in [0,1], we can write $f = \int_0^1 1_{\{f>t\}}\ dt$). Note however that this is a nonlinear operation, since the indicator $1_{\{f>t\}}$ depends nonlinearly on f.

We now claim that the equation $Lu = f$ is not quantitatively linearly solvable, i.e. there is no bounded linear map $S : \ell^\infty(\mathbf{N}) \to M(X)$ such that $LSf = f$ for all $f \in \ell^\infty(\mathbf{N})$. This fact was first observed by Banach and Mazur; we shall give two proofs, one of a "soft analysis" flavour and one of a "hard analysis" flavour.

We begin with the "soft analysis" proof, starting with a measure-theoretic result which is of independent interest.

Theorem 1.7.22 (Nikodym convergence theorem)**.** *Let $(X, \mathcal{B})$ be a measurable space, and let $\sigma_n : \mathcal{B} \to \mathbf{R}$ be a sequence of signed finite measures which is weakly convergent in the sense that $\sigma_n(E)$ converges to some limit $\sigma(E)$ for each $E \in \mathcal{B}$.*

- *The σ_n are* uniformly countably additive*, which means that for any sequence $E_1, E_2, \ldots$ of disjoint measurable sets, the series $\sum_{m=1}^\infty |\sigma_n(E_m)|$ converges uniformly in n.*
- *σ is a signed finite measure.*

Proof. It suffices to prove the first claim, since this easily implies that σ is also countably additive, and is thence a signed finite measure. Suppose for contradiction that the claim failed, then one could find disjoint $E_1, E_2, \ldots$ and $\varepsilon > 0$ such that one has $\limsup_{n\to\infty} \sum_{m=M}^\infty |\sigma_n(E_m)| > \varepsilon$ for all M. We now construct disjoint sets $A_1, A_2, \ldots$, each consisting of the union of a finite collection of the E_j, and an increasing sequence $n_1, n_2, \ldots$ of positive integers, by the following recursive procedure:

0. Initialise $k = 0$.
1. Suppose recursively that $n_1 < \ldots < n_{2k}$ and $A_1, \ldots, A_k$ has already been constructed for some $k \geq 0$.
2. Choose $n_{2k+1} > n_{2k}$ so large that for all $n \geq n_{2k+1}$, $\mu_n(A_1 \cup \ldots \cup A_k)$ differs from $\mu(A_1 \cup \ldots \cup A_k)$ by at most $\varepsilon/10$.

3. Choose M_k so large that M_k is larger than j for any $E_j \subset A_1 \cup \ldots \cup A_k$, and such that $\sum_{m=M_k}^{\infty} |\mu_{n_j}(E_m)| \leq \varepsilon/100^{k+1}$ for all $1 \leq j \leq 2k+1$.
4. Choose $n_{2k+2} > n_{2k+1}$ so that $\sum_{m=M_k}^{\infty} |\mu_{n_{2k+2}}(E_m)| > \varepsilon$.
5. Pick A_{k+1} to be a finite union of the E_j with $j \geq M_k$ such that $|\mu_{n_{2k+2}}(A_{k+1})| > \varepsilon/2$.
6. Increment k to $k+1$ and then return to Step 2.

It is then a routine matter to show that if $A := \bigcup_{j=1}^{\infty} A_j$, then $|\mu_{2k+2}(A) - \mu_{2k+1}(A)| \geq \varepsilon/10$ for all j, contradicting the hypothesis that μ_j is weakly convergent to μ. □

Exercise 1.7.11 (Schur's property for ℓ^1)**.** Show that if a sequence in $\ell^1(\mathbf{N})$ is convergent in the weak topology, then it is convergent in the strong topology.

We return now to the map $S : \ell^\infty(\mathbf{N}) \to M(X)$. Consider the sequence $a_n \in c_0(\mathbf{N}) \subset \ell^\infty$ defined by $a_n := (1_{m \leq n})_{m=1}^{\infty}$, i.e. each a_n is the sequence consisting of n 1's followed by an infinite number of 0's. As the dual of $c_0(\mathbf{N})$ is isomorphic to $\ell^1(\mathbf{N})$, we see from the dominated convergence theorem that a_n is a weakly Cauchy sequence in $c_0(\mathbf{N})$, in the sense that $\lambda(a_n)$ is Cauchy for any $\lambda \in c_0(\mathbf{N})^*$. Applying S, we conclude that $S(a_n)$ is weakly Cauchy in $M(X)$. In particular, using the bounded linear functionals $\mu \mapsto \mu(E)$ on $M(X)$, we see that $S(a_n)(E)$ converges to some limit $\mu(E)$ for all measurable sets E. Applying the Nikodym convergence theorem we see that μ is also a signed finite measure. We then see that $S(a_n)$ converges in the weak topology to μ. (One way to see this is to define $\nu := \sum_{n=1}^{\infty} 2^{-n}|S(a_n)| + |\mu|$, then ν is finite and $S(a_n), \mu$ are all absolutely continuous with respect to ν; now use the Radon-Nikodym theorem (see Section 1.2) and the fact that $L^1(\nu)^* \equiv L^\infty(\nu)$.) On the other hand, as $LS = I$ and L and S are both bounded, S is a Banach space isomorphism between c_0 and $S(c_0)$. Thus $S(c_0)$ is complete, hence closed, hence weakly closed (by the Hahn-Banach theorem), and so $\mu = S(a)$ for some $a \in c_0$. By the Hahn-Banach theorem again, this implies that a_n converges weakly to $a \in c_0$. But this is easily seen to be impossible, since the constant sequence $(1)_{m=1}^{\infty}$ does not lie in c_0, and the claim follows.

Now we give the "hard analysis" proof. Let $e_1, e_2, \ldots$ be the standard basis for $\ell^\infty(\mathbf{N})$, let N be a large number, and consider the random sums

$$S(\varepsilon_1 e_1 + \ldots + \varepsilon_N e_N) \tag{1.70}$$

where $\varepsilon_n \in \{-1, 1\}$ are iid random signs. Since the ℓ^∞ norm of $\varepsilon_1 e_1 + \ldots + \varepsilon_N e_N$ is 1, we have

$$\|S(\varepsilon_1 e_1 + \ldots + \varepsilon_N e_N)\|_{M(X)} \leq C \tag{1.71}$$

for some constant C independent of N. On the other hand, we can write $S(e_n) = f_n \nu$ for some finite measure ν and some $f_n \in L^1(\nu)$ using Radon-Nikodym as in the previous proof, and then

$$\|\varepsilon_1 f_1 + \ldots + \varepsilon_N f_N\|_{L^1(\nu)} \leq C. \tag{1.72}$$

Taking expectations and applying *Khintchine's inequality* we conclude

$$\|(\sum_{n=1}^N |f_n|^2)^{1/2}\|_{L^1(\nu)} \leq C' \tag{1.73}$$

for some constant C' independent of N. By Cauchy-Schwarz, this implies that

$$\|\sum_{n=1}^N |f_n|\|_{L^1(\nu)} \leq C'\sqrt{N}. \tag{1.74}$$

But as $\|f_n\|_{L^1(\nu)} = \|S(e_n)\|_{M(X)} \geq c$ for some constant $c > 0$ independent of N, we obtain a contradiction for N large enough, and the claim follows.

Remark 1.7.23. The phenomenon of nonlinear quantitative solvability actually comes up in many applications of interest. For instance, consider the Fefferman-Stein decomposition theorem[**FeSt1972**], which asserts that any $f \in BMO(\mathbf{R})$ of *bounded mean oscillation* can be decomposed as $f = g + Hh$ for some $g, h \in L^\infty(\mathbf{R})$, where H is the *Hilbert transform*. This theorem was first proven by using the duality of the *Hardy space* $H^1(\mathbf{R})$ and BMO (and by using Exercise 1.5.13), and by using the fact that a function f is in $H^1(\mathbf{R})$ if and only if f and Hf both lie in $L^1(\mathbf{R})$. From the open mapping theorem we know that we can pick g, h so that the L^∞ norms of g, h are bounded by a

multiple of the BMO norm of f. But it turns out not to be possible to pick g and h in a bounded linear manner in terms of f, although this is a little tricky to prove. (Uchiyama[**Uc1982**] famously gave an explicit construction of g, h in terms of f, but the construction was highly nonlinear.)

An example in a similar spirit was given more recently by Bourgain and Brezis[**BoBr2003**], who considered the problem of solving the equation $\operatorname{div} u = f$ on the d-dimensional torus $\mathbf{T}^d$ for some function $f : \mathbf{T}^d \to \mathbf{C}$ on the torus with mean zero, and with some unknown vector field $u : \mathbf{T}^d \to \mathbf{C}^d$, where the derivatives are interpreted in the weak sense. They showed that if $d \geq 2$ and $f \in L^d(\mathbf{T}^d)$, then there existed a solution u to this problem with $u \in W^{1,d} \cap C^0$, despite the failure of *Sobolev embedding* at this endpoint. Again, the open mapping theorem allows one to choose u with norm bounded by a multiple of the norm of f, but Bourgain and Brezis also show that one cannot select u in a bounded *linear* fashion depending on f.

Notes. This lecture first appeared at `terrytao.wordpress.com/2009/02/01`. Thanks to Achille Talon, Phi. Isett, Ulrich, Xiaochuan Liu, and anonymous commenters for corrections.

Let me close with a question. All of the above constructions of non-complemented closed subspaces, or of linear problems that can only be quantitatively solved nonlinearly, were quite involved. Is there a "soft" or "elementary" way to see that closed subspaces of Banach spaces exist which are not complemented, or (equivalently) that surjective continuous linear maps between Banach spaces do not always enjoy a continuous linear right-inverse? I do not have a good answer to this question.

1.8. Compactness in topological spaces

One of the most useful concepts for analysis that arise from topology and metric spaces is the concept of *compactness*. Recall (from Section 1.6) that a space X is compact if every open cover of X has a finite subcover, or equivalently if any collection of closed sets whose finite subcollections have non-empty intersection), has overall non-empty

intersection. (In other words, all families of closed sets obey the *finite intersection property*.)

In these notes, we explore how compactness interacts with other key topological concepts: the *Hausdorff property*, *bases* and *sub-bases*, *product spaces*, and *equicontinuity*, in particular establishing the useful *Tychonoff* and *Arzelá-Ascoli theorems* that give criteria for compactness (or *precompactness*).

Exercise 1.8.1 (Basic properties of compact sets)**.**

- Show that any finite set is compact.
- Show that any finite union of compact subsets of a topological space is still compact.
- Show that any image of a compact space under a continuous map is still compact.

Show that these three statements continue to hold if "compact" is replaced by "sequentially compact".

1.8.1. Compactness and the Hausdorff property. Recall from Section 1.6 that a topological space is *Hausdorff* if every distinct pair x, y of points can be separated by two disjoint open neighbourhoods U, V of x, y respectively; every metric space is Hausdorff, but not every topological space is.

At first glance, the Hausdorff property bears no resemblance to the compactness property. However, they are in some sense "dual" to each other, as the following two exercises show:

Exercise 1.8.2. Let $X = (X, \mathcal{F})$ be a compact topological space.

- Show that every closed subset in X is compact.
- Show that any weaker topology $\mathcal{F}' \subset \mathcal{F}$ on X also yields a compact topological space $(X, \mathcal{F}')$.
- Show that the *trivial topology* on X is always compact.

Exercise 1.8.3. Let X be a Hausdorff topological space.

- Show that every compact subset of X is closed.
- Show that any stronger topology $\mathcal{F}' \supset \mathcal{F}$ on X also yields a Hausdorff topological space $(X, \mathcal{F}')$.

- Show that the *discrete topology* on X is always Hausdorff.

The first exercise asserts that compact topologies tend to be weak, while the second exercise asserts that Hausdorff topologies tend to be strong. The next lemma asserts that the two concepts only barely overlap:

Lemma 1.8.1. *Let $\mathcal{F} \subset \mathcal{F}'$ be a weak and strong topology respectively on a space X. If $\mathcal{F}'$ is compact and $\mathcal{F}$ is Hausdorff, then $\mathcal{F} = \mathcal{F}'$. (In other words, a compact topology cannot be strictly stronger than a Hausdorff one, and a Hausdorff topology cannot be strictly weaker than a compact one.)*

Proof. Since $\mathcal{F} \subset \mathcal{F}'$, every set which is closed in $(X, \mathcal{F})$ is closed in $(X, \mathcal{F}')$, and every set which is compact in $(X, \mathcal{F}')$ is compact in $(X, \mathcal{F})$. But from Exercises 1.8.2, 1.8.3, every set which is closed in $(X, \mathcal{F}')$ is compact in $(X, \mathcal{F}')$, and every set which is compact in $(X, \mathcal{F})$ is closed in $(X, \mathcal{F})$. Putting all this together, we see that $(X, \mathcal{F})$ and $(X, \mathcal{F}')$ have exactly the same closed sets, and thus have exactly the same open sets; in other words, $\mathcal{F} = \mathcal{F}'$. □

Corollary 1.8.2. *Any continuous bijection $f : X \to Y$ from a compact topological space $(X, \mathcal{F}_X)$ to a Hausdorff topological space $(Y, \mathcal{F}_Y)$ is a homeomorphism.*

Proof. Consider the *pullback* $f^\#(\mathcal{F}_Y) := \{f^{-1}(U) : U \in \mathcal{F}_Y\}$ of the topology on Y by f; this is a topology on X. As f is continuous, this topology is weaker than $\mathcal{F}_X$, and thus by Lemma 1.8.1 is equal to $\mathcal{F}_X$. As f is a bijection, this implies that f^{-1} is continuous, and the claim follows. □

One may wish to compare this corollary with Corollary 1.7.14.

Remark 1.8.3. Spaces which are both compact and Hausdorff (e.g. the unit interval $[0, 1]$ with the usual topology) have many nice properties and are moderately common, so much so that the two properties are often concatenated as *CH*. Spaces that are *locally* compact and Hausdorff (e.g. manifolds) are much more common and have nearly as many nice properties, and so these two properties are often concatenated as *LCH*. One should caution that (somewhat confusingly)

in some older literature (particularly those in the French tradition), "compact" is used for "compact Hausdorff".

(Optional) Another way to contrast compactness and the Hausdorff property is via the machinery of *ultrafilters*. Define an *filter* on a space X to be a collection p of sets of 2^X which is closed under finite intersection, is also monotone (i.e. if $E \in p$ and $E \subset F \subset X$, then $F \in p$), and does not contain the empty set. Define an *ultrafilter* to be a filter with the additional property that for any $E \in X$, exactly one of E and $X \backslash E$ lies in p. (See also Section 1.5 of *Structure and Randomness*.)

Exercise 1.8.4 (Ultrafilter lemma)**.** Show that every filter is contained in at least one ultrafilter. (*Hint*: use Zorn's lemma, see Section 2.4.)

Exercise 1.8.5. A collection of subsets of X has the *finite intersection property* if every finite intersection of sets in the collection has non-empty intersection. Show that every filter has the finite intersection property, and that every collection of sets with the finite intersection property is contained in a filter (and hence contained in an ultrafilter, by the ultrafilter lemma).

Given a point $x \in X$ and an ultrafilter p on X, we say that p *converges* to x if every neighbourhood of x belongs to p.

Exercise 1.8.6. Show that a space X is Hausdorff if and only if every ultrafilter has at most one limit. (*Hint*: For the "if" part, observe that if x, y cannot be separated by disjoint neighbourhoods, then the neighbourhoods of x and y together enjoy the finite intersection property.)

Exercise 1.8.7. Show that a space X is compact if and only if every ultrafilter has at least one limit. (*Hint*: use the finite intersection property formulation of compactness and Exercise 1.8.5.)

1.8.2. Compactness and bases. Compactness is the property that every open cover has a finite subcover. This property can be difficult to verify in practice, in part because the class of open sets is very large. However, in many cases one can replace the class of open sets

with a much smaller class of sets. For instance, in metric spaces, a set is open if and only if it is the union of open balls (note that the union may be infinite or even uncountable). We can generalise this notion as follows:

Definition 1.8.4 (Base)**.** Let $(X, \mathcal{F})$ be a topological space. A *base* for this space is a collection $\mathcal{B}$ of open sets such that every open set in X can be expressed as the union of sets in the base. The elements of $\mathcal{B}$ are referred to as *basic open sets*.

Example 1.8.5. The collection of open balls $B(x, r)$ in a metric space forms a base for the topology of that space. As another (rather trivial) example of a base: any topology $\mathcal{F}$ is a base for itself.

This concept should be compared with that of a *basis* of a vector space: every vector in that space can be expressed as a linear combination of vectors in a basis. However, one difference between a base and a basis is that the representation of an open set as the union of basic open sets is almost certainly not unique.

Given a base $\mathcal{B}$, define a *basic open neighbourhood* of a point $x \in X$ to be a basic open set that contains x. Observe that a set U is open if and only if every point in U has a basic open neighbourhood contained in U.

Exercise 1.8.8. Let $\mathcal{B}$ be a collection of subsets of a set X. Show that $\mathcal{B}$ is a basis for some topology $\mathcal{F}$ if and only if it it covers X and has the following additional property: given any $x \in X$ and any two basic open neighbourhoods U, V of x, there exists another basic open neighbourhood W of x that is contained in $U \cap V$. Furthermore, the topology $\mathcal{F}$ is uniquely determined by $\mathcal{B}$.

To verify the compactness property, it suffices to do so for basic open covers (i.e. coverings of the whole space by basic open sets):

Exercise 1.8.9. Let $(X, \mathcal{F})$ be a topological space with a base $\mathcal{B}$. Then the following are equivalent:

- Every open cover has a finite subcover (i.e. X is compact);
- Every basic open cover has a finite subcover.

A useful fact about compact metric spaces is that they are in some sense "countably generated".

Lemma 1.8.6. *Let $X = (X, d_X)$ be a compact metric space.*

(i) *X is* separable *(i.e. it has an at most countably infinite dense subset).*

(ii) *X is* second-countable *(i.e. it has an at most countably infinite base).*

Proof. By Theorem 1.6.8, X is totally bounded. In particular, for every $n \geq 1$, one can cover X by a finite number of balls $B(x_{n,1}, \frac{1}{n}), \ldots, B(x_{n,k_n}, \frac{1}{n})$ of radius $\frac{1}{n}$. The set of points $\{x_{n,i} : n \geq 1; 1 \leq i \leq k_n\}$ is then easily verified to be dense and at most countable, giving (i). Similarly, the set of balls $\{B(x_{n,i}, \frac{1}{n}) : n \geq 1; 1 \leq i \leq k_n\}$ can be easily verified to be a base which is at most countable, giving (ii). □

Remark 1.8.7. One can easily generalise compactness here to σ-compactness; thus for instance finite-dimensional vector spaces $\mathbf{R}^n$ are separable and second-countable. The properties of separability and second-countability are much weaker than σ-compactness in general, but can still serve to provide some constraint as to the "size" or "complexity" of a metric space or topological space in many situations.

We now weaken the notion of a base to that of a *sub-base*.

Definition 1.8.8 (Sub-base)**.** Let $(X, \mathcal{F})$ be a topological space. A *sub-base* for this space is a collection $\mathcal{B}$ of subsets of X such that $\mathcal{F}$ is the weakest topology that makes $\mathcal{B}$ open (i.e. $\mathcal{F}$ is generated by $\mathcal{B}$). Elements of $\mathcal{B}$ are referred to as *sub-basic open sets*.

Observe for instance that every base is a sub-base. The converse is not true: for instance, the half-open intervals $(-\infty, a), (a, +\infty)$ for $a \in \mathbf{R}$ form a sub-base for the standard topology on $\mathbf{R}$, but not a base. In contrast to bases, which need to obey the property in Exercise 1.8.8, no property is required on a collection $\mathcal{B}$ in order for it to be a sub-base; every collection of sets generates a unique topology with respect to which it is a sub-base.

The precise relationship between sub-bases and bases is given by the following exercise.

Exercise 1.8.10. Let $(X, \mathcal{F})$ be a topological space, and let $\mathcal{B}$ be a collection of subsets of X. Then the following are equivalent:

- $\mathcal{B}$ is a sub-base for $(X, \mathcal{F})$.
- The space $\mathcal{B}^* := \{B_1 \cap \ldots \cap B_k : B_1, \ldots, B_k \in \mathcal{B}\}$ of finite intersections of $\mathcal{B}$ (including the whole space X, which corresponds to the case $k = 0$) is a base for $(X, \mathcal{F})$.

Thus a set is open iff it is the union of finite intersections of sub-basic open sets.

Many topological facts involving open sets can often be reduced to verifications on basic or sub-basic open sets, as the following exercise illustrates:

Exercise 1.8.11. Let $(X, \mathcal{F})$ be a topological space, and $\mathcal{B}$ be a sub-base of X, and let $\mathcal{B}^*$ be a base of X.

- Show that a sequence $x_n \in X$ converges to a limit $x \in X$ if and only if every sub-basic open neighbourhood of x contains x_n for all sufficiently large x_n. (Optional: show that an analogous statement is also true for nets.)
- Show that a point $x \in X$ is adherent to a set E if and only if every basic open neighbourhood of x intersects E. Give an example to show that the claim fails for sub-basic open sets.
- Show that a point $x \in X$ is in the interior of a set U if and only if U contains a basic open neighbourhood of x. Give an example to show that the claim fails for sub-basic open sets.
- If Y is another topological space, show that a map $f : Y \to X$ is continuous if and only if the inverse image of every sub-basic open set is open.

There is a useful strengthening of Exercise 1.8.9 in the spirit of the above exercise, namely the *Alexander sub-base theorem*:

Theorem 1.8.9 (Alexander sub-base theorem)**.** *Let $(X, \mathcal{F})$ be a topological space with a sub-base $\mathcal{B}$. Then the following are equivalent:*

- *Every open cover has a finite subcover (i.e. X is compact);*
- *Every sub-basic open cover has a finite subcover.*

Proof. Call an open cover *bad* if it had no finite subcover, and *good* otherwise. In view of Exercise 1.8.9, it suffices to show that if every sub-basic open cover is good, then every basic open cover is good also, where we use the basis $\mathcal{B}^*$ coming from Exercise 1.8.10.

Suppose for contradiction that every sub-basic open cover was good, but at least one basic open cover was bad. If we order the bad basic open covers by set inclusion, observe that every chain of bad basic open covers has an upper bound that is also a bad basic open cover, namely the union of all the covers in the chain. Thus, by Zorn's lemma (Section 2.4), there exists a maximal bad basic open cover $\mathcal{C} = (U_\alpha)_{\alpha \in A}$. Thus this cover has no finite subcover, but if one adds any new basic open set to this cover, then there must now be a finite subcover.

Pick a basic open set U_α in this cover $\mathcal{C}$. Then we can write $U_\alpha = B_1 \cap \ldots \cap B_k$ for some sub-basic open sets $B_1, \ldots, B_k$. We claim that at least one of the $B_1, \ldots, B_k$ also lie in the cover $\mathcal{C}$. To see this, suppose for contradiction that none of the $B_1, \ldots, B_k$ was in $\mathcal{C}$. Then adding any of the B_i to $\mathcal{C}$ enlarges the basic open cover and thus creates a finite subcover; thus B_i together with finitely many sets from $\mathcal{C}$ cover X, or equivalently that one can cover $X \backslash B_i$ with finitely many sets from $\mathcal{C}$. Thus one can also cover $X \backslash U_\alpha = \bigcup_{i=1}^k (X \backslash B_i)$ with finitely many sets from $\mathcal{C}$, and thus X itself can be covered by finitely many sets from $\mathcal{C}$, a contradiction.

From the above discussion and the axiom of choice, we see that for each basic set U_α in $\mathcal{C}$ there exists a sub-basic set B_α containing U_α that also lies in $\mathcal{C}$. (Two different basic sets U_α, U_β could lead to the same sub-basic set $B_\alpha = B_\beta$, but this will not concern us.) Since the U_α cover X, the B_α do also. By hypothesis, a finite number of B_α can cover X, and so $\mathcal{C}$ is good, which gives the desired a contradiction. □

Exercise 1.8.12. (Optional) Use Exercise 1.8.7 to give another proof of the Alexander sub-base theorem.

Exercise 1.8.13. Use the Alexander sub-base theorem to show that the unit interval $[0,1]$ (with the usual topology) is compact, without recourse to the *Heine-Borel* or *Bolzano-Weierstrass* theorems.

Exercise 1.8.14. Let X be a well-ordered set, endowed with the order topology (Exercise 1.6.10); such a space is known as an *ordinal space*. Show that X is Hausdorff, and that X is compact if and only if X has a maximal element.

One of the major applications of the sub-base theorem is to prove *Tychonoff's theorem*, which we turn to next.

1.8.3. Compactness and product spaces. Given two topological spaces $X = (X, \mathcal{F}_X)$ and $Y = (Y, \mathcal{F}_Y)$, we can form the *product space* $X \times Y$, using the cylinder sets $\{U \times Y : U \in \mathcal{F}_X\} \cup \{X \times V : V \in \mathcal{F}_Y\}$ as a sub-base, or equivalently using the open boxes $\{U \times V : U \in \mathcal{F}_X, V \in \mathcal{F}_Y\}$ as a base (cf. Example 1.6.25). One easily verifies that the obvious projection maps $\pi_X : X \times Y \to X$, $\pi_Y : X \times Y \to Y$ are continuous, and that these maps also provide homeomorphisms between $X \times \{y\}$ and X, or between $\{x\} \times Y$ and Y, for every $x \in X, y \in Y$. Also observe that a sequence $(x_n, y_n)_{n=1}^\infty$ (or net $(x_\alpha, y_\alpha)_{\alpha \in A}$) converges to a limit (x, y) in X if and only if $(x_n)_{n=1}^\infty$ and $(y_n)_{n=1}^\infty$ (or $(x_\alpha)_{\alpha \in A}$ and $(y_\alpha)_{\alpha \in A}$) converge in X and Y to x and y respectively.

This operation preserves a number of useful topological properties, for instance

Exercise 1.8.15. Prove that the product of two Hausdorff spaces is still Hausdorff.

Exercise 1.8.16. Prove that the product of two sequentially compact spaces is still sequentially compact.

Proposition 1.8.10. *The product of two compact spaces is compact.*

Proof. By Exercise 1.8.9 it suffices to show that any basic open cover of $X \times Y$ by boxes $(U_\alpha \times V_\alpha)_{\alpha \in A}$ has a finite subcover. For any $x \in X$, this open cover covers $\{x\} \times Y$; by the compactness of $Y \equiv \{x\} \times Y$, we can thus cover $\{x\} \times Y$ by a finite number of open boxes $U_\alpha \times V_\alpha$. Intersecting the U_α together, we obtain a neighbourhood U_x of x

such that $U_x \times Y$ is covered by a finite number of these boxes. But by compactness of X, we can cover X by a finite number of U_x. Thus all of $X \times Y$ can be covered by a finite number of boxes in the cover, and the claim follows. □

Exercise 1.8.17. (Optional) Obtain an alternate proof of this proposition using Exercise 1.6.15.

The above theory for products of two spaces extends without difficulty to products of finitely many spaces. Now we consider infinite products.

Definition 1.8.11 (Product spaces)**.** Given a family $(X_\alpha, \mathcal{F}_\alpha)_{\alpha \in A}$ of topological spaces, let $X := \prod_{\alpha \in A} X_\alpha$ be the Cartesian product, i.e. the space of tuples $(x_\alpha)_{\alpha \in A}$ with $x_\alpha \in X_\alpha$ for all $\alpha \in A$. For each $\alpha \in A$, we have the obvious projection map $\pi_\alpha : X \to X_\alpha$ that maps $(x_\beta)_{\beta \in A}$ to x_α.

- We define the *product topology* on X to be the topology generated by the cylinder sets $\pi_\alpha^{-1}(U_\alpha)$ for $\alpha \in A$ and $U_\alpha \in \mathcal{F}_\alpha$ as a sub-base, or equivalently the weakest topology that makes all of the π_α continuous.
- We define the *box topology* on X to be the topology generated by all the boxes $\prod_{\alpha \in A} U_\alpha$, where $U_\alpha \in \mathcal{F}_\alpha$ for all $\alpha \in A$.

Unless otherwise specified, we assume the product space to be endowed with the product topology rather than the box topology.

When A is finite, the product topology and the box topology coincide. When A is infinite, the two topologies are usually different (as we shall see), but the box topology is always at least as strong as the product topology. Actually, in practice the box topology is too strong to be of much use - there are not enough convergent sequences in it. For instance, in the space $\mathbf{R}^\mathbf{N}$ of real-valued sequences $(x_n)_{n=1}^\infty$, even sequences such as $(\frac{1}{m!} e^{-nm})_{n=1}^\infty$ do not converge to the zero sequence as $m \to \infty$ (why?), despite converging in just about every other sense.

Exercise 1.8.18. Show that the arbitrary product of Hausdorff spaces remains Hausdorff in either the product or the box topology.

Exercise 1.8.19. Let (X_n, d_n) be a sequence of metric spaces. Show that the the function $d : X \times X \to \mathbf{R}^+$ on the product space $X := \prod_n X_n$ defined by

$$d((x_n)_{n=1}^\infty, (y_n)_{n=1}^\infty) := \sum_{n=1}^\infty 2^{-n} \frac{d_n(x_n, y_n)}{1 + d_n(x_n, y_n)}$$

is a metric on X which generates the product topology on X.

Exercise 1.8.20. Let $X = \prod_{\alpha \in A} X_\alpha$ be a product space with the product topology. Show that a sequence x_n in that space converges to a limit $x \in X$ if and only if $\pi_\alpha(x_n)$ converges in X_α to $\pi_\alpha(x)$ for every $\alpha \in A$. (The same statement also holds for nets.) Thus convergence in the product topology is essentially the same concept as pointwise convergence (cf. Example 1.6.24).

The box topology usually does not preserve compactness. For instance, one easily checks that the product of any number of discrete spaces is still discrete in the box topology. On the other hand, a discrete space is compact (or sequentially compact) if and only if it is finite. Thus the infinite product of any number of non-trivial (i.e. having at least two elements) compact discrete spaces will be non-compact, and similarly for sequential compactness.

The situation improves significantly with the product topology, however (which is weaker, and thus more likely to be compact). We begin with the situation for sequential compactness.

Proposition 1.8.12 (Sequential Tychonoff theorem)**.** *Any at most countable product of sequentially compact topological spaces is sequentially compact.*

Proof. We will use the "Arzelá-Ascoli diagonalisation argument". The finite case is already handled by Exercise 1.8.16 (and can in any event be easily deduced from the countable case), so suppose we have a countably infinite sequence $(X_n, \mathcal{F}_n)_{n=1}^\infty$ of sequentially compact spaces, and consider the product space $X = \prod_{n=1}^\infty X_n$ with the product topology. Let $x^{(1)}, x^{(2)}, \ldots$ be a sequence in X, thus each $x^{(m)}$ is itself a sequence $x^{(m)} = (x_n^{(m)})_{n=1}^\infty$ with $x_n^{(m)} \in X_n$ for all n. Our objective is to find a subsequence $x^{(m_j)}$ which converges to some limit $x = (x_n)_{n=1}^\infty$ in the product topology, which by Exercise 1.8.20

is the same as pointwise convergence (i.e. $x_n^{(m_j)} \to x_n$ as $j \to \infty$ for each n).

Consider the first coordinates $x_1^{(m)} \in X_1$ of the sequence $x^{(m)}$. As X_1 is sequentially compact, we can find a subsequence $(x^{(m_{1,j})})_{j=1}^\infty$ in X such that $x_1^{(m_{1,j})}$ converges in X_1 to some limit $x_1 \in X_1$.

Now, in this subsequence, consider the second coordinates $x_2^{(m_{1,j})} \in X_2$. As X_2 is sequentially compact, we can find a further subsequence $(x^{(m_{2,j})})_{j=1}^\infty$ in X such that $x_2^{(m_{2,j})}$ converges in X_2 to some limit $x_2 \in X_1$. Also, we inherit from the preceding subsequence that $x_1^{(m_{2,j})}$ converges in X_1 to x_1.

We continue in this vein, creating nested subsequences $(x^{(m_{i,j})})_{j=1}^\infty$ for $i = 1, 2, 3, \ldots$ whose first i components $x_1^{(m_{i,j})}, \ldots, x_i^{(m_{i,j})}$ converge to $x_1 \in X_1, \ldots, x_i \in X_i$ respectively.

None of these subsequences, by themselves are sufficient to finish the problem. But now we use the diagonalisation trick: we consider the diagonal sequence $(x^{(m_{j,j})})_{j=1}^\infty$. One easily verifies that $x_n^{(m_{j,j})}$ converges in X_n to x_n as $j \to \infty$ for every n, and so we have extracted a sequence that is convergent in the product topology. □

Remark 1.8.13. In the converse direction, if a product of spaces is sequentially compact, then each of the factor spaces must also be sequentially compact, since they are continuous images of the product space and one can apply Exercise 1.8.1.

The sequential Tychonoff theorem breaks down for uncountable products. Consider for instance the product space $X := \{0,1\}^{\{0,1\}^{\mathbf{N}}}$ of functions $f : \{0,1\}^{\mathbf{N}} \to \{0,1\}$. As $\{0,1\}$ (with the discrete topology) is sequentially compact, this is an (uncountable) product of sequentially compact spaces. On the other hand, for each $n \in \mathbf{N}$ we can define the evaluation function $f_n : \{0,1\}^{\mathbf{N}} \to \{0,1\}$ by $f_n : (a_m)_{m=1}^\infty \mapsto a_n$. This is a sequence in X; we claim that it has no convergent subsequence. Indeed, given any $n_j \to \infty$, we can find $x = (x_m)_{m=1}^\infty \in \{0,1\}^\infty$ such that $x_{n_j} = f_{n_j}(x)$ does not converge to a limit as $j \to \infty$, and so f_{n_j} does not converge pointwise (i.e. does not converge in the product topology).

However, we can recover the result for uncountable products as long as we work with topological compactness rather than sequential compactness, leading to *Tychonoff's theorem*:

Theorem 1.8.14 (Tychonoff theorem)**.** *Any product of compact topological spaces is compact.*

Proof. Write $X = \prod_{\alpha \in A} X_\alpha$ for this product of compact topological spaces. By Theorem 1.8.9, it suffices to show that any open cover of X by sub-basic open sets $(\pi_{\alpha_\beta}^{-1}(U_\beta))_{\beta \in B}$ has a finite sub-cover, where B is some index set, and for each $\beta \in B$, $\alpha_\beta \in A$ and U_β is open in X_{α_β}.

For each $\alpha \in A$, consider the sub-basic open sets $\pi_\alpha^{-1}(U_\beta)$ that are associated to those $\beta \in B$ with $\alpha_\beta = \alpha$. If the open sets U_β here cover X_α, then by compactness of X_α, a finite number of the U_β already suffice to cover X_α, and so a finite number of the $\pi_\alpha^{-1}(U_\beta)$ cover X, and we are done. So we may assume that the U_β do not cover X_α, thus there exists $x_\alpha \in X_\alpha$ that avoids all the U_β with $\alpha_\beta = \alpha$. One then sees that the point $(x_\alpha)_{\alpha \in A}$ in X avoids all of the $\pi_\alpha^{-1}(U_\beta)$, a contradiction. The claim follows. □

Remark 1.8.15. The axiom of choice was used in several places in the proof (in particular, via the Alexander sub-base theorem). This turns out to be necessary, because one can use Tychonoff's theorem to establish the axiom of choice. This was first observed by Kelley, and can be sketched as follows. It suffices to show that the product $\prod_{\alpha \in A} X_\alpha$ of non-empty sets is again non-empty. We can make each X_α compact (e.g. by using the trivial topology). We then adjoin an isolated element ∞ to each X_α to obtain another compact space $X_\alpha \cup \{\infty\}$, with X_α closed in $X_\alpha \cup \{\infty\}$. By Tychonoff's theorem, the product $X := \prod_{\alpha \in A}(X_\alpha \cup \{\infty\})$ is compact, and thus every collection of closed sets with finite intersection property has non-empty intersection. But observe that the sets $\pi_\alpha^{-1}(X_\alpha)$ in X, where $\pi_\alpha : X \to X_\alpha \cup \{\infty\}$ is the obvious projection, are closed and has the finite intersection property; thus the intersection of all of these sets is non-empty, and the claim follows.

Remark 1.8.16. From the above discussion, we see that the space $\{0,1\}^{\{0,1\}^{\mathbb{Z}}}$ is compact but not sequentially compact; thus compactness does not necessarily imply sequential compactness.

Exercise 1.8.21. Let us call a topological space $(X, \mathcal{F})$ *first-countable* if, for every $x \in X$, there exists a countable family $B_{x,1}, B_{x,2}, \ldots$ of open neighbourhoods of x such that every neighbourhood of x contains at least one of the $B_{x,j}$.

- Show that every metric space is first-countable.
- Show that every second-countable space is first-countable (see Lemma 1.8.6).
- Show that every separable metric space is second-countable.
- Show that every space which is second-countable, is separable.
- (Optional) Show that every net $(x_\alpha)_{\alpha \in A}$ which converges in X to x, has a convergent subsequence $(x_{\phi(n)})_{n=1}^{\infty}$ (i.e. a subnet whose index set is $\mathbf{N}$).
- Show that any compact space which is first-countable, is also sequentially compact. (The converse is not true: Exercise 1.6.10 provides a counterexample.)

(Optional) There is an alternate proof of the Tychonoff theorem that uses the machinery of *universal nets*. We sketch this approach in a series of exercises.

Definition 1.8.17. A net $(x_\alpha)_{\alpha \in A}$ in a set X is *universal* if for every function $f : X \to \{0,1\}$, the net $(f(x_\alpha))_{\alpha \in A}$ converges to either 0 or 1.

Exercise 1.8.22. Show that a universal net $(x_\alpha)_{\alpha \in A}$ in a compact topological space is necessarily convergent. (*Hint*: show that the collection of closed sets which contain x_α for sufficiently large α enjoys the finite intersection property.)

Exercise 1.8.23 (Kelley's theorem)**.** Every net $(x_\alpha)_{\alpha \in A}$ in a set X has a universal subnet $(x_{\phi(\beta)})_{\beta \in B}$. (*Hint*: First use Exercise 1.8.5 to find an ultrafilter p on A that contains the upsets $\{\beta \in A : \beta \geq \alpha\}$ for all $\alpha \in A$. Now let B be the space of all pairs (U, α), where

$\alpha \in U \in p$, ordered by requiring $(U, \alpha) \leq (U', \alpha')$ when $U \supset U'$ and $\alpha \leq \alpha'$, and let $\phi : B \to A$ be the map $\phi : (U, \alpha) \mapsto \alpha$.)

Exercise 1.8.24. Use the previous two exercises, together with Exercise 1.8.20, to establish an alternate proof of Tychonoff's theorem.

Exercise 1.8.25. Establish yet another proof of Tychonoff's theorem using Exercise 1.8.7 directly (rather than proceeding via Exercise 1.8.12).

1.8.4. Compactness and equicontinuity. We now pause to give an important application of the (sequential) Tychonoff theorem. We begin with some definitions. If $X = (X, \mathcal{F}_X)$ is a topological space and $Y = (Y, d_Y)$ is a metric space, let $BC(X \to Y)$ be the space of bounded continuous functions from X to Y. (If X is compact, this is the same space as $C(X \to Y)$, the space of continuous functions from X to Y.) We can give this space the uniform metric

$$d(f, g) := \sup_{x \in X} d_Y(f(x), g(x)).$$

Exercise 1.8.26. If Y is complete, show that $BC(X \to Y)$ is a complete metric space. (Note that this implies Exercise 1.5.2.)

Note that if $f : X \to Y$ is continuous if and only if, for every $x \in X$ and $\varepsilon > 0$, there exists a neighbourhood U of x such that $d_Y(f(x'), f(x)) \leq \varepsilon$ for all $x' \in U$. We now generalise this concept to families.

Definition 1.8.18. Let X be a topological space, let Y be a metric space, and Let $(f_\alpha)_{\alpha \in A}$ be a family of functions $f_\alpha \in BC(X \to Y)$.

- We say that this family f_α is *pointwise bounded* if for every $x \in X$, the set $\{f_\alpha(x) : \alpha \in A\}$ is bounded in Y.
- We say that this family f_α is *pointwise precompact* if for every $x \in X$, the set $\{f_\alpha(x) : \alpha \in A\}$ is precompact in Y.
- We say that this family f_α is *equicontinuous* if for every $x \in X$ and $\varepsilon > 0$, there exists a neighbourhood U of x such that $d_Y(f_\alpha(x'), f_\alpha(x)) \leq \varepsilon$ for all $\alpha \in A$ and $x' \in U$.
- If $X = (X, d_X)$ is also a metric space, we say that the family f_α is *uniformly equicontinuous* if for every $\varepsilon > 0$ there exists

a $\delta > 0$ such that $d_Y(f_\alpha(x'), f_\alpha(x)) \leq \varepsilon$ for all $\alpha \in A$ and $x', x \in x$ with $d_X(x, x') \leq \delta$.

Remark 1.8.19. From the Heine-Borel theorem, the pointwise boundedness and pointwise precompactness properties are equivalent if Y is a subset of $\mathbf{R}^n$ for some n. Any finite collection of continuous functions is automatically an equicontinuous family (why?), and any finite collection of uniformly continuous functions is automatically a uniformly equicontinuous family; the concept only acquires additional meaning once one considers infinite families of continuous functions.

Example 1.8.20. With $X = [0,1]$ and $Y = \mathbf{R}$, the family of functions $f_n(x) := x^n$ for $n = 1, 2, 3, \ldots$ are pointwise bounded (and thus pointwise precompact), but not equicontinuous. The family of functions $g_n(x) := n$ for $n = 1, 2, 3, \ldots$, on the other hand, are equicontinuous, but not pointwise bounded or pointwise precompact. The family of functions $h_n(x) := \sin nx$ for $n = 1, 2, 3, \ldots$ are pointwise bounded (even uniformly bounded), but not equicontinuous.

Example 1.8.21. With $X = Y = \mathbf{R}$, the functions $f_n(x) = \arctan nx$ are pointwise bounded (even uniformly bounded), are equicontinuous, and are each *individually* uniformly continuous, but are not uniformly equicontinuous.

Exercise 1.8.27. Show that the uniform boundedness principle (Theorem 1.7.5) can be restated as the assertion that any family of bounded linear operators from the unit ball of a Banach space to a normed vector space is pointwise bounded if and only if it is equicontinuous.

Example 1.8.22. A function $f : X \to Y$ between two metric spaces is said to be *Lipschitz* (or *Lipschitz continuous*) if there exists a constant C such that $d_Y(f(x), f(x')) \leq C d_X(x, x')$ for all $x, x' \in X$; the smallest constant C one can take here is known as the *Lipschitz constant* of f. Observe that Lipschitz functions are automatically continuous, hence the name. Also observe that a family $(f_\alpha)_{\alpha \in A}$ of Lipschitz functions with uniformly bounded Lipschitz constant is equicontinuous.

One nice consequence of equicontinuity is that it equates uniform convergence with pointwise convergence, or even pointwise convergence on a dense subset.

Exercise 1.8.28. Let X be a topological space, let Y be a complete metric space, let $f_1, f_2, \ldots \in BC(X \to Y)$ be an equicontinuous family of functions. Show that the following are equivalent:

- The sequence f_n is pointwise convergent.
- The sequence f_n is pointwise convergent on some dense subset of X.

If X is compact, show that the above two statements are also equivalent to

- The sequence f_n is uniformly convergent.

(Compare with Corollary 1.7.7.) Show that no two of the three statements remain equivalent if the hypothesis of equicontinuity is dropped.

We can now use Proposition 1.8.12 to give a useful characterisation of precompactness in $C(X \to Y)$ when X is compact, known as the *Arzelá-Ascoli theorem*:

Theorem 1.8.23 (Arzelá-Ascoli theorem)**.** *Let Y be a metric space, X be a compact metric space, and let $(f_\alpha)_{\alpha \in A}$ be a family of functions $f_\alpha \in BC(X \to Y)$. Then the following are equivalent:*

(i) *$\{f_\alpha : \alpha \in A\}$ is a precompact subset of $BC(X \to Y)$.*

(ii) *$(f_\alpha)_{\alpha \in A}$ is pointwise precompact and equicontinuous.*

(iii) *$(f_\alpha)_{\alpha \in A}$ is pointwise precompact and uniformly equicontinuous.*

Proof. We first show that (i) implies (ii). For any $x \in X$, the evaluation map $f \mapsto f(x)$ is a continuous map from $C(X \to Y)$ to Y, and thus maps precompact sets to precompact sets. As a consequence, any precompact family in $C(X \to Y)$ is pointwise precompact. To show equicontinuity, suppose for contradiction that equicontinuity failed at some point x, thus there exists $\varepsilon > 0$, a sequence $\alpha_n \in A$, and points $x_n \to x$ such that $d_Y(f_{\alpha_n}(x_n), f_{\alpha_n}(x)) > \varepsilon$ for every n. One then verifies that no subsequence of f_{α_n} can converge uniformly to a continuous limit, contradicting precompactness. (Note that in the metric space $C(X \to Y)$, precompactness is equivalent to sequential precompactness.)

Now we show that (ii) implies (iii). It suffices to show that equicontinuity implies uniform equicontinuity. This is a straightforward generalisation of the more familiar argument that continuity implies uniform continuity on a compact domain, and we repeat it here. Namely, fix $\varepsilon > 0$. For every $x \in X$, equicontinuity provides a $\delta_x > 0$ such that $d_Y(f_\alpha(x), f_\alpha(x')) \leq \varepsilon$ whenever $x' \in B(x, \delta_x)$ and $\alpha \in A$. The balls $B(x, \delta_x/2)$ cover X, thus by compactness some finite subcollection $B(x_i, \delta_{x_i}/2)$, $i = 1, \ldots, n$ of these balls cover X. One then easily verifies that $d_Y(f_\alpha(x), f_\alpha(x')) \leq \varepsilon$ whenever $x, x' \in X$ with $d_X(x, x') \leq \min_{1 \leq i \leq n} \delta_{x_i}/2$.

Finally, we show that (iii) implies (i). It suffices to show that any sequence $f_n \in BC(X \to Y)$, $n = 1, 2, \ldots$, which is pointwise precompact and uniformly equicontinuous, has a convergent subsequence. By embedding Y in its metric completion $\overline{Y}$, we may assume without loss of generality that Y is complete. (Note that for every $x \in X$, the set $\{f_n(x) : n = 1, 2, \ldots\}$ is precompact in Y, hence the closure in Y is complete and thus closed in $\overline{Y}$ also. Thus any pointwise limit of the f_n in $\overline{Y}$ will take values in Y.) By Lemma 1.8.6, we can find a countable dense subset $x_1, x_2, \ldots$ of X. For each x_m, we can use pointwise precompactness to find a compact set $K_m \subset Y$ such that $f_\alpha(x_m)$ takes values in K_m. For each n, the tuple $F_n := (f_n(x_m))_{m=1}^\infty$ can then be viewed as a point in the product space $\prod_{n=1}^\infty K_n$. By Proposition 1.8.12, this product space is sequentially compact, hence we may find a subsequence $n_j \to \infty$ such that F_n is convergent in the product topology, or equivalently that f_n pointwise converges on the countable dense set $\{x_1, x_2, \ldots\}$. The claim now follows from Exercise 1.8.28. □

Remark 1.8.24. The above theorem characterises precompact subsets of $BC(X \to Y)$ when X is a compact metric space. One can also characterise compact subsets by observing that a subset of a metric space is compact if and only if it is both precompact and closed.

There are many variants of the Arzelá-Ascoli theorem with stronger or weaker hypotheses or conclusions; for instance, we have

Corollary 1.8.25 (Arzelá-Ascoli theorem, special case)**.** *Let $f_n : X \to \mathbf{R}^m$ be a sequence of functions from a compact metric space*

X to a finite-dimensional vector space $\mathbf{R}^m$ which are equicontinuous and pointwise bounded. Then there is a subsequence f_{n_j} of f_n which converges uniformly to a limit (which is necessarily bounded and continuous).

Thus, for instance, any sequence of uniformly bounded and uniformly Lipschitz functions $f_n : [0,1] \to \mathbf{R}$ will have a uniformly convergent subsequence. This claim fails without the uniform Lipschitz assumption (consider, for instance, the functions $f_n(x) := \sin(nx)$). Thus one needs a "little bit extra" uniform regularity in addition to uniform boundedness in order to force the existence of uniformly convergent subsequences. This is a general phenomenon in infinite-dimensional function spaces: compactness in a strong topology tends to require some sort of uniform control on regularity or decay in addition to uniform bounds on the norm.

Exercise 1.8.29. Show that the equivalence of (i) and (ii) continues to hold if X is assumed to be just a compact Hausdorff space rather than a compact metric space (the statement (iii) no longer makes sense in this setting). *Hint*: X need not be separable any more, however one can still adapt the diagonalisation argument used to prove Proposition 1.8.12. The starting point is the observation that for every $\varepsilon > 0$ and every $x \in X$, one can find a neighbourhood U of x and some subsequence f_{n_j} which only oscillates by at most ε (or maybe 2ε) on U.

Exercise 1.8.30 (Locally compact Hausdorff version of Arzelá-Ascoli)**.** Let X be a locally compact Hausdorff space which is also σ-compact, and let $f_n \in C(X \to \mathbf{R})$ be an equicontinuous, pointwise bounded sequence of functions. Then there exists a subsequence $f_{n_j} \in C(X \to \mathbf{R})$ which converges uniformly on compact subsets of X to a limit $f \in C(X \to \mathbf{R})$. (*Hint*: Express X as a countable union of compact sets K_n, each one contained in the interior of the next. Apply the compact Hausdorff Arzelá-Ascoli theorem on each compact set (Exercise 1.8.29). Then apply the Arzelá-Ascoli argument one last time.)

Remark 1.8.26. The Arzelá-Ascoli theorem (and other compactness theorems of this type) are often used in partial differential equations,

to demonstrate existence of solutions to various equations or variational problems. For instance, one may wish to solve some equation $F(u) = f$, for some function $u : X \to \mathbf{R}^m$. One way to do this is to first construct a sequence u_n of approximate solutions, so that $F(u_n) \to f$ as $n \to \infty$ in some suitable sense. If one can also arrange these u_n to be equicontinuous and pointwise bounded, then the Arzelá-Ascoli theorem allows one to pass to a subsequence that converges to a limit u. Given enough continuity (or *semi-continuity*) properties on F, one can then show that $F(u) = f$ as required.

More generally, the use of compactness theorems to demonstrate existence of solutions in PDE is known as the *compactness method*. It is applicable in a remarkably broad range of PDE problems, but often has the drawback that it is difficult to establish uniqueness of the solutions created by this method (compactness guarantees existence of a limit point, but not uniqueness). Also, in many cases one can only hope for compactness in rather weak topologies, and as a consequence it is often difficult to establish regularity of the solutions obtained via compactness methods.

Notes. This lecture first appeared at `terrytao.wordpress.com/2009/02/09`. Thanks to Nate Chandler, Emmanuel Kowalski, Eric, K. P. Hart, Ke, Luca Trevisan, PDEBeginner, RR, Samir Chomsky, Xiaochuan Liu, and anonymous commenters for corrections.

David Speyer and Eric pointed out that the axiom of choice was used in two different ways in the proof of Tychonoff's theorem; firstly to prove the sub-base theorem, and secondly to select an element x_α from each X_α. Interestingly, it is the latter use which is the more substantial one; the sub-base theorem can be shown to be equivalent to the ultrafilter lemma, which is strictly weaker than the axiom of choice. Furthermore, for Hausdorff spaces, one can establish Tychonoff's theorem purely using ultralimits, which shows that the strange non-Hausdorff nature of the topology in Remark 1.8.15.

1.9. The strong and weak topologies

A *normed vector space* $(X, \|\|_X)$ automatically generates a topology, known as the *norm topology* or *strong topology* on X, generated by

the open balls $B(x,r) := \{y \in X : \|y-x\|_X < r\}$. A sequence x_n in such a space *converges strongly* (or *converges in norm*) to a limit x if and only if $\|x_n - x\|_X \to 0$ as $n \to \infty$. This is the topology we have implicitly been using in our previous discussion of normed vector spaces.

However, in some cases it is useful to work in topologies on vector spaces that are weaker than a norm topology. One reason for this is that many important modes of convergence, such as *pointwise convergence*, *convergence in measure*, *smooth convergence*, or *convergence on compact subsets*, are not captured by a norm topology, and so it is useful to have a more general theory of *topological vector spaces* that contains these modes. Another reason (of particular importance in PDE) is that the norm topology on infinite-dimensional spaces is so strong that very few sets are compact or pre-compact in these topologies, making it difficult to apply *compactness methods* in these topologies (cf. Section 1.6 of *Poincaré's Legacies, Vol. II*). Instead, one often first works in a weaker topology, in which compactness is easier to establish, and then somehow upgrades any weakly convergent sequences obtained via compactness to stronger modes of convergence (or alternatively, one abandons strong convergence and exploits the weak convergence directly). Two basic weak topologies for this purpose are the *weak topology* on a normed vector space X, and the *weak* topology* on a dual vector space X^*. Compactness in the latter topology is usually obtained from the *Banach-Alaoglu theorem* (and its sequential counterpart), which will be a quick consequence of the *Tychonoff theorem* (and its sequential counterpart) from the previous section.

The strong and weak topologies on normed vector spaces also have analogues for the space $B(X \to Y)$ of bounded linear operators from X to Y, thus supplementing the operator norm topology on that space with two weaker topologies, which (somewhat confusingly) are named the *strong operator topology* and the *weak operator topology*.

1.9.1. Topological vector spaces. We begin with the definition of a *topological vector space*, which is a space with suitably compatible topological and vector space structures on it.

Definition 1.9.1. A *topological vector space* $V = (V, \mathcal{F})$ is a real or complex vector space V, together with a topology $\mathcal{F}$ such that the addition operation $+ : V \times V \to V$ and the scalar multiplication operation $\cdot : \mathbf{R} \times V \to V$ or $\cdot : \mathbf{C} \times V \to V$ is jointly continuous in both variables (thus, for instance, $+$ is continuous from $V \times V$ with the product topology to V).

It is an easy consequence of the definitions that the translation maps $x \mapsto x + x_0$ for $x_0 \in V$ and the dilation maps $x \mapsto \lambda \cdot x$ for non-zero scalars λ are homeomorphisms on V; thus for instance the translation or dilation of an open set (or a closed set, a compact set, etc.) is open (resp. closed, compact, etc.). We also have the usual limit laws: if $x_n \to x$ and $y_n \to y$ in a topological vector space, then $x_n + y_n \to x + y$, and if $\lambda_n \to \lambda$ in the field of scalars, then $\lambda_n x_n \to \lambda x$. (Note how we need joint continuity here; if we only had continuity in the individual variables, we could only conclude that $x_n + y_n \to x + y$ (for instance) if one of x_n or y_n was constant.)

We now give some basic examples of topological vector spaces.

Exercise 1.9.1. Show that every normed vector space is a topological vector space, using the balls $B(x, r)$ as the base for the topology. Show that the same statement holds if the vector space is quasi-normed rather than normed.

Exercise 1.9.2. Every semi-normed vector space is a topological vector space, again using the balls $B(x, r)$ as a base for the topology. This topology is Hausdorff if and only if the semi-norm is a norm.

Example 1.9.2. Any linear subspace of a topological vector space is again a topological vector space (with the induced topology).

Exercise 1.9.3. Let V be a vector space, and let $(\mathcal{F}_\alpha)_{\alpha \in A}$ be a (possibly infinite) family of topologies on V, each of which turning V into a topological vector space. Let $\mathcal{F} := \bigvee_{\alpha \in A} \mathcal{F}_\alpha$ be the topology generated by $\bigcup_{\alpha \in A} \mathcal{F}_\alpha$ (i.e. it is the weakest topology that contains all of the $\mathcal{F}_\alpha$. Show that $(V, \mathcal{F})$ is also a topological vector space. Also show that a sequence $x_n \in V$ converges to a limit x in $\mathcal{F}$ if and only if $x_n \to x$ in $\mathcal{F}_\alpha$ for all $\alpha \in A$. (The same statement also holds if sequences are replaced by nets.) In particular, by Exercise 1.9.2, we

can talk about the topological vector space V generated by a family of semi-norms $(\|\|_\alpha)_{\alpha\in A}$ on V.

Exercise 1.9.4. Let $T : V \to W$ be a linear map between vector spaces. Suppose that we give V the topology induced by a family of semi-norms $(\|\|_{V_\alpha})_{\alpha\in A}$, and W the topology induced by a family of semi-norms $(\|\|_{W_\beta})_{\beta\in B}$. Show that T is continuous if and only if, for each $\beta \in B$, there exists a finite subset A_β of A and a constant C_β such that $\|Tf\|_{W_\beta} \leq C_\beta \sum_{\alpha\in A_\beta} \|f\|_{V_\alpha}$ for all $f \in V$.

Example 1.9.3 (Pointwise convergence)**.** Let X be a set, and let $\mathbf{C}^X$ be the space of complex-valued functions $f : X \to \mathbf{C}$; this is a complex vector space. Each point $x \in X$ gives rise to a seminorm $\|f\|_x := |f(x)|$. The topology generated by all of these seminorms is the *topology of pointwise convergence* on $\mathbf{C}^X$ (and is also the product topology on this space); a sequence $f_n \in \mathbf{C}^X$ converges to f in this topology if and only if it converges pointwise. Note that if X has more than one point, then none of the semi-norms individually generate a Hausdorff topology, but when combined together, they do.

Example 1.9.4 (Uniform convergence)**.** Let X be a topological space, and let $C(X)$ be the space of complex-valued continuous functions $f : X \to \mathbf{C}$. If X is not compact, then one does not expect functions in $C(X)$ to be bounded in general, and so the sup norm does not necessarily make $C(X)$ into a normed vector space. Nevertheless, one can still define "balls" $B(f, r)$ in $C(X)$ by

$$B(f, r) := \{g \in C(X) : \sup_{x\in X} |f(x) - g(x)| \leq r\}$$

and verify that these form a base for a topological vector space. A sequence $f_n \in C(X)$ converges in this topology to a limit $f \in C(X)$ if and only if f_n converges uniformly to f, thus $\sup_{x\in X} |f_n(x) - f(x)|$ is finite for sufficiently large n and converges to zero as $n \to \infty$. More generally, one can make a topological vector space out of any "norm", "quasi-norm", or "semi-norm" which is infinite on some portion of the vector space.

Example 1.9.5 (Uniform convergence on compact sets)**.** Let X and $C(X)$ be as in the previous example. For every compact subset K of X, we can define a seminorm $\|\|_{C(K)}$ on $C(X)$ by $\|f\|_{C(K)} :=$

$\sup_{x \in K} |f(x)|$. The topology generated by all of these seminorms (as K ranges over all compact subsets of X) is called the *topology of uniform convergence on compact sets*; it is stronger than the topology of poitnwise convergence but weaker than the topology of uniform convergence. Indeed, a sequence $f_n \in C(X)$ converges to $f \in C(X)$ in this topology if and only if f_n converges uniformly to f on each compact set.

Exercise 1.9.5. Show that an arbitrary product of topological vector spaces (endowed with the product topology) is again a topological vector space[10].

Exercise 1.9.6. Show that a topological vector space is Hausdorff if and only if the origin $\{0\}$ is closed. (*Hint*: first use the continuity of addition to prove the lemma that if V is an open neighbourhood of 0, then there exists another open neighbourhood U of 0 such that $U + U \subset V$, i.e. $u + u' \in V$ for all $u, u' \in U$.)

Example 1.9.6 (Smooth convergence)**.** Let $C^\infty([0,1])$ be the space of smooth functions $f : [0,1] \to \mathbf{C}$. One can define the C^k norm on this space for any non-negative integer k by the formula

$$\|f\|_{C^k} := \sum_{j=0}^{k} \sup_{x \in [0,1]} |f^{(j)}(x)|,$$

where $f^{(j)}$ is the j^{th} derivative of f. The topology generated by all the C^k norms for $k = 0, 1, 2, \ldots$ is the *smooth topology*: a sequence f_n converges in this topology to a limit f if $f_n^{(j)}$ converges uniformly to $f^{(j)}$ for each $j \geq 0$.

Exercise 1.9.7 (Convergence in measure)**.** Let $(X, \mathcal{X}, \mu)$ be a measure space, and let $L(X)$ be the space of measurable functions $f : X \to \mathbf{C}$. Show that the sets

$$B(f, \varepsilon, r) := \{g \in L(X) : \mu(\{x : |f(x) - g(x)| \geq r\} < \varepsilon)\}$$

for $f \in L(X)$, $\varepsilon > 0$, $r > 0$ form the base for a topology that turns $L(X)$ into a topological vector space, and that a sequence $f_n \in L(X)$ converges to a limit f in this topology if and only if it converges in measure.

[10] I am not sure if the same statement is true for the box topology; I believe it is false.

Exercise 1.9.8. Let $[0,1]$ be given the usual Lebesgue measure. Show that the vector space $L^\infty([0,1])$ cannot be given a topological vector space structure in which a sequence $f_n \in L^\infty([0,1])$ converges to f in this topology if and only if it converges almost everywhere. (*Hint*: construct a sequence f_n in $L^\infty([0,1])$ which does not converge pointwise a.e. to zero, but such that every subsequence has a further subsequence that converges a.e. to zero, and use Exercise 1.6.8.) Thus almost everywhere convergence is not "topologisable" in general.

Exercise 1.9.9 (Algebraic topology)**.** Recall that a subset U of a real vector space V is *algebraically open* if the sets $\{t \in \mathbf{R} : x + tv \in U\}$ are open for all $x, v \in V$.

(i) Show that any set which is open in a topological vector space, is also algebraically open.

(ii) Give an example of a set in $\mathbf{R}^2$ which is algebraically open, but not open in the usual topology. (*Hint*: a line intersects the unit circle in at most two points.)

(iii) Show that the collection of algebraically open sets in V is a topology.

(iv) Show that the collection of algebraically open sets in $\mathbf{R}^2$ does *not* give $\mathbf{R}^2$ the structure of a topological vector space.

Exercise 1.9.10 (Quotient topology)**.** Let V be a topological vector space, and let W be a subspace of V. Let $V/W := \{v + W : v \in V\}$ be the space of cosets of W; this is a vector space. Let $\pi : V \to V/W$ be the coset map $\pi(v) := v + W$. Show that the collection of sets $U \subset V/W$ such that $\pi^{-1}(U)$ is open gives V/W the structure of a topological vector space. If V is Hausdorff, show that V/W is Hausdorff if and only if W is closed in V.

Some (but not all) of the concepts that are definable for normed vector spaces, are also definable for the more general category of topological vector spaces. For instance, even though there is no metric structure, one can still define the notion of a Cauchy sequence $x_n \in V$ in a topological vector space: this is a sequence such that $x_n - x_m \to 0$ as $n, m \to \infty$ (or more precisely, for any open neighbourhood U of 0, there exists $N > 0$ such that $x_n - x_m \in U$ for all $n, m \geq N$). It

is then possible to talk about a topological vector space being *complete* (i.e. every Cauchy sequence converges). (From a more abstract perspective, the reason we can define notions such as completeness is because a topological vector space has something better than a topological structure, namely a *uniform structure*.)

Remark 1.9.7. As we have seen in previous lectures, complete normed vector spaces (i.e. Banach spaces) enjoy some very nice properties. Some of these properties (e.g. the *uniform boundedness principle* and the *open mapping theorem*) extend to a slightly larger class of complete topological vector spaces, namely the *Fréchet spaces*. A *Fréchet space* is a complete Hausdorff topological vector space whose topology is generated by an at most countable family of semi-norms; examples include the space $C^\infty([0,1])$ from Exercise 1.9.6 or the uniform convergence on compacta topology from Exercise 1.9.5 in the case when X is σ*-compact*. We will however not study Fréchet spaces systematically here.

One can also extend the notion of a *dual space* V^* from normed vector spaces to topological vector spaces in the obvious manner: the dual space V^* of a topological space is the space of continuous linear functionals from V to the field of scalars (either $\mathbf{R}$ or $\mathbf{C}$, depending on whether V is a real or complex vector space). This is clearly a vector space. Unfortunately, in the absence of a norm on V, one cannot define the analogue of the norm topology on V^*; but as we shall see below, there are some weaker topologies that one can still place on this dual space.

1.9.2. Compactness in the strong topology. We now return to normed vector spaces, and briefly discuss compactness in the strong (or norm) topology on such spaces. In finite dimensions, the *Heine-Borel theorem* tells us that a set is compact if and only if it is closed and bounded. In infinite dimensions, this is not enough, for two reasons. Firstly, compact sets need to be complete, so we are only likely to find many compact sets when the ambient normed vector space is also complete (i.e. it is a Banach space). Secondly, compact sets need to be totally bounded, rather than merely bounded, and

this is quite a stringent condition. Indeed it forces compact sets to be "almost finite-dimensional" in the following sense:

Exercise 1.9.11. Let K be a subset of a Banach space V. Show that the following are equivalent:

(i) K is compact.

(ii) K is sequentially compact.

(iii) K is closed and bounded, and for every $\varepsilon > 0$, K lies in the ε-neighbourhood $\{x \in V : \|x - y\| < \varepsilon \text{ for some } y \in W\}$ of a finite-dimensional subspace W of V.

Suppose furthermore that there is a nested sequence $V_1 \subset V_2 \subset \ldots$ of finite-dimensional subspaces of V such that $\bigcup_{n=1}^{\infty} V_n$ is dense. Show that the following statement is equivalent to the first three:

(iv) K is closed and bounded, and for every $\varepsilon > 0$ there exists an n such that K lies in the ε-neighbourhood of V_n.

Example 1.9.8. Let $1 \leq p < \infty$. In order for a set $K \subset \ell^p(\mathbf{N})$ to be compact in the strong topology, it needs to be closed and bounded, and also *uniformly p^{th}-power integrable at spatial infinity* in the sense that for every $\varepsilon > 0$ there exists $n > 0$ such that

$$(\sum_{m>n} |f(m)|^p)^{1/p} \leq \varepsilon$$

for all $f \in K$. Thus, for instance, the "moving bump" example $\{e_1, e_2, e_3, \ldots\}$, where e_n is the sequence which equals 1 on n and zero elsewhere, is not uniformly p^{th} power integrable and thus not a compact subset of $\ell^p(\mathbf{N})$, despite being closed and bounded.

For "continuous" L^p spaces, such as $L^p(\mathbf{R})$, uniform integrability at spatial infinity is not sufficient to force compactness in the strong topology; one also needs some uniform integrability at very fine scales, which can be described using harmonic analysis tools such as the Fourier transform (Section 1.12). We will not discuss this topic here.

Exercise 1.9.12. Let V be a normed vector space.

- If W is a finite-dimensional subspace of V, and $x \in V$, show that there exists $y \in W$ such that $\|x - y\| \leq \|x - y'\|$ for all

$y' \in W$. Give an example to show that y is not necessarily unique (in contrast to the situation with Hilbert spaces).

- If W is a finite-dimensional proper subspace of V, show that there exists $x \in V$ with $\|x\| = 1$ such that $\|x - y\| \geq 1$ for all $y \in W$. (cf. the *Riesz lemma*.)
- Show that the closed unit ball $\{x \in V : \|x\| \leq 1\}$ is compact in the strong topology if and only if V is finite-dimensional.

1.9.3. The weak and weak* topologies. Let V be a topological vector space. Then, as discussed above, we have the vector space V^* of continuous linear functionals on V. We can use this dual space to create two useful topologies, the *weak topology* on V and the *weak* topology* on V^*:

Definition 1.9.9 (Weak and weak* topologies)**.** Let V be a topological vector space, and let V^* be its dual.

- The *weak topology* on V is the topology generated by the seminorms $\|x\|_\lambda := |\lambda(x)|$ for all $\lambda \in V^*$.
- The *weak* topology* on V^* is the topology generated by the seminorms $\|\lambda\|_x := |\lambda(x)|$ for all $x \in V$.

Remark 1.9.10. It is possible for two non-isomorphic topological vector spaces to have isomorphic duals, but with non-isomorphic weak* topologies. (For instance, $\ell^1(\mathbf{N})$ has a very large number of preduals, which can generate a number of different weak* topologies on $\ell^1(\mathbf{N})$.) So, technically, one cannot talk about *the* weak* topology on a dual space V^*, without specifying exactly what the predual space V is. However, in practice, the predual space is usually clear from context.

Exercise 1.9.13. Show that the weak topology on V is a topological vector space structure on V that is weaker than the strong topology on V. Also, show that the weak* topology on V^* is a topological vector space structure on V^* that is weaker than the weak topology on V^* (which is defined using the double dual $(V^*)^*$. When V is reflexive, show that the weak and weak* topologies on V^* are equivalent.

From the definition, we see that a sequence $x_n \in V$ converges in the weak topology, or *converges weakly* for short, to a limit $x \in V$

if and only if $\lambda(x_n) \to \lambda(x)$ for all $\lambda \in V^*$. This weak convergence is often denoted $x_n \rightharpoonup x$, to distinguish it from strong convergence $x_n \to x$. Similarly, a sequence $\lambda_n \in V^*$ converges in the weak* topology to $\lambda \in V^*$ if $\lambda_n(x) \to \lambda(x)$ for all $x \in V$ (thus λ_n, viewed as a function on V, converges pointwise to λ).

Remark 1.9.11. If V is a Hilbert space, then from the Riesz representation theorem for Hilbert spaces (Theorem 1.4.13) we see that a sequence $x_n \in V$ converges weakly (or in the weak* sense) to a limit $x \in V$ if and only if $\langle x_n, y\rangle \to \langle x, y\rangle$ for all $y \in V$.

Exercise 1.9.14. Show that if V is a normed vector space, then the weak topology on V and the weak* topology on V^* are both Hausdorff. (*Hint*: You will need the Hahn-Banach theorem.) In particular, we conclude the important fact that weak and weak* limits, when they exist, are unique.

The following exercise shows that the strong, weak, and weak* topologies can all differ from each other.

Exercise 1.9.15. Let $V := c_0(\mathbf{N})$, thus $V^* \equiv \ell^1(\mathbf{N})$ and $V^{**} \equiv \ell^\infty(\mathbf{N})$. Let $e_1, e_2, \ldots$ be the standard basis of either V, V^*, or V^{**}.

- Show that the sequence $e_1, e_2, \ldots$ converges weakly in V to zero, but does not converge strongly in V.
- Show that the sequence $e_1, e_2, \ldots$ converges in the weak* sense in V^* to zero, but does not converge in the weak or strong senses in V^*.
- Show that the sequence $\sum_{m=n}^\infty e_m$ for $n = 1, 2, \ldots$ converges in the weak* topology of V^{**} to zero, but does not converge in the weak or strong senses. (*Hint*: use a generalised limit functional).

Remark 1.9.12. Recall from Exercise 1.7.11 that sequences in $V^* \equiv \ell^1(\mathbf{N})$ which converge in the weak topology, also converge in the strong topology. We caution however that the two topologies are not quite equivalent; for instance, the open unit ball in $\ell^1(\mathbf{N})$ is open in the strong topology, but not in the weak.

Exercise 1.9.16. Let V be a normed vector space, and let E be a subset of V. Show that the following are equivalent:

- E is strongly bounded (i.e. E is contained in a ball).
- E is weakly bounded (i.e. $\lambda(E)$ is bounded for all $\lambda \in V^*$).

(*Hint*: use the Hahn-Banach theorem and the uniform boundedness principle.) Similarly, if F is a subset of V^*, and V is a Banach space, show that F is strongly bounded if and only if F is weak* bounded (i.e. $\{\lambda(x) : \lambda \in F\}$ is bounded for each $x \in V$).) Conclude in particular that any sequence which is weakly convergent in V or weak* convergent in V^* is necessarily bounded.

Exercise 1.9.17. Let V be a Banach space, and let $x_n \in V$ converge weakly to a limit $x \in V$. Show that the sequence x_n is bounded, and

$$\|x\|_V \leq \liminf_{n\to\infty} \|x_n\|_V.$$

Observe from Exercise 1.9.15 that strict inequality can hold (cf. Fatou's lemma, Theorem 1.1.21). Similarly, if $\lambda_n \in V^*$ converges in the weak* topology to a limit $\lambda \in V^*$, show that the sequence λ_n is bounded and that

$$\|\lambda\|_{V^*} \leq \liminf_{n\to\infty} \|\lambda_n\|_{V^*}.$$

Again, construct an example to show that strict inequality can hold. Thus we see that weak or weak* limits can lose mass in the limit, as opposed to strong limits (note from the triangle inequality that if x_n converges strongly to x, then $\|x_n\|_V$ converges to $\|x\|_V$).

Exercise 1.9.18. Let H be a Hilbert space, and let $x_n \in H$ converge weakly to a limit $x \in H$. Show that the following statements are equivalent:

- x_n converges strongly to x.
- $\|x_n\|$ converges to $\|x\|$.

Exercise 1.9.19. Let H be a separable Hilbert space. We say that a sequence $x_n \in H$ *converges in the Césaro sense* to a limit $x \in H$ if $\frac{1}{N}\sum_{n=1}^{N} x_n$ converges strongly to x as $n \to \infty$.

- Show that if x_n converges strongly to x, then it also converges in the Césaro sense to x.
- Give examples to show that weak convergence does not imply Césaro convergence, and vice versa. On the other hand,

if a sequence x_n converges both weakly and in the Césaro sense, show that the weak limit is necessarily equal to the Césaro limit.

- Show that if a bounded sequence converges in the Césaro sense to a limit x, then some subsequence converges weakly to x.
- Show that a sequence x_n converges weakly to x if and only if every subsequence has a further subsequence that converges in the Césaro sense to x.

Exercise 1.9.20. Let V be a Banach space. Show that the closed unit ball in V is also closed in the weak topology, and the closed unit ball in V^* is closed in the weak* topology.

Exercise 1.9.21. Let V be a Banach space. Show that the weak* topology on V^* is complete.

Exercise 1.9.22. Let V be a normed vector space, let W be a subspace of V which is closed in the strong topology of V.

- Show that W is closed in the weak topology of V.
- If $w_n \in W$ is a sequence and $w \in W$, show that w_n converges to w in the weak topology of W if and only if it converges to w in the weak topology of V. (Because of this fact, we can often refer to "the weak topology" without specifying the ambient space precisely.)

Exercise 1.9.23. Let $V := c_0(\mathbf{N})$ with the uniform (i.e. ℓ^∞) norm, and identify the dual space V^* with $\ell^1(\mathbf{N})$ in the usual manner.

- Show that a sequence $x_n \in c_0(\mathbf{N})$ converges weakly to a limit $x \in c_0(\mathbf{N})$ if and only if the x_n are bounded in $c_0(\mathbf{N})$ and converge pointwise to x.
- Show that a sequence $\lambda_n \in \ell^1(\mathbf{N})$ converges in the weak* topology to a limit $\lambda \in \ell^1(\mathbf{N})$ if and only if the λ_n are bounded in $\ell^1(\mathbf{N})$ and converge pointwise to λ.
- Show that the weak topology in $c_0(\mathbf{N})$ is not complete.

(More generally, it may help to think of the weak and weak* topologies as being analogous to pointwise convergence topologies.)

One of the main reasons why we use the weak and weak* topologies in the first place is that they have much better compactness properties than the strong topology, thanks to the *Banach-Alaoglu theorem*:

Theorem 1.9.13 (Banach-Alaoglu theorem)**.** *Let V be a normed vector space. Then the closed unit ball of V^* is compact in the weak* topology.*

This result should be contrasted with Exercise 1.9.12.

Proof. Let's say V is a complex vector space (the case of real vector spaces is of course analogous). Let B^* be the closed unit ball of V^*, then any linear functional $\lambda \in B^*$ maps the closed unit ball B of V into the disk $D := \{z \in \mathbf{C} : |z| \leq 1\}$. Thus one can identify B^* with a subset of D^B, the space of functions from B to D. One easily verifies that the weak* topology on B^* is nothing more than the product topology of D^B restricted to B^*. Also, one easily shows that B^* is closed in D^B. But by Tychonoff's theorem, D^B is compact, and so B^* is compact also. □

One should caution that the Banach-Alaoglu theorem does *not* imply that the space V^* is locally compact in the weak* topology, because the norm ball in V has empty interior in the weak* topology unless V is finite dimensional. In fact, we have the following result of Riesz:

Exercise 1.9.24. Let V be a locally compact Hausdorff topological vector space. Show that V is finite dimensional. (*Hint*: If V is locally compact, then there exists an open neighbourhood U of the origin whose closure is compact. Show that $U \subset W + \frac{1}{2}U$ for some finite-dimensional subspace W, where $W + \frac{1}{2}U := \{w + \frac{1}{2}u : w \in W, u \in U\}$. Iterate this to conclude that $U \subset W + \varepsilon U$ for any $\varepsilon > 0$. On the other hand, use the compactness of $\overline{U}$ to show that for any point $x \in V \backslash W$ there exists $\varepsilon > 0$ such that $x - \varepsilon U$ is disjoint from W. Conclude that $U \subset W$ and thence that $V = W$.)

The sequential version of the Banach-Alaoglu theorem is also of importance (particularly in PDE):

Theorem 1.9.14 (Sequential Banach-Alaoglu theorem)**.** *Let V be a separable normed vector space. Then the closed unit ball of V^* is sequentially compact in the weak* topology.*

Proof. The functionals in B^* are uniformly bounded and uniformly equicontinuous on B, which by hypothesis has a countable dense subset Q. By the sequential Tychonoff theorem, any sequence in B^* then has a subsequence which converges pointwise on Q, and thus converges pointwise on B by Exercise 1.8.28, and thus converges in the weak* topology. But as B^* is closed in this topology, we conclude that B^* is sequentially compact as required. □

Remark 1.9.15. One can also deduce the sequential Banach-Alaoglu theorem from the general Banach-Alaoglu theorem by observing that the weak* topology on the dual of a separable space is metrisable. The sequential Banach-Alaoglu theorem can break down for non-separable spaces. For instance, the closed unit ball in $\ell^\infty(\mathbf{N})$ is not sequentially compact in the weak* topology, basically because the space $\beta\mathbf{N}$ of ultrafilters is not sequentially compact (see Exercise 2.3.12 of *Poincaré's Legacies, Vol. I*).

If V is reflexive, then the weak topology on V is identical to the weak* topology on $(V^*)^*$. We thus have

Corollary 1.9.16. *If V is a reflexive normed vector space, then the closed unit ball in V is weakly compact, and (if V^* is separable) is also sequentially weakly compact.*

Remark 1.9.17. If V is a normed vector space that is not separable, then one can show that V^* is not separable either. Indeed, using transfinite induction on first uncountable ordinal, one can construct an uncountable proper chain of closed separable subspaces of the inseparable space V, which by the Hahn-Banach theorem induces an uncountable proper chain of closed subspaces on V^*, which is not compatible with separability. As a consequence, a reflexive space is separable if and only if its dual is separable[11].

[11] On the other hand, separable spaces can have non-separable duals; consider $\ell^1(\mathbf{N})$, for instance.

In particular, any bounded sequence in a reflexive separable normed vector space has a weakly convergent subsequence. This fact leads to the very useful *weak compactness* method in PDE and calculus of variations, in which a solution to a PDE or variational problem is constructed by first constructing a bounded sequence of "near-solutions" or "near-extremisers" to the PDE or variational problem, and then extracting a weak limit. However, it is important to caution that weak compactness can fail for non-reflexive spaces; indeed, for such spaces the closed unit ball in V may not even be weakly complete, let alone weakly compact, as already seen in Exercise 1.9.23. Thus, one should be cautious when applying the weak compactness method to a non-reflexive space such as L^1 or L^∞. (On the other hand, weak* compactness does not need reflexivity, and is thus safer to use in such cases.)

In later notes we will see that the (sequential) Banach-Alaoglu theorem will combine very nicely with the Riesz representation theorem for measures (Section 1.10.2), leading in particular to *Prokhorov's theorem* (Exercise 1.10.29).

1.9.4. The strong and weak operator topologies. Now we turn our attention from function spaces to spaces of operators. Recall that if X and Y are normed vector spaces, then $B(X \to Y)$ is the space of bounded linear transformations from X to Y. This is a normed vector space with the operator norm

$$\|T\|_{\mathrm{op}} := \sup\{\|Tx\|_Y : \|x\|_X \leq 1\}.$$

This norm induces the *operator norm topology* on $B(X \to Y)$. Unfortunately, this topology is so strong that it is difficult for a sequence of operators $T_n \in B(X \to Y)$ to converge to a limit; for this reason, we introduce two weaker topologies.

Definition 1.9.18 (Strong and weak operator topologies)**.** Let X, Y be normed vector spaces. The *strong operator topology* on $B(X \to Y)$ is the topology induced by the seminorms $T \mapsto \|Tx\|_Y$ for all $x \in X$. The *weak operator topology* on $B(X \to Y)$ is the topology induced by the seminorms $T \mapsto |\lambda(Tx)|$ for all $x \in X$ and $\lambda \in Y^*$.

Note that a sequence $T_n \in B(X \to Y)$ converges in the strong operator topology to a limit $T \in B(X \to Y)$ if and only if $T_n x \to Tx$ strongly in Y for all $x \in X$, and T_n converges in the weak operator topology. (In contrast, T_n converges to T in the operator norm topology if and only if $T_n x$ converges to Tx *uniformly* on bounded sets.) One easily sees that the weak operator topology is weaker than the strong operator topology, which in turn is (somewhat confusingly) weaker than the operator norm topology.

Example 1.9.19. When X is the scalar field, then $B(X \to Y)$ is canonically isomorphic to Y. In this case, the operator norm and strong operator topology coincide with the strong topology on Y, and the weak operator norm topology coincides with the weak topology on Y. Meanwhile, $B(Y \to X)$ coincides with Y^*, and the operator norm topology coincides with the strong topology on Y^*, while the strong and weak operator topologies correspond with the weak* topology on Y^*.

We can rephrase the uniform boundedness principle for convergence (Corollary 1.7.7) as follows:

Proposition 1.9.20 (Uniform boundedness principle)**.** *Let $T_n \in B(X \to Y)$ be a sequence of bounded linear operators from a Banach space X to a normed vector space Y, let $T \in B(X \to Y)$ be another bounded linear operator, and let D be a dense subspace of X. Then the following are equivalent:*

- *T_n converges in the strong operator topology of $B(X \to Y)$ to T.*
- *T_n is bounded in the operator norm (i.e. $\|T_n\|_{\mathrm{op}}$ is bounded), and the restriction of T_n to D converges in the strong operator topology of $B(D \to Y)$ to the restriction of T to D.*

Exercise 1.9.25. Let the hypotheses be as in Proposition 1.9.20, but now assume that Y is also a Banach space. Show that the conclusion of Proposition 1.9.20 continues to hold if "strong operator topology" is replaced by "weak operator topology".

Exercise 1.9.26. Show that the operator norm topology, strong operator topology, and weak operator topology, are all Hausdorff. As

these topologies are nested, we thus conclude that it is not possible for a sequence of operators to converge to one limit in one of these topologies and to converge to a different limit in another.

Example 1.9.21. Let $X = L^2(\mathbf{R})$, and for each $t \in \mathbf{R}$, let $T_t : X \to X$ be the translation operator by t: $T_t f(x) := f(x-t)$. If f is continuous and compactly supported, then (e.g. from dominated convergence) we see that $T_t f \to f$ in L^2 as $t \to 0$. Since the space of continuous and compactly supported functions is dense in $L^2(\mathbf{R})$, this implies (from the above proposition, with some obvious modifications to deal with the continuous parameter t instead of the discrete parameter n) that T_t converges in the strong operator topology (and hence weak operator topology) to the identity. On the other hand, T_t does not converge to the identity in the operator norm topology. Indeed, observe for any $t > 0$ that $\|(T_t - I)1_{[0,t]}\|_{L^2(\mathbf{R})} = \sqrt{2}\|1_{[0,t]}\|_{L^2(\mathbf{R})}$, and thus $\|T_t - I\|_{\mathrm{op}} \geq \sqrt{2}$.

In a similar vein, T_t does not converge to anything in the strong operator topology (and hence does not converge in the operator norm topology either) in the limit $t \to \infty$, since $T_t 1_{[0,1]}$ (say) does not converge strongly in L^2. However, one easily verifies that $\langle T_t f, g\rangle \to 0$ as $t \to \infty$ for any compactly supported $f, g \in L^2(\mathbf{R})$, and hence for all $f, g \in L^2(\mathbf{R})$ by the usual limiting argument, and hence T_t converges in the weak operator topology to zero.

The following exercise may help clarify the relationship between the operator norm, strong operator, and weak operator topologies.

Exercise 1.9.27. Let H be a Hilbert space, and let $T_n \in B(H \to H)$ be a sequence of bounded linear operators.

- Show that $T_n \to 0$ in the operator norm topology if and only if $\langle T_n x_n, y_n\rangle \to 0$ for any bounded sequences $x_n, y_n \in H$.
- Show that $T_n \to 0$ in the strong operator topology if and only if $\langle T_n x_n, y_n\rangle \to 0$ for any convergent sequence $x_n \in H$ and any bounded sequence $y_n \in H$.
- Show that $T_n \to 0$ in the weak operator topology if and only if $\langle T_n x_n, y_n\rangle \to 0$ for any convergent sequences $x_n, y_n \in H$.

- Show that $T_n \to 0$ in the operator norm (resp. weak operator) topology if and only if $T_n^\dagger \to 0$ in the operator norm (resp. weak operator) topology. Give an example to show that the corresponding claim for the strong operator topology is false.

There is a counterpart of the Banach-Alaoglu theorem (and its sequential analogue), at least in the case of Hilbert spaces:

Exercise 1.9.28. Let H, H' be Hilbert spaces. Show that the closed unit ball (in the operator norm) in $B(H \to H')$ is compact in the weak operator topology. If H and H' are separable, show that $B(H \to H')$ is sequentially compact in the weak operator topology.

The behaviour of convergence in various topologies with respect to composition is somewhat complicated, as the following exercise shows.

Exercise 1.9.29. Let H be a Hilbert space, let $S_n, T_n \in B(H \to H)$ be sequences of operators, and let $S \in B(H \to H)$ be another operator.

- If $T_n \to 0$ in the operator norm (resp. strong operator or weak operator) topology, show that $ST_n \to 0$ and $T_n S \to 0$ in the operator norm (resp. strong operator or weak operator) topology.
- If $T_n \to 0$ in the operator norm topology, and S_n is bounded in the operator norm topology, show that $S_n T_n \to 0$ and $T_n S_n \to 0$ in the operator norm topology.
- If $T_n \to 0$ in the strong operator topology, and S_n is bounded in the operator norm topology, show that $S_n T_n \to 0$ in the strong operator norm topology.
- Give an example where $T_n \to 0$ in the strong operator topology, and $S_n \to 0$ in the weak operator topology, but $T_n S_n$ does not converge to zero even in the weak operator topology.

Exercise 1.9.30. Let H be a Hilbert space. An operator $T \in B(H \to H)$ is said to be *finite rank* if its image $T(H)$ is finite dimensional. T is said to be *compact* if the image of the unit ball is

precompact. Let $K(H \to H)$ denote the space of compact operators on H.

- Show that $T \in B(H \to H)$ is compact if and only if it is the limit of finite rank operators in the operator norm topology. Conclude in particular that $K(H \to H)$ is a closed subset of $B(H \to H)$ in the operator norm topology.
- Show that an operator $T \in B(H \to H)$ is compact if and only if $T^\dagger$ is compact.
- If H is separable, show that every $T \in B(H \to H)$ is the limit of finite rank operators in the strong operator topology.
- If $T \in K(H \to H)$, show that T maps weakly convergent sequences to strongly convergent sequences. (This property is known as *complete continuity*.)
- Show that $K(H \to H)$ is a subspace of $B(H \to H)$, which is closed with respect to left and right multiplication by elements of $B(H \to H)$. (In other words, the space of compact operators is an two-ideal in the algebra of bounded operators.)

The weak operator topology plays a particularly important role on the theory of *von Neumann algebras*, which we will not discuss here.

Notes. This lecture first appeared at `terrytao.wordpress.com/2009/02/21`. Thanks to Eric, etale, less than epsilon, Matt Daws, PDEBeginner, Sebastian Scholtes, Xiaochuan Liu, Yasser Taima, and anonymous commenters for corrections.

1.10. Continuous functions on locally compact Hausdorff spaces

A key theme in real analysis is that of studying general functions $f : X \to \mathbf{R}$ or $f : X \to \mathbf{C}$ by first approximating them by "simpler" or "nicer" functions. But the precise class of "simple" or "nice" functions may vary from context to context. In measure theory, for instance, it is common to approximate measurable functions by indicator functions or simple functions. But in other parts of analysis, it

is often more convenient to approximate rough functions by continuous or smooth functions (perhaps with compact support, or some other decay condition), or by functions in some algebraic class, such as the class of polynomials or *trigonometric polynomials*.

In order to approximate rough functions by more continuous ones, one of course needs tools that can generate continuous functions with some specified behaviour. The two basic tools for this are *Urysohn's lemma*, which approximates indicator functions by continuous functions, and the *Tietze extension theorem*, which extends continuous functions on a subdomain to continuous functions on a larger domain. An important consequence of these theorems is the *Riesz representation theorem* for linear functionals on the space of compactly supported continuous functions, which describes such functionals in terms of *Radon measures*.

Sometimes, approximation by continuous functions is not enough; one must approximate continuous functions in turn by an even smoother class of functions. A useful tool in this regard is the *Stone-Weierstrass theorem*, that generalises the classical *Weierstrass approximation theorem* to more general algebras of functions.

As an application of this theory (and of many of the results accumulated in previous lecture notes), we will present (in an optional section) the commutative *Gelfand-Neimark theorem* classifying all commutative unital C^**-algebras*.

1.10.1. Urysohn's lemma. Let X be a topological space. An indicator function 1_E in this space will not typically be a continuous function (indeed, if X is connected, this only happens when E is the empty set or the whole set). Nevertheless, for certain topological spaces, it is possible to approximate an indicator function by a continuous function, as follows.

Lemma 1.10.1 (Urysohn's lemma)**.** *Let X be a topological space. Then the following are equivalent:*

(i) *Every pair of disjoint closed sets K, L in X can be separated by disjoint open neighbourhoods $U \supset K$, $V \supset L$.*

(ii) *For every closed set K in X and every open neighbourhood U of K, there exists an open set V and a closed set L such that $K \subset V \subset L \subset U$.*

(iii) *For every pair of disjoint closed sets K, L in X, there exists a continuous function $f : X \to [0,1]$ which equals 1 on K and 0 on L.*

(iv) *For every closed set K in X and every open neighbourhood U of K, there exists a continuous function $f : X \to [0,1]$ such that $1_K(x) \leq f(x) \leq 1_U(x)$ for all $x \in X$.*

A topological space which obeys any (and hence all) of (i-iv) is known as a *normal space*; definition (i) is traditionally taken to be the standard definition of normality. We will give some examples of normal spaces shortly.

Proof. The equivalence of (iii) and (iv) is clear, as the complement of a closed set is an open set and vice versa. The equivalence of (i) and (ii) follows similarly.

To deduce (i) from (iii), let K, L be disjoint closed sets, let f be as in (iii), and let U, V be the open sets $U := \{x \in X : f(x) > 2/3\}$ and $V := \{x \in X : f(x) < 1/3\}$.

The only remaining task is to deduce (iv) from (ii). Suppose we have a closed set $K = K_1$ and an open set $U = U_0$ with $K_1 \subset U_0$. Applying (ii), we can find an open set $U_{1/2}$ and a closed set $K_{1/2}$ such that

$$K_1 \subset U_{1/2} \subset K_{1/2} \subset U_0.$$

Applying (ii) two more times, we can find more open sets $U_{1/4}, U_{3/4}$ and closed sets $K_{1/4}, K_{3/4}$ such that

$$K_1 \subset U_{3/4} \subset K_{3/4} \subset U_{1/2} \subset K_{1/2} \subset U_{1/4} \subset K_{1/4} \subset U_0.$$

Iterating this process, we can construct open sets U_q and closed sets K_q for every dyadic rational $q = a/2^n$ in $(0,1)$ such that $U_q \subset K_q$ for all $0 < q < 1$, and $K_{q'} \subset U_q$ for any $0 \leq q < q' \leq 1$.

If we now define $f(x) := \sup\{q : x \in U_q\} = \inf\{q : x \in K_q\}$, where q ranges over dyadic rationals between 0 and 1, and with the convention that the empty set has sup 1 and inf 0, one easily verifies

that the sets $\{f(x) > \alpha\} = \bigcup_{q>\alpha} U_q$ and $\{f(x) < \alpha\} = \bigcup_{q<\alpha} X \backslash K_q$ are open for every real number α, and so f is continuous as required. □

The definition of normality is very similar to the *Hausdorff property*, which separates pairs of points instead of closed sets. Indeed, if every point in X is closed (a property known as the T_1 *property*), then normality clearly implies the Hausdorff property. The converse is not always true, but (as the term suggests) in practice most topological spaces one works with in real analysis are normal. For instance:

Exercise 1.10.1. Show that every metric space is normal.

Exercise 1.10.2. Let X be a Hausdorff space.

- Show that a compact subset of X and a point disjoint from that set can always be separated by open neighbourhoods.
- Show that a pair of disjoint compact subsets of X can always be separated by open neighbourhoods.
- Show that every compact Hausdorff space is normal.

Exercise 1.10.3. Let $\mathbf{R}$ be the real line with the usual topology $\mathcal{F}$, and let $\mathcal{F}'$ be the topology on $\mathbf{R}$ generated by $\mathcal{F}$ and the rationals. Show that $(\mathbf{R}, \mathcal{F}')$ is Hausdorff, with every point closed, but is not normal.

The above example was a simple but somewhat artificial example of a non-normal space. One can create more "natural" examples of non-normal Hausdorff spaces (with every point closed), but establishing non-normality becomes more difficult. The following example is due to Stone[**St1948**].

Exercise 1.10.4. Let $\mathbf{N}^{\mathbf{R}}$ be the space of natural number-valued tuples $(n_x)_{x\in\mathbf{R}}$, endowed with the product topology (i.e. the topology of pointwise convergence).

- Show that $\mathbf{N}^{\mathbf{R}}$ is Hausdorff, and every point is closed.
- For $j = 1, 2$, let K_j be the set of all tuples $(n_x)_{x\in\mathbf{R}}$ such that $n_x = j$ for all x outside of a countable set, and such that $x \mapsto n_x$ is injective on this finite set (i.e. there do not exist

distinct x, x' such that $n_x = n_{x'} \neq j$). Show that K_1, K_2 are disjoint and closed.

- Show that given any open neighbourhood U of K_1, there exists disjoint finite subsets $A_1, A_2, \ldots$ of $\mathbf{R}$ and an injective function $f : \bigcup_{i=1}^{\infty} A_i \to \mathbf{N}$ such that for any $j \geq 0$, any $(m_x)_{x \in \mathbf{R}}$ such that $m_x = f(x)$ for all $x \in A_1 \cup \ldots \cup A_j$ and is identically 1 on A_{j+1}, lies in U.
- Show that any open neighbourhood of K_1 and any open neighbourhood of K_2 necessarily intersect, and so $\mathbf{N}^{\mathbf{R}}$ is not normal.
- Conclude that $\mathbf{R}^{\mathbf{R}}$ with the product topology is not normal.

The property of being normal is a topological one, thus if one topological space is normal, then any other topological space homeomorphic to it is also normal. However, (unlike, say, the Hausdorff property), the property of being normal is not preserved under passage to subspaces:

Exercise 1.10.5. Given an example of a subspace of a normal space which is not normal. (*Hint*: use Exercise 1.10.4, possibly after replacing $\mathbf{R}$ with a homeomorphic equivalent.)

Let $C_c(X \to \mathbf{R})$ be the space of real continuous compactly supported functions on X. Urysohn's lemma generates a large number of useful elements of $C_c(X \to \mathbf{R})$, in the case when X is *locally compact Hausdorff* (LCH):

Exercise 1.10.6. Let X be a locally compact Hausdorff space, let K be a compact set, and let U be an open neighbourhood of K. Show that there exists $f \in C_c(X \to \mathbf{R})$ such that $1_K(x) \leq f(x) \leq 1_U(x)$ for all $x \in X$. (*Hint*: First use the local compactness of X to find a neighbourhood of K with compact closure; then restrict U to this neighbourhood. The closure of U is now a compact set; restrict everything to this set, at which point the space becomes normal.)

One consequence of this exercise is that $C_c(X \to \mathbf{R})$ tends to be dense in many other function spaces. We give an important example here:

Definition 1.10.2 (Radon measure)**.** Let X be a locally compact Hausdorff space that is also *σ-compact*, and let $\mathcal{B}$ be the Borel σ-algebra. An (unsigned) *Radon measure* is a unsigned measure $\mu : \mathcal{B} \to \mathbf{R}^+$ with the following properties:

- (Local finiteness) For any compact subset K of X, $\mu(K)$ is finite.
- (Outer regularity) For any Borel set E of X, $\mu(E) = \inf\{\mu(U) : U \supset E; U \text{ open}\}$.
- (Inner regularity) For any Borel set E of X, $\mu(E) = \sup\{\mu(K) : K \subset E; K \text{ compact}\}$.

Example 1.10.3. Lebesgue measure m on $\mathbf{R}^n$ is a Radon measure, as is any absolutely continuous unsigned measure m_f, where $f \in L^1(\mathbf{R}^n, dm)$. More generally, if μ is Radon and ν is a finite unsigned measure which is absolutely continuous with respect to μ, then ν is Radon. On the other hand, counting measure on $\mathbf{R}^n$ is not Radon (it is not locally finite). It is possible to define Radon measures on Hausdorff spaces that are not σ-compact or locally compact, but the theory is more subtle and will not be considered here. We will study Radon measures more thoroughly in the next section.

Proposition 1.10.4. *Let X be a locally compact Hausdorff space which is also σ-compact, and let μ be a Radon measure on X. Then for any $0 < p < \infty$, $C_c(X \to \mathbf{R})$ is a dense subset in (real-valued) $L^p(X, \mu)$. In other words, every element of $L^p(X, \mu)$ can be expressed as a limit (in $L^p(X, \mu)$) of continuous functions of compact support.*

Proof. Since continuous functions of compact support are bounded, and compact sets have finite measure, we see that $C_c(X)$ is a subspace of $L^p(X, \mu)$. We need to show that the closure $\overline{C_c(X)}$ of this space contains all of $L^p(X, \mu)$.

Let K be a compact set, and let $E \subset K$ be a Borel set, then E has finite measure. Applying inner and outer regularity, we can find a sequence of compact sets $K_n \subset E$ and open sets $U_n \supset E$ such that $\mu(E \backslash K_n), \mu(U_n \backslash E) \to 0$. Applying Exercise 1.10.6, we can then find $f_n \in C_c(X \to \mathbf{R})$ such that $1_{K_n}(x) \leq f_n(x) \leq 1_{U_n}(x)$. In particular, this implies (by the *squeeze theorem*) that f_n converges in $L^p(X, \mu)$

to 1_E (here we use the finiteness of p); thus 1_E lies in $\overline{C_c(X \to \mathbf{R})}$ for any measurable subset E of K. By linearity, all simple functions supported on K also lie in $\overline{C_c(X \to \mathbf{R})}$; taking closures, we see that any L^p function supported in K also lies in $\overline{C_c(X \to \mathbf{R})}$. As X is σ-finite, one can express any non-negative L^p function as a monotone limit of compactly supported functions, and thus every non-negative L^p function lies in $\overline{C_c(X \to \mathbf{R})}$, and thus all L^p functions lie in this space, and the claim follows. □

Of course, the real-valued version of the above proposition immediately implies a complex-valued analogue. On the other hand, the claim fails when $p = \infty$:

Exercise 1.10.7. Let X be a locally compact Hausdorff space that is σ-compact, and let μ be a Radon measure. Show that the closure of $C_c(X \to \mathbf{R})$ in $L^\infty(X, \mu)$ is $C_0(X \to \mathbf{R})$, the space of continuous real-valued functions which vanish at infinity (i.e. for every $\varepsilon > 0$ there exists a compact set K such that $|f(x)| \leq \varepsilon$ for all $x \in K$). Thus, in general, $C_c(X \to \mathbf{R})$ is not dense in $L^\infty(X, \mu)$.

Thus we see that the L^∞ norm is strong enough to preserve continuity in the limit, whereas the L^p norms are (locally) weaker and permit discontinuous functions to be approximated by continuous ones.

Another important consequence of Urysohn's lemma is the *Tietze extension theorem*:

Theorem 1.10.5 (Tietze extension theorem)**.** *Let X be a normal topological space, let $[a, b] \subset \mathbf{R}$ be a bounded interval, let K be a closed subset of X, and let $f : K \to [a, b]$ be a continuous function. Then there exists a continuous function $\tilde{f} : X \to [a, b]$ which extends f, i.e. $\tilde{f}(x) = f(x)$ for all $x \in K$.*

Proof. It suffices to find an continuous extension $\tilde{f} : X \to \mathbf{R}$ taking values in the real line rather than in $[a, b]$, since one can then replace $\tilde{f}$ by $\min(\max(\tilde{f}, a), b)$ (note that min and max are continuous operations).

Let $T : BC(X \to \mathbf{R}) \to BC(K \to \mathbf{R})$ be the restriction map $Tf := f \restriction_K$. This is clearly a continuous linear map; our task is to show that it is surjective, i.e. to find a solution to the equation

$Tg = f$ for each $f \in BC(X \to \mathbf{R})$. We do this by the standard analysis trick of getting an approximate solution to $Tg = f$ first, and then using iteration to boost the approximate solution to an exact solution.

Let $f : K \to \mathbf{R}$ have sup norm 1, thus f takes values in $[-1, 1]$. To solve the problem $Tg = f$, we approximate f by $\frac{1}{3}1_{f \geq 1/3} - \frac{1}{3}1_{f \leq -1/3}$. By Urysohn's lemma, we can find a continuous function $g : X \to [-1/3, 1/3]$ such that $g = 1/3$ on the closed set $\{x \in K : f \geq 1/3\}$ and $g = -1/3$ on the closed set $\{x \in K : f \leq -1/3\}$. Now, Tg is not quite equal to f; but observe from construction that $f - Tg$ has sup norm $2/3$.

Scaling this fact, we conclude that, given any $f \in BC(K \to \mathbf{R})$, we can find a decomposition $f = Tg + f'$, where $\|g\|_{BC(X \to \mathbf{R})} \leq \frac{1}{3}\|f\|_{BC(K \to \mathbf{R})}$ and $\|f'\|_{BC(K \to \mathbf{R})} \leq \frac{2}{3}\|f\|_{BC(K \to \mathbf{R})}$.

Starting with any $f = f_0 \in BC(K \to \mathbf{R})$, we can now iterate this construction to express $f_n = Tg_n + f_{n+1}$ for all $n = 0, 1, 2, \ldots$, where $\|f_n\|_{BC(K \to R)} \leq (\frac{2}{3})^n \|f\|_{BC(K \to \mathbf{R})}$ and $\|g_n\|_{BC(X \to \mathbf{R})} \leq \frac{1}{3}(\frac{2}{3})^n \|f\|_{BC(K \to \mathbf{R})}$. As $BC(X \to \mathbf{R})$ is a Banach space, we see that $\sum_{n=0}^{\infty} g_n$ converges absolutely to some limit $g \in BC(X \to \mathbf{R})$, and that $Tg = f$, as desired. □

Remark 1.10.6. Observe that Urysohn's lemma can be viewed the special case of the Tietze extension theorem when K is the union of two disjoint closed sets, and f is equal to 1 on one of these sets and equal to 0 on the other.

Remark 1.10.7. One can extend the Tietze extension theorem to finite-dimensional vector spaces: if K is a closed subset of a normal vector space X and $f : K \to \mathbf{R}^n$ is bounded and continuous, then one has a bounded continuous extension $\overline{f} : K \to \mathbf{R}^n$. Indeed, one simply applies the Tietze extension theorem to each component of f separately. However, if the range space is replaced by a space with a non-trivial topology, then there can be topological obstructions to continuous extension. For instance, a map $f : \{0, 1\} \to Y$ from a two-point set into a topological space Y is always continuous, but can be extended to a continuous map $\tilde{f} : \mathbf{R} \to Y$ if and only if $f(0)$ and $f(1)$ lie in the same *path-connected* component of Y. Similarly, if $f : S^1 \to Y$ is a map from the unit circle into a topological space Y,

then a continuous extension from S^1 to $\mathbf{R}^2$ exists if and only if the closed curve $f : S^1 \to Y$ is *contractible* to a point in Y. These sorts of questions require the machinery of algebraic topology to answer them properly, and are beyond the scope of this course.

There are analogues for the Tietze extension theorem in some other categories of functions. For instance, in the Lipschitz category, we have

Exercise 1.10.8. Let X be a metric space, let K be a subset of X, and let $f : K \to \mathbf{R}$ be a *Lipschitz continuous* map with some Lipschitz constant A (thus $|f(x) - f(y)| \leq Ad(x,y)$ for all $x, y \in K$). Show that there exists an extension $\tilde{f} : X \to \mathbf{R}$ of f which is Lipschitz continuous with the same Lipschitz constant A. (*Hint*: A "greedy" algorithm will work here: pick $\tilde{f}$ to be as large as one can get away with (or as small as one can get away with).)

One can also remove the requirement that the function f be bounded in the Tietze extension theorem:

Exercise 1.10.9. Let X be a normal topological space, let K be a closed subset of X, and let $f : K \to \mathbf{R}$ be a continuous map (not necessarily bounded). Then there exists an extension $\tilde{f} : X \to \mathbf{R}$ of f which is still continuous. (*Hint*: first "compress" f to be bounded by working with, say, $\arctan(f)$ (other choices are possible), and apply the usual Tietze extension theorem. There will be some sets in which one cannot invert the compression function, but one can deal with this by a further appeal to Urysohn's lemma to damp the extension out on such sets.)

There is also a *locally compact Hausdorff* version of the Tietze extension theorem:

Exercise 1.10.10. Let X be locally compact Hausdorff, let K be compact, and let $f \in C(K \to \mathbf{R})$. Then there exists $\tilde{f} \in C_c(X \to \mathbf{R})$ which extends f.

Proposition 1.10.4 shows that measurable functions in L^p can be approximated by continuous functions of compact support (cf. *Littlewood's second principle*). Another approximation result in a similar spirit is *Lusin's theorem*:

Theorem 1.10.8 (Lusin's theorem)**.** *Let X be a locally compact Hausdorff space that is σ-compact, and let μ be a Radon measure. Let $f : X \to \mathbf{R}$ be a measurable function supported on a set of finite measure, and let $\varepsilon > 0$. Then there exists $g \in C_c(X \to \mathbf{R})$ which agrees with f outside of a set of measure at most ε.*

Proof. Observe that as f is finite everywhere, it is bounded outside of a set of arbitrarily small measure. Thus we may assume without loss of generality that f is bounded. Similarly, as X is σ-compact (or by inner regularity), the support of f differs from a compact set by a set of arbitrarily small measure; so we may assume that f is also supported on a compact set K. By Theorem 1.10.5, it then suffices to show that f is continuous on the complement of an open set of arbitrarily small measure; by outer regularity, we may delete the adjective "open" from the preceding sentence.

As f is bounded and compactly supported, f lies in $L^p(X, \mu)$ for every $0 < p < \infty$, and using Proposition 1.10.4 and Chebyshev's inequality, it is not hard to find, for each $n = 1, 2, \ldots$, a function $f_n \in C_c(X \to \mathbf{R})$ which differs from f by at most $1/2^n$ outside of a set of measure at most $\varepsilon/2^{n+2}$ (say). In particular, f_n converges uniformly to f outside of a set of measure at most $\varepsilon/4$, and f is therefore continuous outside this set. The claim follows. □

Another very useful application of Urysohn's lemma is to create *partitions of unity.*

Lemma 1.10.9 (Partitions of unity)**.** *Let X be a normal topological space, and let $(K_\alpha)_{\alpha \in A}$ be a collection of closed sets that cover X. For each $\alpha \in A$, let U_α be an open neighbourhood of K_α, which are finitely overlapping in the sense that each $x \in X$ belongs to at most finitely many of the U_α. Then there exists a continuous function $f_\alpha : X \to [0, 1]$ supported on U_α for each $\alpha \in A$ such that $\sum_{\alpha \in A} f_\alpha(x) = 1$ for all $x \in X$.*

If X is locally compact Hausdorff instead of normal, and the K_α are compact, then one can take the f_α to be compactly supported.

Proof. Suppose first that X is normal. By Urysohn's lemma, one can find a continuous function $g_\alpha : X \to [0, 1]$ for each $\alpha \in A$ which

is supported on U_α and equals 1 on the closed set K_α. Observe that the function $g := \sum_{\alpha \in A} g_\alpha$ is well-defined, continuous and bounded below by 1. The claim then follows by setting $f_\alpha := g_\alpha / g$.

The final claim follows by using Exercise 1.10.6 instead of Urysohn's lemma. □

Exercise 1.10.11. Let X be a topological space. A function $f : X \to \mathbf{R}$ is said to be *upper semi-continuous* if $f^{-1}((-\infty, a))$ is open for all real a, and *lower semi-continuous* if $f^{-1}((a, +\infty))$ is open for all real a.

- Show that an indicator function 1_E is upper semi-continuous if and only if E is closed, and lower semi-continuous if and only if E is open.
- If X is normal, show that a function f is upper semi-continuous if and only if $f(x) = \inf\{g(x) : g \in C(X \to \mathbf{R}), g \geq f\}$ for all $x \in X$, and lower semi-continuous if and only if $f(x) = \sup\{g(x) : g \in C(X \to \mathbf{R}), g \leq f\}$ for all $x \in X$, where we write $f \leq g$ if $f(x) \leq g(x)$ for all $x \in X$.

1.10.2. The Riesz representation theorem. Let X be a locally compact Hausdorff space which is also σ-compact. In Definition 1.10.2 we defined the notion of a Radon measure. Such measures are quite common in real analysis. For instance, we have the following result.

Theorem 1.10.10. *Let μ be a non-negative finite Borel measure on a compact metric space X. Then μ is a Radon measure.*

Proof. As μ is finite, it is locally finite, so it suffices to show inner and outer regularity. Let $\mathcal{A}$ be the collection of all Borel subsets E of X such that

$$\sup\{\mu(K) : K \subset E, \text{ closed}\} = \inf\{\mu(U) : U \supset E, \text{ open}\} = \mu(E),$$

It will then suffice to show that every Borel set lies in $\mathcal{A}$ (note that as X is compact, a subset K of X is closed if and only if it is compact).

Clearly $\mathcal{A}$ contains the empty set and the whole set X, and is closed under complements. It is also closed under finite unions and intersections. Indeed, given two sets $E, F \in \mathcal{A}$, we can find a sequences $K_n \subset E \subset U_n$, $L_n \subset F \subset V_n$ of closed sets K_n, L_n and open sets

U_n, V_n such that $\mu(K_n), \mu(U_n) \to \mu(E)$ and $\mu(L_n), \mu(V_n) \to \mu(F)$. Since

$$\begin{aligned}\mu(K_n \cap L_n) + \mu(K_n \cup L_n) &= \mu(K_n) + \mu(L_n) \\ &\to \mu(E) + \mu(F) \\ &= \mu(E \cap F) + \mu(E \cup F)\end{aligned}$$

we have (by monotonicity of μ) that

$$\mu(K_n \cap L_n) \to \mu(E \cap F); \quad \mu(K_n \cup L_n) \to \mu(E \cup F)$$

and similarly

$$\mu(U_n \cap V_n) \to \mu(E \cap F); \quad \mu(U_n \cup V_n) \to \mu(E \cup F)$$

and so $E \cap F, E \cup F \in \mathcal{A}$.

One can also show that $\mathcal{A}$ is closed under countable disjoint unions and is thus a σ-algebra. Indeed, given disjoint sets $E_n \in \mathcal{A}$ and $\varepsilon > 0$, pick a closed $K_n \subset E_n$ and open $U_n \supset E_n$ such that $\mu(E_n \backslash K_n), \mu(U_n \backslash E_n) \leq \varepsilon/2^n$; then

$$\mu(\bigcup_{n=1}^{\infty} E_n) \leq \mu(\bigcup_{n=1}^{\infty} U_n) \leq \sum_{n=1}^{\infty} \mu(E_n) + \varepsilon$$

and

$$\mu(\bigcup_{n=1}^{\infty} E_n) \geq \mu(\bigcup_{n=1}^{\infty} K_n) \geq \sum_{n=1}^{N} \mu(E_n) - \varepsilon$$

for any N, and the claim follows from the squeeze test.

To finish the claim it suffices to show that every open set V lies in $\mathcal{A}$. For this it will suffice to show that V is a countable union of closed sets. But as X is a compact metric space, it is separable (Lemma 1.8.6), and so V has a countable dense subset $x_1, x_2, \ldots$. One then easily verifies that every point in the open set V is contained in a closed ball of rational radius centred at one of the x_i that is in turn contained in V; thus V is the countable union of closed sets as desired. □

This result can be extended to more general spaces than compact metric spaces, for instance to Polish spaces (provided that the measure remains finite). For instance:

Exercise 1.10.12. Let X be a locally compact metric space which is σ-compact, and let μ be an unsigned Borel measure which is finite on every compact set. Show that μ is a Radon measure.

When the assumptions of X are weakened, then it is possible to find locally finite Borel measures that are not Radon measures, but they are somewhat pathological in nature.

Exercise 1.10.13. Let X be a locally compact Hausdorff space which is σ-compact, and let μ be a Radon measure. Define a F_σ set to be a countable union of closed sets, and a G_δ set to be a countable intersection of open sets. Show that every Borel set can be expressed as the union of an F_σ set and a null set, and as a G_δ set with a null subset removed.

If μ is a Radon measure on X, then we can define the integral $I_\mu(f) := \int_X f \, d\mu$ for every $f \in C_c(X \to \mathbf{R})$, since μ assigns every compact set a finite measure. Furthermore, I_μ is a *linear functional* on $C_c(X \to \mathbf{R})$ which is *positive* in the sense that $I_\mu(f) \geq 0$ whenever f is non-negative. If we place the uniform norm on $C_c(X \to \mathbf{R})$, then I_μ is continuous if and only if μ is finite; but we will not use continuity for now, relying instead on positivity.

The fundamentally important *Riesz representation theorem* for such spaces asserts that this is the *only* way to generate such linear functionals:

Theorem 1.10.11 (Riesz representation theorem for $C_c(X \to \mathbf{R})$, unsigned version)**.** *Let X be a locally compact Hausdorff space which is also σ-compact. Let $I : C_c(X \to \mathbf{R}) \to \mathbf{R}$ be a positive linear functional. Then there exists a unique Radon measure μ on X such that $I = I_\mu$.*

Remark 1.10.12. The σ-compactness hypothesis can be dropped (after relaxing the inner regularity condition to only apply to open sets, rather than to all sets); but I will restrict attention here to the σ-compact case (which already covers a large fraction of the applications of this theorem) as the argument simplifies slightly.

Proof. We first prove the uniqueness, which is quite easy due to all the properties that Radon measures enjoy. Suppose we had two

Radon measures μ, μ' such that $I = I_\mu = I_{\mu'}$; in particular, we have

$$\int_X f \, d\mu = \int_X f \, d\mu' \tag{1.75}$$

for all $f \in C_c(X \to \mathbf{R})$. Now let K be a compact set, and let U be an open neighbourhood of K. By Exercise 1.10.6, we can find $f \in C_c(X \to \mathbf{R})$ with $1_K \leq f \leq 1_U$; applying this to (1.75), we conclude that

$$\mu(U) \geq \mu'(K).$$

Taking suprema in K and using inner regularity, we conclude that $\mu(U) \geq \mu'(U)$; exchanging μ and μ' we conclude that μ and μ' agree on open sets; by outer regularity we then conclude that μ and μ' agree on all Borel sets.

Now we prove existence, which is significantly trickier. We will initially make the simplifying assumption that X is compact (so in particular $C_c(X \to \mathbf{R}) = C(X \to \mathbf{R}) = BC(X \to \mathbf{R})$), and remove this assumption at the end of the proof.

Observe that I is monotone on $C(X \to \mathbf{R})$, thus $I(f) \leq I(g)$ whenever $f \leq g$.

We would like to define the measure μ on Borel sets E by defining $\mu(E) := I(1_E)$. This does not work directly, because 1_E is not continuous. To get around this problem we shall begin by extending the functional I to the class $BC_{lsc}(X \to \mathbf{R}^+)$ of bounded lower *semi-continuous* non-negative functions. We define $I(f)$ for such functions by the formula

$$I(f) := \sup\{I(g) : g \in C_c(X \to \mathbf{R}); 0 \leq g \leq f\}$$

(cf. Exercise 1.10.11). This definition agrees with the existing definition of $I(f)$ in the case when f is continuous. Since $I(1)$ is finite and I is monotone, one sees that $I(f)$ is finite (and non-negative) for all $f \in BC_{lsc}(X \to \mathbf{R}^+)$. One also easily sees that I is monotone on $BC_{lsc}(X \to \mathbf{R}^+)$: $I(f) \leq I(g)$ whenever $f, g \in BC_{lsc}(X \to \mathbf{R}^+)$ and $f \leq g$, and homogeneous in the sense that $I(cf) = cI(f)$ for all $f \in BC_{lsc}(X \to \mathbf{R}^+)$ and $c > 0$. It is also easy to verify the super-additivity property $I(f + f') \geq I(f) + I(f')$ for $f, f' \in BC_{lsc}(X \to$

$\mathbf{R}^+$); this simply reflects the linearity of I on $C_c(X \to \mathbf{R})$, together with the fact that if $0 \leq g \leq f$ and $0 \leq g' \leq f'$, then $0 \leq g + g' \leq f + f'$.

We now complement the super-additivity property with a countably sub-additive one: if $f_n \in BC_{lsc}(X \to \mathbf{R}^+)$ is a sequence, and $f \in BC_{lsc}(X \to \mathbf{R}^+)$ is such that $f(x) \leq \sum_{n=1}^\infty f_n(x)$ for all $x \in X$, then $I(f) \leq \sum_{n=1}^\infty I(f_n)$.

Pick a small $0 < \varepsilon < 1$. It will suffice to show that $I(g) \leq \sum_{n=1}^\infty I(f_n) + O(\varepsilon^{1/2})$ (say) whenever $g \in C_c(X \to \mathbf{R})$ is such that $0 \leq g \leq f$, and $O(\varepsilon^{1/2})$ denotes a quantity bounded in magnitude by $C\varepsilon^{1/2}$, where C is a quantity that is independent of ε.

Fix g. For every $x \in X$, we can find a neighbourhood U_x of x such that $|g(y) - g(x)| \leq \varepsilon$ for all $y \in U_x$; we can also find $N_x > 0$ such that $\sum_{n=1}^{N_x} f_n(x) \geq f(x) - \varepsilon$. By shrinking U_x if necessary, we see from the lower semicontinuity of the f_n and f that we can also ensure that $f_n(y) \geq f_n(x) - \varepsilon/2^n$ for all $1 \leq n \leq N_x$ and $y \in U_x$.

By normality, we can find open neighbourhoods V_x of x whose closure lies in U_x. The V_x form an open cover of X. Since we are assuming X to be compact, we can thus find a finite subcover $V_{x_1}, \ldots, V_{x_k}$ of X. Applying Lemma 1.10.9, we can thus find a partition of unity $1 = \sum_{j=1}^k \psi_j$, where each ψ_j is supported on U_{x_j}.

Let $x \in X$ be such that $g(x) \geq \sqrt{\varepsilon}$. Then we can write $g(x) = \sum_{j:x\in U_{x_j}} g(x)\psi_j(x)$. If j is in this sum, then $|g(x_j) - g(x)| \leq \varepsilon$, and thus (for ε small enough) $g(x_j) \geq \sqrt{\varepsilon}/2$, and hence $f(x_j) \geq \sqrt{\varepsilon}/2$. We can then write

$$1 \leq \sum_{n=1}^{N_{x_j}} \frac{f_n(x_j)}{f(x_j)} + O(\sqrt{\varepsilon})$$

and thus

$$g(x) \leq \sum_{n=1}^\infty \sum_{j:f(x_j)\geq\sqrt{\varepsilon}/2;N_{x_j}\geq n} \frac{f_n(x_j)}{f(x_j)} g(x_j)\psi_j(x) + O(\sqrt{\varepsilon})$$

(here we use the fact that $\sum_j \psi_j(x) = 1$ and that the continuous compactly supported function g is bounded). Observe that only finitely

many summands are non-zero. We conclude that

$$I(g) \leq \sum_{n=1}^{\infty} I(\sum_{j: f(x_j) \geq \sqrt{\varepsilon}/2; N_{x_j} \geq n} \frac{f_n(x_j)}{f(x_j)} g(x_j) \psi_j) + O(\sqrt{\varepsilon})$$

(here we use that $1 \in C_c(X)$ and so $I(1)$ is finite). On the other hand, for any $x \in X$ and any n, the expression

$$\sum_{j: f(x_j) \geq \sqrt{\varepsilon}/2; N_{x_j} \geq n} \frac{f_n(x_j)}{f(x_j)} g(x_j) \psi_j(x)$$

is bounded from above by

$$\sum_j f_n(x_j) \psi_j(x);$$

since $f_n(x) \geq f_n(x_j) - \varepsilon/2^n$ and $\sum_j \psi_j(x) = 1$, this is bounded above in turn by

$$\varepsilon/2^n + f_n(x).$$

We conclude that

$$I(g) \leq \sum_{n=1}^{\infty} [I(f_n) + O(\varepsilon/2^n)] + O(\sqrt{\varepsilon})$$

and the sub-additivity claim follows.

Combining sub-additivity and super-additivity we see that I is additive: $I(f+g) = I(f) + I(g)$ for $f, g \in BC_{lsc}(X \to \mathbf{R}^+)$.

Now that we are able to integrate lower semi-continuous functions, we can start defining the Radon measure μ. When U is open, we define $\mu(U)$ by

$$\mu(U) := I(1_U),$$

which is well-defined and non-negative since 1_U is bounded, non-negative and lower semi-continuous. When K is closed we define $\mu(K)$ by complementation:

$$\mu(K) := \mu(X) - \mu(X \backslash K);$$

this is compatible with the definition of μ on open sets by additivity of I, and is also non-negative. The monotonicity of I implies monotonicity of μ: in particular, if a closed set K lies in an open set U, then $\mu(K) \leq \mu(U)$.

Given any set $E \subset X$, define the *outer measure*

$$\mu^+(E) := \inf\{\mu(U) : E \subset U, \text{ open}\}$$

and the *inner measure*

$$\mu^-(E) := \sup\{\mu(K) : E \supset K, \text{ closed}\};$$

thus $0 \leq \mu^-(E) \leq \mu^+(E) \leq \mu(X)$. We call a set E *measurable* if $\mu^-(E) = \mu^+(E)$. By arguing as in the proof of Theorem 1.10.10, we see that the class of measurable sets is a *Boolean algebra*. Next, we claim that every open set U is measurable. Indeed, unwrapping all the definitions we see that

$$\mu(U) = \sup\{I(f) : f \in C_c(X \to \mathbf{R}); 0 \leq f \leq 1_U\}.$$

Each f in this supremum is supported in some closed subset K of U, and from this one easily verifies that $\mu^+(U) = \mu(U) = \mu^-(U)$. Similarly, every closed set K is measurable. We can now extend μ to measurable sets by declaring $\mu(E) := \mu^+(E) = \mu^-(E)$ when E is measurable; this is compatible with the previous definitions of μ.

Next, let $E_1, E_2, \ldots$ be a countable sequence of disjoint measurable sets. Then for any $\varepsilon > 0$, we can find open neighbourhoods U_n of E_n and closed sets K_n in E_n such that $\mu(E_n) \leq \mu(U_n) \leq \mu(E_n) + \varepsilon/2^n$ and $\mu(E_n) - \varepsilon/2^n \leq \mu(K_n) \leq \mu(E_n)$. Using the sub-additivity of I on $BC(X \to \mathbf{R}^+)$, we have $\mu(\bigcup_{n=1}^\infty U_n) \leq \sum_{n=1}^\infty \mu(U_n) \leq \sum_{n=1}^\infty \mu(E_n) + \varepsilon$. Similarly, from the additivity of I we have $\mu(\sum_{n=1}^N K_n) = \sum_{n=1}^N \mu(K_n) \geq \sum_{n=1}^N \mu(E_n) - \varepsilon$. Letting $\varepsilon \to 0$, we conclude that $\bigcup_{n=1}^\infty E_n$ is measurable with $\mu(\bigcup_{n=1}^\infty E_n) = \sum_{n=1}^\infty \mu(E_n)$. Thus the Boolean algebra of measurable sets is in fact a σ-algebra, and μ is a countably additive measure on it. From construction we also see that it is finite, outer regular, and inner regular, and therefore is a Radon measure. The only remaining thing to check is that $I(f) = I_\mu(f)$ for all $f \in C(X \to \mathbf{R})$. If f is a finite non-negative linear combination of indicator functions of open sets, the claim is clear from the construction of μ and the additivity of I on $BC(X \to \mathbf{R}^+)$; taking uniform limits, we obtain the claim for non-negative continuous functions, and then by linearity we obtain it for all functions.

This concludes the proof in the case when X is compact. Now suppose that X is σ-compact. Then we can find a partition of unity

$1 = \sum_{n=0}^{\infty} \psi_n$ into continuous compactly supported functions $\psi_n \in C_c(X \to \mathbf{R}^+)$, with each $x \in X$ being contained in the support of finitely many ψ_n. (Indeed, from σ-compactness and the locally compact Hausdorff property one can find a nested sequence $K_1 \subset K_2 \subset \ldots$ of compact sets, with each K_n in the interior of K_{n+1}, such that $\bigcup_n K_n = X$. Using Exercise 1.10.6, one can find functions $\eta_n \in C_c(X \to \mathbf{R}^+)$ that equal 1 on K_n and are supported on K_{n+1}; now take $\psi_n := \eta_{n+1} - \eta_n$ and $\psi_0 := \eta_0$.) Observe that $I(f) = \sum_n I(\psi_n f)$ for all $f \in C_c(X \to \mathbf{R})$. From the compact case we see that there exists a finite Radon measure μ_n such that $I(\psi_n f) = I_{\mu_n}(f)$ for all $f \in C_c(X \to \mathbf{R})$; setting $\mu := \sum_n \mu_n$ one can verify (using the monotone convergence theorem, Theorem 1.1.21) that μ obeys the required properties. □

Remark 1.10.13. One can also construct the Radon measure μ using the Carátheodory extension theorem (Theorem 1.1.17); this proof of the Riesz representation theorem can be found in many real analysis texts. A third method is to first create the space L^1 by taking the completion of $C_c(X \to \mathbf{R})$ with respect to the L^1 norm $\|f\|_{L^1} := I(|f|)$, and then define $\mu(E) := \|1_E\|_{L^1}$. It seems to me that all three proofs are about equally lengthy, and ultimately rely on the same ingredients; they all seem to have their strengths and weaknesses, and involve at least one tricky computation somewhere (in the above argument, the most tricky thing is the countable subadditivity of I on lower semicontinuous functions). I have yet to find a proof of this theorem which is both clean and conceptual, and would be happy to learn of other proofs of this theorem.

Remark 1.10.14. One can use the Riesz representation theorem to provide an alternate construction of Lebesgue measure, say on $\mathbf{R}$. Indeed, the *Riemann integral* already provides a positive linear functional on $C_0(\mathbf{R} \to \mathbf{R})$, which by the Riesz representation theorem must come from a Radon measure, which can be easily verified to assign the value $b - a$ to every interval $[a, b]$ and thus must agree with Lebesgue measure. The same approach lets one define volume measures on manifolds with a volume form.

Exercise 1.10.14. Let X be a locally compact Hausdorff space which is σ-compact, and let μ be a Radon measure. For any non-negative

Borel measurable function f, show that

$$\int_X f\ d\mu = \inf\{\int_X g\ d\mu : g \geq f; g \text{ lower semi-continuous}\}$$

and

$$\int_X f\ d\mu = \sup\{\int_X g\ d\mu : 0 \leq g \leq f; g \text{ upper semi-continuous}\}.$$

Similarly, for any non-negative lower semi-continuous function g, show that

$$\int_X g\ d\mu = \sup\{\int_X h\ d\mu : 0 \leq h \leq g; h \in C_c(X \to \mathbf{R})\}.$$

Now we consider signed functionals on $C_c(X \to \mathbf{R})$, which we now turn into a normed vector space using the uniform norm. The key lemma here is the following variant of the Jordan decomposition theorem (Exercise 1.2.5).

Lemma 1.10.15 (Jordan decomposition for functions)**.** *Let $I \in C_c(X \to \mathbf{R})^*$ be a (real) continuous linear functional. Then there exist positive linear functions $I^+, I^- \in C_c(X \to \mathbf{R})^*$ such that $I = I^+ - I^-$.*

Proof. For $f \in C_c(X \to \mathbf{R}^+)$, we define

$$I^+(f) := \sup\{I(g) : g \in C_c(X \to \mathbf{R}) : 0 \leq g \leq f\}.$$

Clearly $0 \leq I^+(f) \leq I(f)$ for $f \in C_c(X \to \mathbf{R}^+)$; one also easily verifies the homogeneity property $I^+(cf) = cI^+(f)$ and super-additivity property $I^+(f_1 + f_2) \geq I^+(f_1) + I^+(f_2)$ for $c > 0$ and $f, f_1, f_2 \in C_c(X \to \mathbf{R}^+)$. On the other hand, if $g, f_1, f_2 \in C_c(X \to \mathbf{R}^+)$ are such that $g \leq f_1 + f_2$, then we can decompose $g = g_1 + g_2$ for some $g_1, g_2 \in C_c(X \to \mathbf{R}^+)$ with $g_1 \leq f_1$ and $g_2 \leq f_2$; for instance we can take $g_1 := \min(g, f_1)$ and $g_2 := g - g_1$. From this we can complement super-additivity with sub-additivity and conclude that $I^+(f_1 + f_2) = I^+(f_1) + I^+(f_2)$.

Every function in $C_c(X \to \mathbf{R})$ can be expressed as the difference of two functions in $C_c(X \to \mathbf{R}^+)$. From the additivity and homogeneity of I^+ on $C_c(X \to \mathbf{R}^+)$ we may thus extend I^+ uniquely to be a linear functional on $C_c(X \to \mathbf{R})$. Since I is bounded on $C_c(X \to \mathbf{R})$, we see that I^+ is also. If we then define $I^- := I^+ - I$, one quickly verifies all the required properties. □

Exercise 1.10.15. Show that the functionals I^+, I^- appearing in the above lemma are unique.

Define a *signed Radon measure* on a σ-compact, locally compact Hausdorff space X to be a signed Borel measure μ whose positive and negative variations are both Radon. It is easy to see that a signed Radon measure μ generates a linear functional I_μ on $C_c(X \to \mathbf{R})$ as before, and I_μ is continuous if μ is finite. We have a converse:

Exercise 1.10.16 (Riesz representation theorem, signed version)**.** Let X be a locally compact Hausdorff space which is also σ-compact, and let $I \in C_c(X \to \mathbf{R})^*$ be a continuous linear functional. Then there exists a unique signed finite Radon measure μ such that $I = I_\mu$. (*Hint*: combine Theorem 1.10.11 with Lemma 1.10.15.)

The space of signed finite Radon measures on X is denoted $M(X \to \mathbf{R})$, or $M(X)$ for short.

Exercise 1.10.17. Show that the space $M(X)$, with the total variation norm $\|\mu\|_{M(X)} := |\mu|(X)$, is a real Banach space, which is isomorphic to the dual of both $C_c(X \to \mathbf{R})$ and its completion $C_0(X \to \mathbf{R})$, thus

$$C_c(X \to \mathbf{R})^* \equiv C_0(X \to \mathbf{R})^* \equiv M(X).$$

Remark 1.10.16. Note that the previous exercise generalises the identifications $c_c(\mathbf{N})^* \equiv c_0(\mathbf{N})^* \equiv \ell^1(\mathbf{N})$ from previous notes. For compact Hausdorff spaces X, we have $C(X \to \mathbf{R}) = C_0(X \to \mathbf{R})$, and thus $C(X \to \mathbf{R})^* \equiv M(X)$. For locally compact Hausdorff spaces that are σ-compact but not compact, we instead have $C(X \to \mathbf{R})^* \equiv M(\beta X)$, where βX is the *Stone-Čech compactification* of X, which we will discuss in Section 2.5.

Remark 1.10.17. One can of course also define complex Radon measures to be those *complex* finite Borel measures whose real and imaginary parts are signed Radon measures, and define $M(X \to \mathbf{C})$ to be the space of all such measures; then one has analogues of the above identifications. We omit the details.

Exercise 1.10.18. Let X, Y be two locally compact Hausdorff spaces that are also σ-compact, and let $f : X \to Y$ be a continuous map.

If μ is an unsigned Radon measure on X, show that the *pushforward measure* $f_{\#}\mu$ on Y, defined by $f_{\#}\mu(E) := \mu(f^{-1}(E))$, is a Radon measure on Y. Establish the same fact for signed Radon measures.

Let X be locally compact Hausdorff and σ-compact. As $M(X)$ is equivalent to the dual of the Banach space $C_0(X \to \mathbf{R})$, it acquires a *weak* topology* (see Section 1.9), known as the *vague topology*. A sequence of Radon measures $\mu_n \in M(X)$ then converges vaguely to a limit $\mu \in M(X)$ if and only if $\int_X f\ d\mu_n \to \int_X f\ d\mu$ for all $f \in C_0(X \to \mathbf{R})$.

Exercise 1.10.19. Let m be Lebesgue measure on the real line (with the usual topology).

- Show that the measures $nm \downharpoonright_{[0,1/n]}$ converge vaguely as $n \to \infty$ to the *Dirac mass* δ_0 at the origin 0.
- Show that the measures $\frac{1}{n}\sum_{i=1}^n \delta_{i/n}$ converge vaguely as $n \to \infty$ to the measure $m \downharpoonright_{[0,1]}$. (*Hint*: Continuous, compactly supported functions are *Riemann integrable*.)
- Show that the measures δ_n converge vaguely as $n \to \infty$ to the zero measure 0.

Exercise 1.10.20. Let X be locally compact Hausdorff and σ-compact. Show that for every unsigned Radon measure μ, the map $\iota : L^1(\mu) \to M(X)$ defined by sending $f \in L^1(\mu)$ to the measure μ_f is an *isometry*, thus $L^1(\mu)$ can be identified with a subspace of $M(X)$. Show that this subspace is closed in the norm topology, but give an example to show that it need not be closed in the vague topology. Show that $M(X) = \bigcup_\mu L^1(\mu)$, where μ ranges over all unsigned Radon measures on X; thus one can think of $M(X)$ as many L^1's "glued together".

Exercise 1.10.21. Let X be a locally compact Hausdorff space which is σ-compact. Let $f_n \in C_0(X \to \mathbf{R})$ be a sequence of functions, and let $f \in C_0(X \to \mathbf{R})$ be another function. Show that f_n converges weakly to f in $C_0(X \to \mathbf{R})$ if and only if the f_n are uniformly bounded and converge pointwise to f.

Exercise 1.10.22. Let X be a locally compact metric space which is σ-compact.

- Show that the space of finitely supported measures in $M(X)$ is a dense subset of $M(X)$ in the vague topology.
- Show that a Radon probability measure in $M(X)$ can be expressed as the vague limit of a sequence of discrete (i.e. finitely supported) probability measures.

1.10.3. The Stone-Weierstrass theorem. We have already seen how rough functions (e.g. L^p functions) can be approximated by continuous functions. Now we study in turn how continuous functions can be approximated by even more special functions, such as polynomials. The natural topology to work with here is the uniform topology (since uniform limits of continuous functions are continuous).

For non-compact spaces, such as $\mathbf{R}$, it is usually not possible to approximate continuous functions uniformly by a smaller class of functions. For instance, the function $\sin(x)$ cannot be approximated uniformly by polynomials on $\mathbf{R}$, since $\sin(x)$ is bounded, the only bounded polynomials are the constants, and constants cannot converge to anything other than another constant. On the other hand, on a compact domain such as $[-1,1]$, one can easily approximate $\sin(x)$ uniformly by polynomials, for instance by using *Taylor series*. So we will focus instead on compact Hausdorff spaces X such as $[-1,1]$, in which continuous functions are automatically bounded.

The space $\mathcal{P}([-1,1])$ of (real-valued) polynomials is a subspace of the Banach space $C([-1,1])$. But it is also closed under pointwise multiplication $f,g \mapsto fg$, making $\mathcal{P}([-1,1])$ an *algebra*, and not merely a vector space. We can then rephrase the classical *Weierstrass approximation theorem* as the assertion that $\mathcal{P}([-1,1])$ is dense in $C([-1,1])$.

One can then ask the more general question of when a sub-algebra $\mathcal{A}$ of $C(X)$ - i.e. a subspace closed under pointwise multiplication - is dense. Not every sub-algebra is dense: the algebra of constants, for instance, will not be dense in $C(X)$ when X has at least two points. Another example in a similar spirit: given two distinct points x_1, x_2 in X, the space $\{f \in C(X) : f(x_1) = f(x_2)\}$ is a sub-algebra of $C(X)$, but it is not dense, because it is already closed, and cannot *separate*

x_1 and x_2 in the sense that it cannot produce a function that assigns different values to x_1 and x_2.

The remarkable *Stone-Weierstrass theorem* shows that this inability to separate points is the *only* obstruction to density, at least for algebras with the identity.

Theorem 1.10.18 (Stone-Weierstrass theorem, real version)**.** *Let X be a compact Hausdorff space, and let $\mathcal{A}$ be a sub-algebra of $C(X \to \mathbf{R})$ which contains the constant function 1 and separates points (i.e. for every distinct $x_1, x_2 \in X$, there exists at least one f in $\mathcal{A}$ such that $f(x_1) \neq f(x_2)$. Then $\mathcal{A}$ is dense in $C(X \to \mathbf{R})$.*

Remark 1.10.19. Observe that this theorem contains the Weierstrass approximation theorem as a special case, since the algebra of polynomials clearly separates points. Indeed, we will use (a very special case) of the Weierstrass approximation theorem in the proof.

Proof. It suffices to verify the claim for algebras $\mathcal{A}$ which are closed in the $C(X \to \mathbf{R})$ topology, since the claim follows in the general case by replacing $\mathcal{A}$ with its closure (note that the closure of an algebra is still an algebra).

Observe from the *Weierstrass approximation theorem* that on any bounded interval $[-K, K]$, the function $|x|$ can be expressed as the uniform limit of polynomials $P_n(x)$; one can even write down explicit formulae for such a P_n, though we will not need such formulae here. Since continuous functions on the compact space X are bounded, this implies that for any $f \in \mathcal{A}$, the function $|f|$ is the uniform limit of polynomial combinations $P_n(f)$ of f. As $\mathcal{A}$ is an algebra, the $P_n(f)$ lie in $\mathcal{A}$; as $\mathcal{A}$ is closed; we see that $|f|$ lies in $\mathcal{A}$.

Using the identities $\max(f,g) = \frac{f+g}{2} + |\frac{f-g}{2}|$, $\min(f,g) = \frac{f+g}{2} - |\frac{f-g}{2}|$, we conclude that $\mathcal{A}$ is a *lattice* in the sense that one has $\max(f,g), \min(f,g) \in \mathcal{A}$ whenever $f, g \in \mathcal{A}$.

Now let $f \in C(X \to \mathbf{R})$ and $\varepsilon > 0$. We would like to find $g \in \mathcal{A}$ such that $|f(x) - g(x)| \leq \varepsilon$ for all $x \in X$.

Given any two points $x, y \in X$, we can at least find a function $g_{xy} \in \mathcal{A}$ such that $g_{xy}(x) = f(x)$ and $g_{xy}(y) = f(y)$; this follows since the vector space $\mathcal{A}$ separates points and also contains the identity

function (the case $x = y$ needs to be treated separately). We now use these functions g_{xy} to build the approximant g. First, observe from continuity that for every $x, y \in X$ there exists an open neighbourhood V_{xy} of y such that $g_{xy}(y') \geq f(y') - \varepsilon$ for all $y' \in V_{xy}$. By compactness, for any fixed x we can cover X by a finite number of these V_{xy}. Taking the max of all the g_{xy} associated to this finite subcover, we create another function $g_x \in \mathcal{A}$ such that $g_x(x) = f(x)$ and $g_x(y) \geq f(y) - \varepsilon$ for all $y \in X$. By continuity, we can find an open neighbourhood U_x of x such that $g_x(x') \leq f(x') + \varepsilon$ for all $x' \in U_x$. Again applying compactness, we can cover X by a finite number of the U_x; taking the min of all the g_x associated to this finite subcover we obtain $g \in \mathcal{A}$ with $f(x) - \varepsilon \leq g(x) \leq f(x) + \varepsilon$ for all $x \in X$, and the claim follows. □

There is an analogue of the Stone-Weierstrass theorem for algebras that do not contain the identity:

Exercise 1.10.23. Let X be a compact Hausdorff space, and let $\mathcal{A}$ be a closed sub-algebra of $C(X \to \mathbf{R})$ which separates points but does not contain the identity. Show that there exists a unique $x_0 \in X$ such that $\mathcal{A} = \{f \in C(X \to \mathbf{R}) : f(x_0) = 0\}$.

The Stone-Weierstrass theorem is not true as stated in the complex case. For instance, the space $C(\mathbb{D} \to \mathbf{C})$ of complex-valued functions on the closed unit disk $\mathbb{D} := \{z \in \mathbf{C} : |z| \leq 1\}$ has a closed proper sub-algebra that separates points, namely the algebra $\mathcal{H}(\mathbb{D})$ of functions in $C(\mathbb{D} \to \mathbf{C})$ that are *holomorphic* on the interior of this disk. Indeed, by *Cauchy's theorem* and its converse (*Morera's theorem*), a function $f \in C(\mathbb{D} \to \mathbf{C})$ lies in $\mathcal{H}(\mathbb{D})$ if and only if $\int_\gamma f = 0$ for every closed contour γ in $\mathbb{D}$, and one easily verifies that this implies that $\mathcal{H}(\mathbb{D})$ is closed; meanwhile, the holomorphic function $z \mapsto z$ separates all points. However, the Stone-Weierstrass theorem can be recovered in the complex case by adding one further axiom, namely that the algebra be closed under conjugation:

Exercise 1.10.24 (Stone-Weierstrass theorem, complex version)**.** Let X be a compact Hausdorff space, and let $\mathcal{A}$ be a complex sub-algebra of $C(X \to \mathbf{C})$ which contains the constant function 1, separates

points, and is closed under the conjugation operation $f \mapsto \overline{f}$. Then $\mathcal{A}$ is dense in $C(X \to \mathbf{C})$.

Exercise 1.10.25. Let $\mathcal{T} \subset C([0,1] \to \mathbf{C})$ be the space of trigonometric polynomials $x \mapsto \sum_{n=-N}^{N} c_n e^{2\pi i n x}$, where $N \geq 0$ and the c_n are complex numbers. Show that $\mathcal{T}$ is dense in $C([0,1] \to \mathbf{C})$ (with the uniform topology), and that $\mathcal{T}$ is dense in $L^p([0,1] \to \mathbf{C})$ (with the L^p topology) for all $0 < p < \infty$.

Exercise 1.10.26. Let X be a locally compact Hausdorff space that is σ-compact, and let $\mathcal{A}$ be a sub-algebra of $C(X \to \mathbf{R})$ which separates points and contains the identity function. Show that for every function $f \in C(X \to \mathbf{R})$ there exists a sequence $f_n \in \mathcal{A}$ which converges to f uniformly on compact subsets of X.

Exercise 1.10.27. Let X, Y be compact Hausdorff spaces. Show that every function $f \in C(X \times Y \to \mathbf{R})$ can be expressed as the uniform limit of functions of the form $(x,y) \mapsto \sum_{j=1}^{k} f_j(x) g_j(y)$, where $f_j \in C(X \to \mathbf{R})$ and $g_j \in C(Y \to \mathbf{R})$.

Exercise 1.10.28. Let $(X_\alpha)_{\alpha \in A}$ be a family of compact Hausdorff spaces, and let $X := \prod_{\alpha \in A} X_\alpha$ be the product space (with the product topology). Let $f \in C(X \to \mathbf{R})$. Show that f can be expressed as the uniform limit of continuous functions f_n, each of which only depend on finitely many of the coordinates in A, thus there exists a finite subset A_n of A and a continuous function $g_n \in C(\prod_{\alpha \in A_n} X_\alpha \to \mathbf{R})$ such that $f_n((x_\alpha)_{\alpha \in A}) = g_n((x_\alpha)_{\alpha \in A_n})$ for all $(x_\alpha)_{\alpha \in A} \in X$.

One useful application of the Stone-Weierstrass theorem is to demonstrate separability of spaces such as $C(X)$.

Proposition 1.10.20. *Let X be a compact metric space. Then $C(X \to \mathbf{C})$ and $C(X \to \mathbf{R})$ are separable.*

Proof. It suffices to show that $C(X \to \mathbf{R})$ is separable. By Lemma 1.8.6, X has a countable dense subset $x_1, x_2, \ldots$. By Urysohn's lemma, for each $n, m \geq 1$ we can find a function $\psi_{n,m} \in C(X \to \mathbf{R})$ which equals 1 on $B(x_n, 1/m)$ and is supported on $B(x_n, 2/m)$. The $\psi_{n,m}$ can then easily be verified to separate points, and so by the Stone-Weierstrass theorem, the algebra of polynomial combinations of the

$\psi_{n,m}$ in $C(X \to \mathbf{R})$ are dense; this implies that the algebra of *rational* polynomial combinations of the $\psi_{n,m}$ are dense, and the claim follows. □

Combining this with the Riesz representation theorem and the sequential Banach-Alaoglu theorem (Theorem 1.9.14), we obtain

Corollary 1.10.21. *If X is a compact metric space, then $M(X)$ is sequentially compact.*

Combining this with Theorem 1.10.10, we conclude a special case of *Prokhorov's theorem*:

Corollary 1.10.22 (Prokhorov's theorem, compact case)**.** *Let X be a compact metric space, and let μ_n be a sequence of Borel (hence Radon) probability measures on X. Then there exists a subsequence of μ_n which converge vaguely to another Borel probability measure μ.*

Exercise 1.10.29 (Prokhorov's theorem, non-compact case)**.** Let X be a locally compact metric space which is σ-compact, and let μ_n be a sequence of Borel probability measures. We assume that the sequence μ_n is *tight*, which means that for every $\varepsilon > 0$ there exists a compact set K such that $\mu_n(X \backslash K) \leq \varepsilon$ for all n. Show that there is a subsequence of μ_n which converges vaguely to another Borel probability measure μ. If tightness is not assumed, show that there is a subsequence which converges vaguely to a non-negative Borel measure μ, but give an example to show that this measure need not be a probability measure.

This theorem can be used to establish *Helly's selection theorem*:

Exercise 1.10.30 (Helly's selection theorem)**.** Let $f_n : \mathbf{R} \to \mathbf{R}$ be a sequence of functions whose *total variation* is uniformly bounded in n, and which is bounded at one point $x_0 \in \mathbf{R}$ (i.e. $\{f_n(x_0) : n = 1, 2, \ldots\}$ is bounded). Show that there exists a subsequence of f_n which converges uniformly on compact subsets of $\mathbf{R}$. (*Hint*: one can deduce this from Prokhorov's theorem using the *fundamental theorem of calculus* for functions of bounded variation.)

1.10.4. The commutative Gelfand-Naimark theorem (optional). One particularly beautiful application of the machinery developed in

the last few notes is the commutative *Gelfand-Naimark theorem*, that classifies commutative C^**-algebras*, and is of importance in spectral theory, operator algebras, and quantum mechanics.

Definition 1.10.23. A *complex Banach algebra* is a complex Banach space A which is also a complex algebra, such that $\|xy\| \leq \|x\|\|y\|$ for all $x, y \in A$. An algebra is *unital* if it contains a multiplicative identity 1, and *commutative* if $xy = yx$ for all $x, y \in A$. A C^**-algebra* is a complex Banach algebra with an anti-linear map $x \mapsto x^*$ from A to A which is an isometry (thus $\|x^*\| = \|x\|$ for all $x \in A$), an involution (thus $(x^*)^* = x$ for all $x \in A$), and obeys the C^* *identity* $\|x^*x\| = \|x\|^2$ for all $x \in A$.

A *homomorphism* $\phi : A \to B$ between two C^*-algebras is a continuous algebra homomorphism such that $\phi(x^*) = \phi(x)^*$ for all $x \in X$. An *isomorphism* is an homomorphism whose inverse exists and is also a homomorphism; two C^*-algebras are *isomorphic* if there exists an isomorphism between them.

Exercise 1.10.31. If H is a Hilbert space, and $B(H \to H)$ is the algebra of bounded linear operators on this space, with the adjoint map $T \mapsto T^*$ and the operator norm, show that $B(H \to H)$ is a unital C^*-algebra (not necessarily commutative). Indeed, one can think of C^*-algebras as an abstraction of a space of bounded linear operators on a Hilbert space (this is basically the content of the non-commutative *Gelfand-Naimark theorem*, which we will not discuss here).

Exercise 1.10.32. If X is a compact Hausdorff space, show that $C(X \to \mathbf{C})$ is a unital commutative C^*-algebra, with involution $f^* := \overline{f}$.

The remarkable (unital commutative) Gelfand-Naimark theorem asserts the converse statement to Exercise 1.10.32:

Theorem 1.10.24 (Unital commutative Gelfand-Naimark theorem)**.** *Every unital commutative C^*-algebra A is isomorphic to $C(X \to \mathbf{C})$ for some compact Hausdorff space X.*

There are analogues of this theorem for non-unital or non-commutative C^*-algebras, but for simplicity we shall restrict attention to the unital commutative case. We first need some spectral theory.

Exercise 1.10.33. Let A be a unital Banach algebra. Show that if $x \in A$ is such that $\|x - 1\| < 1$, then x is invertible. (*Hint*: use *Neumann series*.) Conclude that the space $A^\times \subset A$ of invertible elements of A is open.

Define the spectrum $\sigma(x)$ of an element $x \in A$ to be the set of all $z \in \mathbf{C}$ such that $x - z1$ is not invertible.

Exercise 1.10.34. If A is a unital Banach algebra and $x \in A$, show that $\sigma(x)$ is a compact subset of $\mathbf{C}$ that is contained inside the disk $\{z \in \mathbf{C} : |z| \leq \|x\|\}$.

Exercise 1.10.35 (Beurling-Gelfand spectral radius formula)**.** If A is a unital Banach algebra and $x \in A$, show that $\sigma(x)$ is non-empty with $\sup\{|z| : z \in \sigma(x)\} = \lim_{n\to\infty} \|x^n\|^{1/n}$. (*Hint*: To get the upper bound, observe that if $x^n - z^n 1$ is invertible for some $n \geq 1$, then so is $x - zI$, then use Exercise 1.10.34. To get the lower bound, first observe that for any $\lambda \in A^*$, the function $f_\lambda : z \mapsto \lambda((x - zI)^{-1})$ is holomorphic on the complement of $\sigma(x)$, which is already enough (with *Liouville's theorem*) to show that σ is non-empty. Let $r > \sup\{|z| : z \in \sigma(x)\}$ be arbitrary, then use Laurent series to show that $\lambda(x^n) \leq C_{\lambda,r} r^n$ for all n and some $C_{\lambda,r}$ independent of n. Then divide by r^n and use the uniform boundedness principle to conclude.)

Exercise 1.10.36 (C^*-algebra spectral radius formula)**.** Let A be a unital C^*-algebra. Show that

$$\|x\| = \|(x^*x)^{2^n}\|^{1/2^{n+1}} = \|(xx^*)^{2^n}\|^{1/2^{n+1}}$$

for all $n \geq 1$ and $x \in A$. Conclude that any homomorphism between C^*-algebras has operator norm at most 1. Also conclude that

$$\sup\{|z| : z \in \sigma(x)\} = \|x\|.$$

The next important concept is that of a *character*.

Definition 1.10.25. Let A be a unital commutative C^*-algebra. A *character* of A is be an element $\lambda \in A^*$ in the dual Banach space such

that $\lambda(xy) = \lambda(x)\lambda(y)$, $\lambda(1) = 1$, and $\lambda(x^*) = \overline{\lambda(x)}$ for all $x, y \in A$; equivalently, a character is a homomorphism from A to $\mathbf{C}$ (viewed as a (unital) C^* algebra). We let $\hat{A} \subset A^*$ be the space of all characters; this space is known as the *spectrum* of A.

Exercise 1.10.37. If A is a unital commutative C^*-algebra, show that $\hat{A}$ is a compact Hausdorff subset of A^* in the weak-* topology. (*Hint*: first use the spectral radius formula to show that all characters have operator norm 1, then use the Banach-Alaoglu theorem.)

Exercise 1.10.38. Define an *ideal* of a unital commutative C^*-algebra A to be a proper subspace I of A such that $xy, yx \in I$ for all $x \in I$ and $y \in A$. Show that if $\lambda \in \hat{A}$, then the *kernel* $\lambda^{-1}(\{0\})$ is a maximal ideal in A; conversely, if I is a maximal ideal in A, show that I is closed, and there is exactly one $\lambda \in \hat{A}$ such that $I = \lambda^{-1}(\{0\})$. Thus the spectrum of A can be canonically identified with the space of maximal ideals in A.

Exercise 1.10.39. Let X be a compact Hausdorff space, and let A be the C^*-algebra $A := C(X \to \mathbf{C})$. Show that for each $x \in X$, the operation $\lambda_x : f \mapsto f(x)$ is a character of A. Show that the map $\lambda : x \mapsto \lambda_x$ is a homeomorphism from X to $\hat{A}$; thus the spectrum of $C(X \to \mathbf{C})$ can be canonically identified with X. (*Hint*: use Exercise 1.10.23 to show the surjectivity of λ, Urysohn's lemma to show injectivity, and Corollary 1.8.2 to show the homeomorphism property.)

Inspired by the above exercise, we define the *Gelfand representation* $\hat{}: A \mapsto C(\hat{A} \to \mathbf{C})$, by the formula $\hat{x}(\lambda) := \lambda(x)$.

Exercise 1.10.40. Show that if A is a unital commutative C^*-algebra, then the Gelfand representation is a homomorphism of C^*-algebras.

Exercise 1.10.41. Let x be a non-invertible element of a unital commutative C^*-algebra A. Show that $\hat{x}$ vanishes at some $\lambda \in \hat{A}$. (*Hint*: the set $\{xy : y \in A\}$ is a proper ideal of A, and thus by Zorn's lemma (Section 2.4) is contained in a maximal ideal.)

Exercise 1.10.42. Show that if A is a unital commutative C^*-algebra, then the Gelfand representation is an isometry. (*Hint*: use Exercise 1.10.36 and Exercise 1.10.41.)

Exercise 1.10.43. Use the complex Stone-Weierstrass theorem and Exercises 1.10.40, 1.10.42 to conclude the proof of Theorem 1.10.24.

Notes. This lecture first appeared at `terrytao.wordpress.com/2009/03/02`. Thanks to Anush Tserunyan, Haokun Xu, Max Baroi, mmailliw/william, PDEbeginner, and anonymous commenters for corrections.

Eric noted another example of a locally compact Hausdorff space which was not normal, namely $(\omega + 1) \times (\omega_1 + 1) \backslash (\omega, \omega_1)$, where ω is the first infinite ordinal, and ω_1 is the first uncountable ordinal (endowed with the order topology, of course).

1.11. Interpolation of L^p spaces

In the previous sections, we have been focusing largely on the "soft" side of real analysis, which is primarily concerned with "qualitative" properties such as convergence, compactness, measurability, and so forth. In contrast, we will now emphasise the "hard" side of real analysis, in which we study estimates and upper and lower bounds of various quantities, such as norms of functions or operators. (Of course, the two sides of analysis are closely connected to each other; an understanding of both sides and their interrelationships, are needed in order to get the broadest and most complete perspective for this subject; see Section 1.3 of *Structure and Randomness* for more discussion.)

One basic tool in hard analysis is that of *interpolation*, which allows one to start with a hypothesis of two (or more) "upper bound" estimates, e.g. $A_0 \leq B_0$ and $A_1 \leq B_1$, and conclude a family of intermediate estimates $A_\theta \leq B_\theta$ (or maybe $A_\theta \leq C_\theta B_\theta$, where C_θ is a constant) for any choice of parameter $0 < \theta < 1$. Of course, interpolation is not a magic wand; one needs various hypotheses (e.g. linearity, sublinearity, convexity, or complexifiability) on A_i, B_i in order for interpolation methods to be applicable. Nevertheless, these techniques are available for many important classes of problems, most notably that of establishing boundedness estimates such as $\|Tf\|_{L^q(Y,\nu)} \leq C\|f\|_{L^p(X,\mu)}$ for linear (or "linear-like") operators T from one Lebesgue space $L^p(X, \mu)$ to another $L^q(Y, \nu)$. (Interpolation can also be performed for many other normed vector spaces

than the Lebesgue spaces, but we will just focus on Lebesgue spaces in these notes to focus the discussion.) Using interpolation, it is possible to reduce the task of proving such estimates to that of proving various "endpoint" versions of these estimates. In some cases, each endpoint only faces a portion of the difficulty that the interpolated estimate did, and so by using interpolation one has split the task of proving the original estimate into two or more simpler subtasks. In other cases, one of the endpoint estimates is very easy, and the other one is significantly more difficult than the original estimate; thus interpolation does not really simplify the task of proving estimates in this case, but at least clarifies the relative difficulty between various estimates in a given family.

As is the case with many other tools in analysis, interpolation is not captured by a single "interpolation theorem"; instead, there are a family of such theorems, which can be broadly divided into two major categories, reflecting the two basic methods that underlie the principle of interpolation. The *real interpolation method* is based on a divide and conquer strategy: to understand how to obtain control on some expression such as $\|Tf\|_{L^q(Y,\nu)}$ for some operator T and some function f, one would divide f into two or more components, e.g. into components where f is large and where f is small, or where f is oscillating with high frequency or only varying with low frequency. Each component would be estimated using a carefully chosen combination of the extreme estimates available; optimising over these choices and summing up (using whatever linearity-type properties on T are available), one would hope to get a good estimate on the original expression. The strengths of the real interpolation method are that the linearity hypotheses on T can be relaxed to weaker hypotheses, such as sublinearity or quasilinearity; also, the endpoint estimates are allowed to be of a weaker "type" than the interpolated estimates. On the other hand, the real interpolation often concedes a multiplicative constant in the final estimates obtained, and one is usually obligated to keep the operator T fixed throughout the interpolation process. The proofs of real interpolation theorems are also a little bit messy, though in many cases one can simply invoke a standard instance of such theorems (e.g. the Marcinkiewicz interpolation theorem) as a black box in applications.

The *complex interpolation method* instead proceeds by exploiting the powerful tools of complex analysis, in particular the *maximum modulus principle* and its relatives (such as the *Phragmén-Lindelöf principle*). The idea is to rewrite the estimate to be proven (e.g. $\|Tf\|_{L^q(Y,\nu)} \leq C\|f\|_{L^p(X,\mu)}$) in such a way that it can be embedded into a family of such estimates which depend holomorphically on a complex parameter s in some domain (e.g. the strip $\{\sigma+it : t \in \mathbf{R}, \sigma \in [0,1]\}$. One then exploits things like the maximum modulus principle to bound an estimate corresponding to an interior point of this domain by the estimates on the boundary of this domain. The strengths of the complex interpolation method are that it typically gives cleaner constants than the real interpolation method, and also allows the underlying operator T to vary holomorphically with respect to the parameter s, which can significantly increase the flexibility of the interpolation technique. The proofs of these methods are also very short (if one takes the maximum modulus principle and its relatives as a black box), which make the method particularly amenable for generalisation to more intricate settings (e.g. multilinear operators, mixed Lebesgue norms, etc.). On the other hand, the somewhat rigid requirement of holomorphicity makes it much more difficult to apply this method to non-linear operators, such as sublinear or quasilinear operators; also, the interpolated estimate tends to be of the same "type" as the extreme ones, so that one does not enjoy the upgrading of weak type estimates to strong type estimates that the real interpolation method typically produces. Also, the complex method runs into some minor technical problems when target space $L^q(Y,\nu)$ ceases to be a Banach space (i.e. when $q < 1$) as this makes it more difficult to exploit duality.

Despite these differences, the real and complex methods tend to give broadly similar results in practice, especially if one is willing to ignore constant losses in the estimates or epsilon losses in the exponents.

The theory of both real and complex interpolation can be studied abstractly, in general normed or quasi-normed spaces; see e.g. [**BeLo1976**] for a detailed treatment. However in these notes we shall focus exclusively on interpolation for Lebesgue spaces L^p (and

their cousins, such as the weak Lebesgue spaces $L^{p,\infty}$ and the Lorentz spaces $L^{p,r}$).

1.11.1. Interpolation of scalars. As discussed in the introduction, most of the interesting applications of interpolation occur when the technique is applied to operators T. However, in order to gain some intuition as to why interpolation works in the first place, let us first consider the significantly simpler (though rather trivial) case of interpolation in the case of scalars or functions.

We begin first with scalars. Suppose that A_0, B_0, A_1, B_1 are non-negative real numbers such that

$$A_0 \leq B_0 \tag{1.76}$$

and

$$A_1 \leq B_1. \tag{1.77}$$

Then clearly we will have

$$A_\theta \leq B_\theta \tag{1.78}$$

for every $0 \leq \theta \leq 1$, where we define

$$A_\theta := A_0^{1-\theta} A_1^\theta \tag{1.79}$$

and

$$B_\theta := B_0^{1-\theta} B_1^\theta; \tag{1.80}$$

indeed one simply raises (1.76) to the power $1-\theta$, (1.77) to the power θ, and multiplies the two inequalities together. Thus for instance, when $\theta = 1/2$ one obtains the geometric mean of (1.76) and (1.77):

$$A_0^{1/2} A_1^{1/2} \leq B_0^{1/2} B_1^{1/2}.$$

One can view A_θ and B_θ as the unique log-linear functions of θ (i.e. $\log A_\theta$, $\log B_\theta$ are (affine-)linear functions of θ) which equal their boundary values A_0, A_1 and B_0, B_1 respectively as $\theta = 0, 1$.

Example 1.11.1. If $A_0 = AL^{1/p_0}$ and $A_1 = AL^{1/p_1}$ for some $A, L > 0$ and $0 < p_0, p_1 \leq \infty$, then the log-linear interpolant A_θ is given by $A_\theta = AL^{1/p_\theta}$, where $0 < p_\theta \leq \infty$ is the quantity such that $\frac{1}{p_\theta} = \frac{1-\theta}{p_0} + \frac{\theta}{p_1}$.

The deduction of (1.78) from (1.76), (1.77) is utterly trivial, but there are still some useful lessons to be drawn from it. For instance, let us take $A_0 = A_1 = A$ for simplicity, so we are interpolating two upper bounds $A \leq B_0$, $A \leq B_1$ on the same quantity A to give a new bound $A \leq B_\theta$. But actually we have a refinement available to this bound, namely

$$A_\theta \leq B_\theta \min(\frac{B_0}{B_1}, \frac{B_1}{B_0})^\varepsilon \tag{1.81}$$

for any sufficiently small $\varepsilon > 0$ (indeed one can take any ε less than or equal to $\min(\theta, 1-\theta)$). Indeed one sees this simply by applying (1.78) with θ with $\theta-\varepsilon$ and $\theta+\varepsilon$ and taking minima. Thus we see that (1.78) is only sharp when the two original bounds B_0, B_1 are comparable; if instead we have $B_1 \sim 2^n B_0$ for some integer n, then (1.81) tells us that we can improve (1.78) by an exponentially decaying factor of $2^{-\varepsilon|n|}$. The geometric series formula tells us that such factors are absolutely summable, and so in practice it is often a useful heuristic to pretend that the $n = O(1)$ cases dominate so strongly that the other cases can be viewed as negligible by comparison.

Also, one can trivially extend the deduction of (1.78) from (1.76), (1.77) as follows: if $\theta \to A_\theta$ is a function from $[0,1]$ to $\mathbf{R}^+$ which is log-convex (thus $\theta \mapsto \log A_\theta$ is a convex function of θ, and (1.76), (1.77) hold for some $B_0, B_1 > 0$, then (1.78) holds for all intermediate θ also, where B_θ is of course defined by (1.80). Thus one can interpolate upper bounds on log-convex functions. However, one certainly cannot interpolate lower bounds: lower bounds on a log-convex function $\theta \to A_\theta$ at $\theta = 0$ and $\theta = 1$ yield no information about the value of, say, $A_{1/2}$. Similarly, one cannot extrapolate upper bounds on log-convex functions: an upper bound on, say, A_0 and $A_{1/2}$ does not give any information about A_1. (However, an upper bound on A_0 coupled with a *lower* bound on $A_{1/2}$ gives a lower bound on A_1; this is the contrapositive of an interpolation statement.)

Exercise 1.11.1. Show that the sum $f+g$, product fg, or pointwise maximum $\max(f,g)$ of two log-convex functions $f, g : [0,1] \to \mathbf{R}^+$ is log-convex.

Remark 1.11.2. Every non-negative log-convex function $\theta \mapsto A_\theta$ is convex, thus in particular $A_\theta \leq (1-\theta)A_0 + \theta A_1$ for all $0 \leq \theta \leq 1$ (note

that this generalises the arithmetic mean-geometric mean inequality). Of course, the converse statement is not true.

Now we turn to the complex version of the interpolation of log-convex functions, a result known as *Lindelöf's theorem*:

Theorem 1.11.3 (Lindelöf's theorem)**.** *Let $s \mapsto f(s)$ be a holomorphic function on the strip $S := \{\sigma + it : 0 \leq \sigma \leq 1; t \in \mathbf{R}\}$, which obeys the bound*

$$|f(\sigma + it)| \leq A \exp(\exp((\pi - \delta)t)) \tag{1.82}$$

for all $\sigma + it \in S$ and some constants $A, \delta > 0$. Suppose also that $|f(0 + it)| \leq B_0$ and $|f(1 + it)| \leq B_1$ for all $t \in \mathbf{R}$. Then we have $|f(\theta + it)| \leq B_\theta$ for all $0 \leq \theta \leq 1$ and $t \in \mathbf{R}$, where B_θ is of course defined by (1.80)*.*

Remark 1.11.4. The hypothesis (1.82) is a qualitative hypothesis rather than a quantitative one, since the exact values of A, σ do not show up in the conclusion. It is quite a mild condition; any function of exponential growth in t, or even with such super-exponential growth as $O(|t|^{|t|})$ or $O(e^{|t|^{O(1)}})$, will obey (1.82). The principle however fails without this hypothesis, as one can see for instance by considering the holomorphic function $f(s) := \exp(-i \exp(\pi i s))$.

Proof. Observe that the function $s \mapsto B_0^{1-s} B_1^s$ is holomorphic and non-zero on S, and has magnitude exactly B_θ on the line $\mathrm{Re}(s) = \theta$ for each $0 \leq \theta \leq 1$. Thus, by dividing f by this function (which worsens the qualitative bound (1.82) slightly) we may reduce to the case when $B_\theta = 1$ for all $0 \leq \theta \leq 1$.

Suppose we temporarily assume that $f(\sigma + it) \to 0$ as $|\sigma + it| \to \infty$. Then by the maximum modulus principle (applied to a sufficiently large rectangular portion of the strip), it must then attain a maximum on one of the two sides of the strip. But $|f| \leq 1$ on these two sides, and so $|f| \leq 1$ on the interior as well.

To remove the assumption that f goes to zero at infinity, we use the trick of giving ourselves an epsilon of room (Section 2.7). Namely, we multiply $f(s)$ by the holomorphic function $g_\varepsilon(s) := \exp(\varepsilon i \exp(i[(\pi - \delta/2)s + \delta/4]))$ for some $\varepsilon > 0$. A little complex arithmetic shows

that the function $f(s)g_\varepsilon(s)g_\varepsilon(1-s)$ goes to zero at infinity in S (the $g_\varepsilon(s)$ factor decays fast enough to damp out the growth of f as $\mathrm{Im}(s) \to -\infty$, while the $g_\varepsilon(1-s)$ damps out the growth as $\mathrm{Im}(s) \to +\infty$), and is bounded in magnitude by 1 on both sides of the strip S. Applying the previous case to this function, then taking limits as $\varepsilon \to 0$, we obtain the claim. □

Exercise 1.11.2. With the notation and hypotheses of Theorem 1.11.3, show that the function $\sigma \mapsto \sup_{t\in\mathbf{R}} |f(\sigma+it)|$ is log-convex on $[0,1]$.

Exercise 1.11.3 (Hadamard three-circles theorem)**.** Let f be a holomorphic function on an annulus $\{z \in \mathbf{C} : R_1 \leq |z| \leq R_2\}$. Show that the function $r \mapsto \sup_{\theta\in[0,2\pi]} |f(re^{i\theta})|$ is log-convex on $[R_1, R_2]$.

Exercise 1.11.4 (Phragmén-Lindelöf principle)**.** Let f be as in Theorem 1.11.3, but suppose that we have the bounds $f(0+it) \leq C(1+|t|)^{a_0}$ and $f(1+it) \leq C(1+|t|)^{a_1}$ for all $t \in \mathbf{R}$ and some exponents $a_0, a_1 \in \mathbf{R}$ and a constant $C > 0$. Show that one has $f(\sigma+it) \leq C'(1+|t|)^{(1-\sigma)a_0+\sigma a_1}$ for all $\sigma+it \in S$ and some constant C' (which is allowed to depend on the constants A, δ in (1.82)). (*Hint*: it is convenient to work first in a half-strip such as $\{\sigma+it \in S : t \geq T\}$ for some large T. Then multiply f by something like $\exp(-((1-z)a_0+za_1)\log(-iz))$ for some suitable branch of the logarithm and apply a variant of Theorem 1.11.3 for the half-strip. A more refined estimate in this regard is due to Rademacher [**Ra1959**].) This particular version of the principle gives the *convexity bound* for Dirichlet series such as the Riemann zeta function. Bounds which exploit the deeper properties of these functions to improve upon the convexity bound are known as *subconvexity bounds* and are of major importance in analytic number theory, which is of course well outside the scope of this course.

1.11.2. Interpolation of functions. We now turn to the interpolation in function spaces, focusing particularly on the Lebesgue spaces $L^p(X)$ and the weak Lebesgue spaces $L^{p,\infty}(X)$. Here, $X = (X, \mathcal{X}, \mu)$ is a fixed measure space. It will not matter much whether we deal with real or complex spaces; for sake of concretness we work with complex spaces. Then for $0 < p < \infty$, recall (see Section 1.3) that

$L^p(X)$ is the space of all functions $f : X \to \mathbf{C}$ whose L^p norm

$$\|f\|_{L^p(X)} := (\int_X |f|^p \, d\mu)^{1/p}$$

is finite, modulo almost everywhere equivalence. The space $L^\infty(X)$ is defined similarly, but where $\|f\|_{L^\infty(X)}$ is the essential supremum of $|f|$ on X.

A simple test case in which to understand the L^p norms better is that of a *step function* $f = A1_E$, where A is a non-negative number and E a set of finite measure. Then one has $\|f\|_{L^p(X)} = A\mu(E)^{1/p}$ for $0 < p \leq \infty$. Observe that this is a log-convex function of $1/p$. This is a general phenomenon:

Lemma 1.11.5 (Log-convexity of L^p norms)**.** *Let that $0 < p_0 < p_1 \leq \infty$ and $f \in L^{p_0}(X) \cap L^{p_1}(X)$. Then $f \in L^p(X)$ for all $p_0 \leq p \leq p_1$, and furthermore we have*

$$\|f\|_{L^{p_\theta}(X)} \leq \|f\|^{1-\theta}_{L^{p_0}(X)} \|f\|^{\theta}_{L^{p_1}(X)}$$

for all $0 \leq \theta \leq 1$, where the exponent p_θ is defined by $1/p_\theta := (1-\theta)/p_0 + \theta/p_1$.

In particular, we see that the function $1/p \mapsto \|f\|_{L^p(X)}$ is log-convex whenever the right-hand side is finite (and is in fact log-convex for all $0 \leq 1/p < \infty$, if one extends the definition of log-convexity to functions that can take the value $+\infty$). In other words, we can interpolate any two bounds $\|f\|_{L^{p_0}(X)} \leq B_0$ and $\|f\|_{L^{p_1}(X)} \leq B_1$ to obtain $\|f\|_{L^{p_\theta}(X)} \leq B_\theta$ for all $0 \leq \theta \leq 1$.

Let us give several proofs of this lemma. We will focus on the case $p_1 < \infty$; the endpoint case $p_1 = \infty$ can be proven directly, or by modifying the arguments below, or by using an appropriate limiting argument, and we leave the details to the reader.

The first proof is to use *Hölder's inequality*

$$\|f\|^{p_\theta}_{L^{p_\theta}(X)} = \int_X |f|^{(1-\theta)p_\theta} |f|^{\theta p_\theta} \, d\mu \leq \||f|^{(1-\theta)p_\theta}\|_{L^{p_0/((1-\theta)p_\theta)}} \||f|^{\theta p_\theta}\|_{L^{p_1/(\theta p_\theta)}}$$

when p_1 is finite (with some minor modifications in the case $p_1 = \infty$).

Another (closely related) proof proceeds by using the log-convexity inequality

$$|f(x)|^{p_\theta} \leq (1-\alpha)|f(x)|^{p_0} + \alpha |f(x)|^{p_1}$$

for all x, where $0 < \alpha < 1$ is the quantity such that $p_\theta = (1-\alpha)p_0 + \alpha p_1$. If one integrates this inequality in x, one already obtains the claim in the normalised case when $\|f\|_{L^{p_0}(X)} = \|f\|_{L^{p_1}(X)} = 1$. To obtain the general case, one can multiply the function f and the measure μ by appropriately chosen constants to obtain the above normalisation; we leave the details as an exercise to the reader. (The case when $\|f\|_{L^{p_0}(X)}$ or $\|f\|_{L^{p_1}(X)}$ vanishes is of course easy to handle separately.)

A third approach is more in the spirit of the real interpolation method, avoiding the use of convexity arguments. As in the second proof, we can reduce to the normalised case $\|f\|_{L^{p_0}(X)} = \|f\|_{L^{p_1}(X)} = 1$. We then split $f = f1_{|f|\leq 1} + f1_{|f|>1}$, where $1_{|f|\leq 1}$ is the indicator function to the set $\{x : |f(x)| \leq 1\}$, and similarly for $1_{|f|>1}$. Observe that

$$\|f1_{|f|\leq 1}\|_{L^{p_\theta}(X)}^{p_\theta} = \int_{|f|\leq 1} |f|^{p_\theta}\ d\mu \leq \int_X |f|^{p_0}\ d\mu = 1$$

and similarly

$$\|f1_{|f|>1}\|_{L^{p_\theta}(X)}^{p_\theta} = \int_{|f|>1} |f|^{p_\theta}\ d\mu \leq \int_X |f|^{p_1}\ d\mu = 1$$

and so by the quasi-triangle inequality (or triangle inequality, when $p_\theta \geq 1$)

$$\|f\|_{L^{p_\theta}(X)} \leq C$$

for some constant C depending on p_θ. Note, by the way, that this argument gives the inclusions

$$L^{p_0}(X) \cap L^{p_1}(X) \subset L^{p_\theta}(X) \subset L^{p_0}(X) + L^{p_1}(X). \tag{1.83}$$

This is off by a constant factor by what we want. But one can eliminate this constant by using the *tensor power trick* (Section 1.9 of *Structure and Randomness*). Indeed, if one replaces X with a Cartesian power X^M (with the product σ-algebra $\mathcal{X}^M$ and product measure μ^M), and replace f by the tensor power $f^{\otimes M} : (x_1, \ldots, x_m) \mapsto f(x_1)\ldots f(x_m)$, we see from many applications of the Fubini-Tonelli theorem that

$$\|f^{\otimes M}\|_{L^p(X)} = \|f\|_{L^p(X)}^M$$

for all p. In particular, $f^{\otimes M}$ obeys the same normalisation hypotheses as f, and thus by applying the previous inequality to $f^{\otimes M}$, we obtain

$$\|f\|^M_{L^{p_\theta}(X)} \leq C$$

for every M, where it is key to note that the constant C on the right is independent of M. Taking M^{th} roots and then sending $M \to \infty$, we obtain the claim.

Finally, we give a fourth proof in the spirit of the complex interpolation method. By replacing f by $|f|$ we may assume f is non-negative. By expressing non-negative measurable functions as the monotone limit of simple functions and using the monotone convergence theorem (Theorem 1.1.21), we may assume that f is a simple function, which is then necessarily of finite measure support from the L^p finiteness hypotheses. Now consider the function $s \mapsto \int_X |f|^{(1-s)p_0+sp_1}\, d\mu$. Expanding f out in terms of step functions we see that this is an analytic function of f which grows at most exponentially in s; also, by the triangle inequality this function has magnitude at most $\int_X |f|^{p_0}$ when $s = 0 + it$ and magnitude $\int_X |f|^{p_1}$ when $s = 1 + it$. Applying Theorem 1.11.3 and specialising to $s := \theta$ we obtain the claim.

Exercise 1.11.5. If $0 < \theta < 1$, show that equality holds in Lemma 1.11.5 if and only if $|f|$ is a step function.

Now we consider variants of interpolation in which the "strong" L^p spaces are replaced by their "weak" counterparts $L^{p,\infty}$. Given a measurable function $f : X \to \mathbf{C}$, we define the *distribution function* $\lambda_f : \mathbf{R}^+ \to [0, +\infty]$ by the formula

$$\lambda_f(t) := \mu(\{x \in X : |f(x)| \geq t\}) = \int_X 1_{|f|\geq t}\, d\mu.$$

This distribution function is closely connected to the L^p norms. Indeed, from the calculus identity

$$|f(x)|^p = p \int_0^\infty 1_{|f|\geq t} t^p\, \frac{dt}{t}$$

and the Fubini-Tonelli theorem, we obtain the formula

$$\|f\|^p_{L^p(X)} = p \int_0^\infty \lambda_f(t) t^p\, \frac{dt}{t} \tag{1.84}$$

for all $0 < p < \infty$, thus the L^p norms are essentially moments of the distribution function. The L^∞ norm is of course related to the distribution function by the formula

$$\|f\|_{L^\infty(X)} = \inf\{t \geq 0 : \lambda_f(t) = 0\}.$$

Exercise 1.11.6. Show that we have the relationship

$$\|f\|_{L^p(X)}^p \sim_p \sum_{n \in \mathbf{Z}} \lambda_f(2^n) 2^{np}$$

for any measurable $f : X \to \mathbf{C}$ and $0 < p < \infty$, where we use $X \sim_p Y$ to denote a pair of inequalities of the form $c_p Y \leq X \leq C_p Y$ for some constants $c_p, C_p > 0$ depending only on p. (*Hint*: $\lambda_f(t)$ is non-increasing in t.) Thus we can relate the L^p norms of f to the dyadic values $\lambda_f(2^n)$ of the distribution function; indeed, for any $0 < p \leq \infty$, $\|f\|_{L^p(X)}$ is comparable (up to constant factors depending on p) to the $\ell^p(\mathbf{Z})$ norm of the sequence $n \mapsto 2^n \lambda_f(2^n)^{1/p}$.

Another relationship between the L^p norms and the distribution function is given by observing that

$$\|f\|_{L^p(X)}^p = \int_X |f|^p \, d\mu \geq \int_{|f| \geq t} t^p \, d\mu = t^p \lambda_f(t)$$

for any $t > 0$, leading to *Chebyshev's inequality*

$$\lambda_f(t) \leq \frac{1}{t^p} \|f\|_{L^p(X)}^p.$$

(The $p = 1$ version of this inequality is also known as *Markov's inequality*. In probability theory, Chebyshev's inequality is often specialised to the case $p = 2$, and with f replaced by a normalised function $f - \mathbf{E}f$. Note that, as with many other Cyrillic names, there are also a large number of alternative spellings of Chebyshev in the Roman alphabet.)

Chebyshev's inequality motivates one to define the *weak L^p norm* $\|f\|_{L^{p,\infty}(X)}$ of a measurable function $f : X \to \mathbf{C}$ for $0 < p < \infty$ by the formula

$$\|f\|_{L^{p,\infty}(X)} := \sup_{t > 0} t \lambda_f(t)^{1/p},$$

thus Chebyshev's inequality can be expressed succinctly as

$$\|f\|_{L^{p,\infty}(X)} \leq \|f\|_{L^p(X)}.$$

It is also natural to adopt the convention that $\|f\|_{L^{\infty,\infty}(X)} = \|f\|_{L^\infty(X)}$. If $f, g : X \to \mathbf{C}$ are two functions, we have the inclusion

$$\{|f+g| \geq t\} \subset \{|f| \geq t/2\} \cup \{|g| \geq t/2\}$$

and hence

$$\lambda_{f+g}(t) \leq \lambda_f(t/2) + \lambda_g(t/2);$$

this easily leads to the quasi-triangle inequality

$$\|f+g\|_{L^{p,\infty}(X)} \lesssim_p \|f\|_{L^{p,\infty}(X)} + \|f\|_{L^{p,\infty}(X)}$$

where we use $X \lesssim_p Y$ as shorthand for the inequality $X \leq C_p Y$ for some constant C_p depending only on p (it can be a different constant at each use of the $\lesssim_p$ notation). [Note: in analytic number theory, it is more customary to use $\ll_p$ instead of $\lesssim_p$, following Vinogradov. However, in analysis $\ll$ is sometimes used instead to denote "much smaller than", e.g. $X \ll Y$ denotes the assertion $X \leq cY$ for some sufficiently small constant c.]

Let $L^{p,\infty}(X)$ be the space of all $f : X \to \mathbf{C}$ which have finite $L^{p,\infty}(X)$, modulo almost everywhere equivalence; this space is also known as weak $L^p(X)$. The quasi-triangle inequality soon implies that $L^{p,\infty}(X)$ is a quasi-normed vector space with the $L^{p,\infty}(X)$ quasi-norm, and Chebyshev's inequality asserts that $L^{p,\infty}(X)$ contains $L^p(X)$ as a subspace (though the L^p norm is not a restriction of the $L^{p,\infty}(X)$ norm).

Example 1.11.6. If $X = \mathbf{R}^n$ with the usual measure, and $0 < p < \infty$, then the function $f(x) := |x|^{-n/p}$ is in weak L^p, but not strong L^p. It is also not in strong or weak L^q for any other q. But the "local" component $|x|^{-n/p} 1_{|x| \leq 1}$ of f is in strong and weak L^q for all $q > p$, and the "global" component $|x|^{-n/p} 1_{|x|>1}$ of f is in strong and weak L^q for all $q > p$.

Exercise 1.11.7. For any $0 < p, q \leq \infty$ and $f : X \to \mathbf{C}$, define the *(dyadic) Lorentz norm* $\|f\|_{L^{p,q}(X)}$ to be $\ell^q(\mathbf{Z})$ norm of the sequence $n \mapsto 2^n \lambda_f(2^n)^{1/p}$, and define the *Lorentz space* $L^{p,q}(X)$ be the space of functions f with $\|f\|_{L^{p,q}(X)}$ finite, modulo almost everywhere equivalence. Show that $L^{p,q}(X)$ is a quasi-normed space, which is equivalent to $L^{p,\infty}(X)$ when $q = \infty$ and to $L^p(X)$ when $q = p$. Lorentz spaces arise naturally in more refined applications of

the real interpolation method, and are useful in certain "endpoint" estimates that fail for Lebesgue spaces, but which can be rescued by using Lorentz spaces instead. However, we will not pursue these applications in detail here.

Exercise 1.11.8. Let X be a finite set with counting measure, and let $f : X \to \mathbf{C}$ be a function. For any $0 < p < \infty$, show that

$$\|f\|_{L^{p,\infty}(X)} \leq \|f\|_{L^p(X)} \lesssim_p \log(1 + |X|)\|f\|_{L^{p,\infty}(X)}.$$

(*Hint*: to prove the second inequality, normalise $\|f\|_{L^{p,\infty}(X)} = 1$, and then manually dispose of the regions of X where f is too large or too small.) Thus, in some sense, weak L^p and strong L^p are equivalent "up to logarithmic factors".

One can interpolate weak L^p bounds just as one can strong L^p bounds: if $\|f\|_{L^{p_0,\infty}(X)} \leq B_0$ and $\|f\|_{L^{p_1,\infty}(X)} \leq B_1$, then

$$\|f\|_{L^{p_\theta,\infty}(X)} \leq B_\theta \tag{1.85}$$

for all $0 \leq \theta \leq 1$. Indeed, from the hypotheses we have

$$\lambda_f(t) \leq \frac{B_0^{p_0}}{t^{p_0}}$$

and

$$\lambda_f(t) \leq \frac{B_1^{p_1}}{t^{p_1}}$$

for all $t > 0$, and hence by scalar interpolation (using an interpolation parameter $0 < \alpha < 1$ defined by $p_\theta = (1-\alpha)p_0 + \alpha p_1$, and after doing some algebra) we have

$$\lambda_f(t) \leq \frac{B_\theta^{p_\theta}}{t^{p_\theta}} \tag{1.86}$$

for all $0 < \theta < 1$.

As remarked in the previous section, we can improve upon (1.86); indeed, if we define t_0 to be the unique value of t where $B_0^{p_0}/t^{p_0}$ and $B_1^{p_1}/t^{p_1}$ are equal, then we have

$$\lambda_f(t) \leq \frac{B_\theta^{p_\theta}}{t^{p_\theta}} \min(t/t_0, t_0/t)^\varepsilon$$

for some $\varepsilon > 0$ depending on p_0, p_1, θ. Inserting this improved bound into (1.84) we see that we can improve the weak-type bound (1.85)

to a strong-type bound

$$\|f\|_{L^{p_\theta}(X)} \leq C_{p_0,p_1,\theta} B_\theta \tag{1.87}$$

for some constant $C_{p_0,p_1,\theta}$. Note that one cannot use the tensor power trick this time to eliminate the constant $C_{p_0,p_1,\theta}$ as the weak L^p norms do not behave well with respect to tensor product. Indeed, the constant $C_{p_0,p_1,\theta}$ must diverge to infinity in the limit $\theta \to 0$ if $p_0 \neq \infty$, otherwise it would imply that the L^{p_0} norm is controlled by the $L^{p_0,\infty}$ norm, which is false by Example 1.11.6; similarly one must have a divergence as $\theta \to 1$ if $p_1 \neq \infty$.

Exercise 1.11.9. Let $0 < p_0 < p_1 \leq \infty$ and $0 < \theta < 1$. Refine the inclusions in (1.83) to

$$L^{p_0}(X) \cap L^{p_1}(X) \subset L^{p_0,\infty}(X) \cap L^{p_1,\infty}(X) \subset L^{p_\theta}(X) \subset$$
$$\subset L^{p_\theta,\infty}(X) \subset L^{p_0}(X) + L^{p_1}(X) \subset L^{p_0,\infty}(X) + L^{p_1,\infty}(X).$$

Define the *strong type diagram* of a function $f : X \to \mathbf{C}$ to be the set of all $1/p$ for which f lies in strong L^p, and the *weak type diagram* to be the set of all $1/p$ for which f lies in weak L^p. Then both the strong and weak type diagrams are connected subsets of $[0, +\infty)$, and the strong type diagram is contained in the weak type diagram, and contains in turn the interior of the weak type diagram. By experimenting with linear combinations of the examples in Example 1.11.6 we see that this is basically everything one can say about the strong and weak type diagrams, without further information on f or X.

Exercise 1.11.10. Let $f : X \to \mathbf{C}$ be a measurable function which is finite almost everywhere. Show that there exists a unique non-increasing left-continuous function $f^* : \mathbf{R}^+ \to \mathbf{R}^+$ such that $\lambda_{f^*}(t) = \lambda_f(t)$ for all $t \geq 0$, and in particular $\|f\|_{L^p(X)} = \|f^*\|_{L^p(\mathbf{R}^+)}$ for all $0 < p \leq \infty$, and $\|f\|_{L^{p,\infty}(X)} = \|f^*\|_{L^{p,\infty}(\mathbf{R}^+)}$. (*Hint*: first look for the formula that describes $f^*(x)$ for some $x > 0$ in terms of $\lambda_f(t)$.) The function f^* is known as the *non-increasing rearrangement* of f, and the spaces $L^p(X)$ and $L^{p,\infty}(X)$ are examples of *rearrangement-invariant spaces*. There are a class of useful *rearrangement inequalities* that relate f to its rearrangements, and which can be used to clarify the structure of rearrangement-invariant spaces, but we will not pursue this topic here.

Exercise 1.11.11. Let $(X, \mathcal{X}, \mu)$ be a σ-finite measure space, let $1 < p < \infty$, and $f : X \to \mathbf{C}$ be a measurable function. Show that the following are equivalent:

- f lies in $L^{p,\infty}(X)$, thus $\|f\|_{L^{p,\infty}(X)} \leq C$ for some finite C.
- There exists a constant C' such that $|\int_X f 1_E \, d\mu| \leq C' \mu(E)^{1/p'}$ for all sets E of finite measure.

Furthermore show that the best constants C, C' in the above statements are equivalent up to multiplicative constants depending on p, thus $C \sim_p C'$. Conclude that the modified weak $L^{p,\infty}(X)$ norm $\|f\|_{\tilde{L}^{p,\infty}(X)} := \sup_E \mu(E)^{-1/p'} |\int_X f 1_E \, d\mu|$, where E ranges over all sets of positive finite measure, is a genuine norm on $L^{p,\infty}(X)$ which is equivalent to the $L^{p,\infty}(X)$ quasinorm.

Exercise 1.11.12. Let $n > 1$ be an integer. Find a probability space $(X, \mathcal{X}, \mu)$ and functions $f_1, \ldots, f_n : X \to \mathbf{R}$ with $\|f_j\|_{L^{1,\infty}(X)} \leq 1$ for $j = 1, \ldots, n$ such that $\|\sum_{j=1}^n f_j\|_{L^{1,\infty}(X)} \geq cn \log n$ for some absolute constant $c > 0$. (*Hint*: exploit the logarithmic divergence of the harmonic series $\sum_{j=1}^\infty \frac{1}{j}$.) Conclude that there exists a probability space X such that the $L^{1,\infty}(X)$ quasi-norm is not equivalent to an actual norm.

Exercise 1.11.13. Let $(X, \mathcal{X}, \mu)$ be a σ-finite measure space, let $0 < p < \infty$, and $f : X \to \mathbf{C}$ be a measurable function. Show that the following are equivalent:

- f lies in $L^{p,\infty}(X)$.
- There exists a constant C such that for every set E of finite measure, there exists a subset E' with $\mu(E') \geq \frac{1}{2}\mu(E)$ such that $|\int_X f 1_{E'} \, d\mu| \leq C\mu(E)^{1/p'}$.

Exercise 1.11.14. Let $(X, \mathcal{X}, \mu)$ be a measure space of finite measure, and $f : X \to \mathbf{C}$ be a measurable function. Show that the following two statements are equivalent:

- There exists a constant $C > 0$ such that $\|f\|_{L^p(X)} \leq Cp$ for all $1 \leq p < \infty$.
- There exists a constant $c > 0$ such that $\int_X e^{c|f|} \, d\mu < \infty$.

1.11.3. Interpolation of operators. We turn at last to the central topic of these notes, which is interpolation of operators T between functions on two fixed measure spaces $X = (X, \mathcal{X}, \mu)$ and $Y = (Y, \mathcal{Y}, \nu)$. To avoid some (very minor) technicalities we will make the mild assumption throughout that X and Y are both σ-finite, although much of the theory here extends to the non-σ-finite setting.

A typical situation is that of a linear operator T which maps one $L^{p_0}(X)$ space to another $L^{q_0}(Y)$, and also maps $L^{p_1}(X)$ to $L^{q_1}(Y)$ for some exponents $0 < p_0, p_1, q_0, q_1 \leq \infty$; thus (by linearity) T will map the larger vector space $L^{p_0}(X) + L^{p_1}(X)$ to $L^{q_0}(Y) + L^{q_1}(Y)$, and one has some estimates of the form

$$\|Tf\|_{L^{q_0}(Y)} \leq B_0 \|f\|_{L^{p_0}(X)} \tag{1.88}$$

and

$$\|Tf\|_{L^{q_1}(Y)} \leq B_1 \|f\|_{L^{p_1}(X)} \tag{1.89}$$

for all $f \in L^{p_0}(X), f \in L^{p_1}(X)$ respectively, and some $B_0, B_1 > 0$. We would like to then interpolate to say something about how T maps $L^{p_\theta}(X)$ to $L^{q_\theta}(Y)$.

The complex interpolation method gives a satisfactory result as long as the exponents allow one to use duality methods, a result known as the *Riesz-Thorin theorem*:

Theorem 1.11.7 (Riesz-Thorin theorem)**.** *Let $0 < p_0, p_1 \leq \infty$ and $1 \leq q_0, q_1 \leq \infty$. Let $T : L^{p_0}(X) + L^{p_1}(X) \to L^{q_0}(Y) + L^{q_1}(Y)$ be a linear operator obeying the bounds* (1.88), (1.89) *for all $f \in L^{p_0}(X), f \in L^{p_1}(X)$ respectively, and some $B_0, B_1 > 0$. Then we have*

$$\|Tf\|_{L^{q_\theta}(Y)} \leq B_\theta \|f\|_{L^{p_\theta}(X)}$$

for all $0 < \theta < 1$ and $f \in L^{p_\theta}(X)$, where $1/p_\theta := 1 - \theta/p_0 + \theta/p_1$, $1/q_\theta := 1 - \theta/q_0 + \theta/q_1$, and $B_\theta := B_0^{1-\theta} B_1^\theta$.

Remark 1.11.8. When X is a point, this theorem essentially collapses to Lemma 1.11.5 (and when Y is a point, this is a dual formulation of that lemma); and when X and Y are both points; this collapses to interpolation of scalars.

Proof. If $p_0 = p_1$ then the claim follows from Lemma 1.11.5, so we may assume $p_0 \neq p_1$, which in particular forces p_θ to be finite. By symmetry we can take $p_0 < p_1$. By multiplying the measures μ and ν (or the operator T) by various constants, we can normalise $B_0 = B_1 = 1$ (the case when $B_0 = 0$ or $B_1 = 0$ is trivial). Thus we have $B_\theta = 1$ also.

By Hölder's inequality, the bound (1.88) implies that

$$|\int_Y (Tf)g \, d\nu| \leq \|f\|_{L^{p_0}(X)} \|g\|_{L^{q_0'}(Y)} \tag{1.90}$$

for all $f \in L^{p_0}(X)$ and $g \in L^{q_0'}(Y)$, where q_0' is the dual exponent of q_0. Similarly we have

$$|\int_Y (Tf)g \, d\nu| \leq \|f\|_{L^{p_1}(X)} \|g\|_{L^{q_1'}(Y)} \tag{1.91}$$

for all $f \in L^{p_1}(X)$ and $g \in L^{q_1'}(Y)$.

We now claim that

$$|\int_Y (Tf)g \, d\nu| \leq \|f\|_{L^{p_\theta}(X)} \|g\|_{L^{q_\theta'}(Y)} \tag{1.92}$$

for all f, g that are simple functions with finite measure support. To see this, we first normalise $\|f\|_{L^{p_\theta}(X)} = \|g\|_{L^{q_\theta'}(Y)} = 1$. Observe that we can write $f = |f| \operatorname{sgn}(f)$, $g = |g| \operatorname{sgn}(g)$ for some functions $\operatorname{sgn}(f), \operatorname{sgn}(g)$ of magnitude at most 1. If we then introduce the quantity

$$F(s) := \int_Y (T[|f|^{(1-s)p_\theta/p_0 + sp_\theta/p_1} \operatorname{sgn}(f)])[|g|^{(1-s)q_\theta'/q_0' + sq_\theta'/q_1'} \operatorname{sgn}(g)] \, d\nu$$

(with the conventions that $q_\theta'/q_0', q_\theta'/q_1' = 1$ in the endpoint case $q_0' = q_1' = q_\theta' = \infty$) we see that F is a holomorphic function of s of at most exponential growth which equals $\int_Y (Tf)g \, d\nu$ when $s = \theta$. When instead $s = 0 + it$, an application of (1.90) shows that $|F(s)| \leq 1$; a similar claim obtains when $s = 1 + it$ using (1.91). The claim now follows from Theorem 1.11.3.

The estimate (1.92) has currently been established for simple functions f, g with finite measure support. But one can extend the claim to any $f \in L^{p_\theta}(X)$ (keeping g simple with finite measure support) by decomposing f into a bounded function and a function of finite measure support, approximating the former in $L^{p_\theta}(X) \cap L^{p_1}(X)$

by simple functions of finite measure support, and approximating the latter in $L^{p_\theta}(X) \cap L^{p_0}(X)$ by simple functions of finite measure support, and taking limits using (1.90), (1.91) to justify the passage to the limit. One can then also allow arbitrary $g \in L^{q_\theta'}(Y)$ by using the monotone convergence theorem (Theorem 1.1.21). The claim now follows from the duality between $L^{q_1}(Y)$ and $L^{q_1'}(Y)$. □

Suppose one has a linear operator T that maps simple functions of finite measure support on X to measurable functions on Y (modulo almost everywhere equivalence). We say that such an operator is of *strong type* (p, q) if it can be extended in a continuous fashion to an operator on $L^p(X)$ to an operator on $L^q(Y)$; this is equivalent to having an estimate of the form $\|Tf\|_{L^q(Y)} \leq B\|f\|_{L^p(X)}$ for all simple functions f of finite measure support. (The extension is unique if p is finite or if X has finite measure, due to the density of simple functions of finite measure support in those cases. Annoyingly, uniqueness fails for L^∞ of an infinite measure space, though this turns out not to cause much difficulty in practice, as the conclusions of interpolation methods are usually for finite exponents p.) Define the *strong type diagram* to be the set of all $(1/p, 1/q)$ such that T is of strong type (p, q). The Riesz-Thorin theorem tells us that if T is of strong type (p_0, q_0) and (p_1, q_1) with $0 < p_0, p_1 \leq \infty$ and $1 \leq q_0, q_1 \leq \infty$, then T is also of strong type (p_θ, q_θ) for all $0 < \theta < 1$; thus the strong type diagram contains the closed line segment connecting $(1/p_0, 1/q_0)$ with $(1/p_1, 1/q_1)$. Thus the strong type diagram of T is convex in $[0, +\infty) \times [0, 1]$ at least. (As we shall see later, it is in fact convex in all of $[0, +\infty)^2$.) Furthermore, on the intersection of the strong type diagram with $[0, 1] \times [0, +\infty)$, the operator norm $\|T\|_{L^p(X) \to L^q(Y)}$ is a log-convex function of $(1/p, 1/q)$.

Exercise 1.11.15. If $X = Y = [0, 1]$ with the usual measure, show that the strong type diagram of the identity operator is the triangle $\{(1/p, 1/q) \in [0, +\infty) \times [0, +\infty) : 1/p \leq 1/q\}$. If instead $X = Y = \mathbf{Z}$ with the usual counting measure, show that the strong type diagram of the identity operator is the triangle $\{(1/p, 1/q) \in [0, +\infty) \times [0, +\infty) : 1/p \geq 1/q\}$. What is the strong type diagram of the identity when $X = Y = \mathbf{R}$ with the usual measure?

Exercise 1.11.16. Let T (resp. T^*) be a linear operator from simple functions of finite measure support on Y (resp. X) to measurable functions on Y (resp. X) modulo a.e. equivalence that are absolutely integrable on finite measure sets. We say T, T^* are *formally adjoint* if we have $\int_Y (Tf)\overline{g}\ d\nu = \int_X f\overline{T^*g}\ d\mu$ for all simple functions f, g of finite measure support on X, Y respectively. If $1 \leq p, q \leq \infty$, show that T is of strong type (p, q) if and only if T^* is of strong type (q', p'). Thus, taking formal adjoints reflects the strong type diagram around the line of duality $1/p + 1/q = 1$, at least inside the Banach space region $[0, 1]^2$.

Remark 1.11.9. There is a powerful extension of the Riesz-Thorin theorem known as the *Stein interpolation theorem*, in which the single operator T is replaced by a family of operators T_s for $s \in S$ that vary holomorphically in s in the sense that $\int_Y (T_s 1_E) 1_F\ d\nu$ is a holomorphic function of s for any sets E, F of finite measure. Roughly speaking, the Stein interpolation theorem asserts that if T_{j+it} is of strong type (p_j, q_j) for $j = 0, 1$ with a bound growing at most exponentially in t, and T_s itself grows at most exponentially in t in some sense, then T_θ will be of strong type (p_θ, q_θ). A precise statement of the theorem and some applications can be found in [**St1993**].

Now we turn to the real interpolation method. Instead of linear operators, it is now convenient to consider *sublinear operators* T mapping simple functions $f : X \to \mathbf{C}$ of finite measure support in X to $[0, +\infty]$-valued measurable functions on Y (modulo almost everywhere equivalence, as usual), obeying the homogeneity relationship

$$|T(cf)| = |c||Tf|$$

and the pointwise bound

$$|T(f + g)| \leq |Tf| + |Tg|$$

for all $c \in \mathbf{C}$, and all simple functions f, g of finite measure support.

Every linear operator is sublinear; also, the absolute value $Tf := |Sf|$ of a linear (or sublinear) operator is also sublinear. More generally, any *maximal operator* of the form $Tf := \sup_{\alpha \in A} |S_\alpha f|$, where $(S_\alpha)_{\alpha \in A}$ is a family of linear operators, is also a non-negative sublinear operator; note that one can also replace the supremum here by any

other norm in α, e.g. one could take an ℓ^p norm $(\sum_{\alpha \in A} |S_\alpha f|^p)^{1/p}$ for any $1 \leq p \leq \infty$. (After $p = \infty$ and $p = 1$, a particularly common case is when $p = 2$, in which case T is known as a *square function*.)

The basic theory of sublinear operators is similar to that of linear operators in some respects. For instance, continuity is still equivalent to boundedness:

Exercise 1.11.17. Let T be a sublinear operator, and let $0 < p, q \leq \infty$. Then the following are equivalent:

- T can be extended to a continuous operator from $L^p(X)$ to $L^q(Y)$.
- There exists a constant $B > 0$ such that $\|Tf\|_{L^q(Y)} \leq B\|f\|_{L^p(X)}$ for all simple functions f of finite measure support.
- T can be extended to a operator from $L^p(X)$ to $L^q(Y)$ such that $\|Tf\|_{L^q(Y)} \leq B\|f\|_{L^p(X)}$ for all $f \in L^p(X)$ and some $B > 0$.

Show that the extension mentioned above is unique of p is finite, or if X has finite measure. Finally, show that the same equivalences hold if $L^q(Y)$ is replaced by $L^{q,\infty}(Y)$ throughout.

We say that T is of *strong type* (p, q) if any of the above equivalent statements (for $L^q(Y)$) hold, and of *weak type* (p, q) if any of the above equivalent statements (for $L^{q,\infty}(Y)$) hold. We say that a linear operator S is of strong or weak type (p, q) if its non-negative counterpart $|S|$ is; note that this is compatible with our previous definition of strong type for such operators. Also, Chebyshev's inequality tells us that strong type (p, q) implies weak type (p, q).

We now give the real interpolation counterpart of the Riesz-Thorin theorem, namely the *Marcinkeiwicz interpolation theorem*:

Theorem 1.11.10 (Marcinkiewicz interpolation theorem)**.** *Let $0 < p_0, p_1, q_0, q_1 \leq \infty$ and $0 < \theta < 1$ be such that $q_0 \neq q_1$, and $p_i \leq q_i$ for $i = 0, 1$. Let T be a sublinear operator which is of weak type (p_0, q_0) and of weak type (p_1, q_1). Then T is of strong type (p_θ, q_θ).*

Remark 1.11.11. Of course, the same claim applies to linear operators S by setting $T := |S|$. One can also extend the argument to *quasi-linear* operators, in which the pointwise bound $|T(f+g)| \leq |Tf|+|Tg|$ is replaced by $|T(f+g)| \leq C(|Tf|+|Tg|)$ for some constant $C > 0$, but this generalisation only appears occasionally in applications. The conditions $p_0 \leq q_0, p_1 \leq q_1$ can be replaced by the variant condition $p_\theta \leq q_\theta$ (see Exercise 1.11.19, Exercise 1.11.21), but cannot be eliminated entirely: see Exercise 1.11.20. The precise hypotheses required on $p_0, p_1, q_0, q_1, p_\theta, q_\theta$ are rather technical and I recommend that they be ignored on a first reading.

Proof. For notational reasons it is convenient to take q_0, q_1 finite; however the arguments below can be modified without much difficulty to deal with the infinite case (or one can use a suitable limiting argument); we leave this to the interested reader.

By hypothesis, there exist constants $B_0, B_1 > 0$ such that

$$\lambda_{Tf}(t) \leq B_0^{q_0} \|f\|_{L^{p_0}(X)}^{q_0} / t^{q_0} \tag{1.93}$$

and

$$\lambda_{Tf}(t) \leq B_1^{q_1} \|f\|_{L^{p_1}(X)}^{q_1} / t^{q_1} \tag{1.94}$$

for all simple functions f of finite measure support, and all $t > 0$. Let us write $A \lesssim B$ to denote $A \leq C_{p_0,p_1,q_0,q_1,\theta,B_0,B_1} B$ for some constant $C_{p_0,p_1,q_0,q_1,\theta,B_0,B_1}$ depending on the indicated parameters. By (1.84), it will suffice to show that

$$\int_0^\infty \lambda_{Tf}(t) t^{q_\theta} \frac{dt}{t} \lesssim \|f\|_{L^{p_\theta}(X)}^{q_\theta}.$$

By homogeneity we can normalise $\|f\|_{L^{p_\theta}(X)} = 1$.

Actually, it will be more slightly convenient to work with the dyadic version of the above estimate, namely

$$\sum_{n \in \mathbf{Z}} \lambda_{Tf}(2^n) 2^{q_\theta n} \lesssim 1; \tag{1.95}$$

see Exercise 1.11.6. The hypothesis $\|f\|_{L^{p_\theta}(X)} = 1$ similarly implies that

$$\sum_{m \in \mathbf{Z}} \lambda_f(2^m) 2^{p_\theta m} \lesssim 1. \tag{1.96}$$

The basic idea is then to get enough control on the numbers $\lambda_{Tf}(2^n)$ in terms of the numbers $\lambda_f(2^m)$ that one can deduce (1.95) from (1.96).

When $p_0 = p_1$, the claim follows from direct substitution of (1.91), (1.94) (see also the discussion in the previous section about interpolating strong L^p bounds from weak ones), so let us assume $p_0 \neq p_1$; by symmetry we may take $p_0 < p_1$, and thus $p_0 < p_\theta < p_1$. In this case we cannot directly apply (1.91), (1.94) because we only control f in L^{p_θ}, not L^{p_0} or L^{p_1}. To get around this, we use the basic real interpolation trick of *decomposing* f into pieces. There are two basic choices for what decomposition to pick. On one hand, one could adopt a "minimalistic" approach and just decompose into two pieces

$$f = f_{\geq s} + f_{<s}$$

where $f_{\geq s} := f1_{|f|\geq s}$ and $f_{<s} := f1_{|f|<s}$, and the threshold s is a parameter (depending on n) to be optimised later. Or we could adopt a "maximalistic" approach and perform the dyadic decomposition

$$f = \sum_{m\in\mathbf{Z}} f_m$$

where $f_m = f1_{2^m\leq|f|<2^{m+1}}$. (Note that only finitely many of the f_m are non-zero, as we are assuming f to be a simple function.) We will adopt the latter approach, in order to illustrate the dyadic decomposition method; the former approach also works, but we leave it as an exercise to the interested reader.

From sublinearity we have the pointwise estimate

$$Tf \leq \sum_m Tf_m$$

which implies that

$$\lambda_{Tf}(2^n) \leq \sum_m \lambda_{Tf_m}(c_{n,m}2^n)$$

whenever $c_{n,m}$ are positive constants such that $\sum_m c_{n,m} = 1$, but for which we are otherwise at liberty to choose. We will set aside the problem of deciding what the optimal choice of $c_{n,m}$ is for now, and continue with the proof.

From (1.91), (1.94), we have two bounds for the quantity $\lambda_{Tf_m}(c_{n,m}2^n)$, namely

$$\lambda_{Tf_m}(c_{n,m}2^n) \lesssim c_{n,m}^{-q_0} 2^{-nq_0} \|f_m\|_{L^{p_0}(X)}^{q_0}$$

and

$$\lambda_{Tf_m}(c_{n,m}2^n) \lesssim c_{n,m}^{-q_1} 2^{-nq_1} \|f_m\|_{L^{p_1}(X)}^{q_1}.$$

From construction of f_m we can bound

$$\|f_m\|_{L^{p_0}(X)} \lesssim 2^m \lambda_f(2^m)^{1/p_0}$$

and similarly for p_1, and thus we have

$$\lambda_{Tf_m}(c_{n,m}2^n) \lesssim c_{n,m}^{-q_i} 2^{-nq_i} 2^{mq_i} \lambda_f(2^m)^{q_i/p_i}.$$

for $i = 0, 1$. To prove (1.95), it thus suffices to show that

$$\sum_n 2^{nq_\theta} \sum_m \min_{i=0,1} c_{n,m}^{-q_i} 2^{-nq_i} 2^{mq_i} \lambda_f(2^m)^{q_i/p_i} \lesssim 1.$$

It is convenient to introduce the quantities $a_m := \lambda_f(2^m)2^{mp_\theta}$ appearing in (1.96), thus

$$\sum_m a_m \lesssim 1$$

and our task is to show that

$$\sum_n 2^{nq_\theta} \sum_m \min_{i=0,1} c_{n,m}^{-q_i} 2^{-nq_i} 2^{mq_i} 2^{-mq_i p_\theta/p_i} a_m^{q_i/p_i} \lesssim 1.$$

Since $p_i \leq q_i$, we have $a_m^{q_i/p_i} \lesssim a_m$, and so we are reduced to the purely numerical task of locating constants $c_{n,m}$ with $\sum_m c_{n,m} \leq 1$ for all n such that

$$\sum_n 2^{nq_\theta} \sum_m \min_{i=0,1} c_{n,m}^{-q_i} 2^{-nq_i} 2^{mq_i} 2^{-mq_i p_\theta/p_i} \lesssim 1 \tag{1.97}$$

for all m.

We can simplify this expression a bit by collecting terms and making some substitutions. The points $(1/p_0, 1/q_0), (1/p_\theta, 1/q_\theta), (1/p_1, 1/q_1)$ are collinear, and we can capture this by writing

$$\frac{1}{p_i} = \frac{1}{p_\theta} + x_i; \quad \frac{1}{q_i} = \frac{1}{q_\theta} + \alpha x_i$$

for some $x_0 > 0 > x_1$ and some $\alpha \in \mathbf{R}$. We can then simplify the left-hand side of (1.97) to

$$\sum_m \min_{i=0,1} (c_{n,m}^{-1} 2^{n\alpha q_\theta - mp_\theta})^{q_i x_i}.$$

Note that $q_0 x_0$ is positive and $q_1 x_1$ is negative. If we then pick $c_{n,m}$ to be a sufficiently small multiple of $2^{|n\alpha q_\theta - m p_\theta|/2}$ (say), we obtain the claim by summing geometric series. □

Remark 1.11.12. A closer inspection of the proof (or a rescaling argument to reduce to the normalised case $B_0 = B_1 = 1$, as in preceding sections) reveals that one establishes the estimate

$$\|Tf\|_{L^{q_\theta}(Y)} \leq C_{p_0,p_1,q_0,q_1,\theta,C} B_0^{1-\theta} B_1^\theta \|f\|_{L^{p_\theta}(X)}$$

for all simple functions f of finite measure support (or for all $f \in L^{p_\theta}(X)$, if one works with the continuous extension of T to such functions), and some constant $C_{p_0,p_1,q_0,q_1,\theta,C} > 0$. Thus the conclusion here is weaker by a multiplicative constant from that in the Riesz-Thorin theorem, but the hypotheses are weaker too (weak-type instead of strong-type). Indeed, we see that the constant $C_{p_0,p_1,q_0,q_1,\theta}$ must blow up as $\theta \to 0$ or $\theta \to 1$.

The power of the Marcinkiewicz interpolation theorem, as compared to the Riesz-Thorin theorem, is that it allows one to weaken the hypotheses on T from strong type to weak type. Actually, it can be weakened further. We say that a non-negative sublinear operator T is *restricted weak-type* (p,q) for some $0 < p,q \leq \infty$ if there is a constant $B > 0$ such that

$$\|Tf\|_{L^{q,\infty}(Y)} \leq B\mu(E)^{1/p}$$

for all sets E of finite measure and all simple functions f with $|f| \leq 1_E$. Clearly restricted weak-type (p,q) is implied by weak-type (p,q), and thus by strong-type (p,q). (One can also define the notion of *restricted strong-type* (p,q) by replacing $L^{q,\infty}(Y)$ with $L^q(Y)$; this is between strong-type (p,q) and restricted weak-type (p,q), but is incomparable to weak-type (p,q).)

Exercise 1.11.18. Show that the Marcinkiewicz interpolation theorem continues to hold if the weak-type hypotheses are replaced by restricted weak-type hypothesis. (*Hint*: where were the weak-type hypotheses used in the proof?)

We thus see that the strong-type diagram of T contains the interior of the restricted weak-type or weak-type diagrams of T, at least in the triangular region $\{(1/p, 1/q) \in [0,+\infty)^2 : p \geq q\}$.

Exercise 1.11.19. Suppose that T is a sublinear operator of restricted weak-type (p_0, q_0) and (p_1, q_1) for some $0 < p_0, p_1, q_0, q_1 \leq \infty$. Show that T is of restricted weak-type (p_θ, q_θ) for any $0 < \theta < 1$, or in other words the restricted type diagram is convex in $[0, +\infty)^2$. (This is an easy result requiring only interpolation of scalars.) Conclude that the hypotheses $p_0 \leq q_0, p_1 \leq q_1$ in the Marcinkiewicz interpolation theorem can be replaced by the variant $p_\theta < q_\theta$.

Exercise 1.11.20. For any $\alpha \in \mathbf{R}$, let X_α be the natural numbers $\mathbf{N}$ with the weighted counting measure $\sum_{n \in \mathbf{N}} 2^{\alpha n} \delta_n$, thus each point n has mass $2^{\alpha n}$. Show that if $\alpha > \beta > 0$, then the identity operator from X_α to X_β is of weak-type (p, q) but not strong-type (p, q) when $1 < p, q < \infty$ and $\alpha/p = \beta/q$. Conclude that the hypotheses $p_0 \leq q_0, p_1 \leq q_1$ cannot be dropped entirely.

Exercise 1.11.21. Suppose we are in the situation of the Marcinkiewicz interpolation theorem, with the hypotheses $p_0 \leq q_0, p_1 \leq q_1$ replaced by $p_0 \neq p_1$. Show that for all $0 < \theta < 1$ and $1 \leq r \leq \infty$ there exists a $B > 0$ such that

$$\|Tf\|_{L^{q_\theta, r}(Y)} \leq B \|f\|_{L^{p_\theta, r}(X)}$$

for all simple functions f of finite measure support, where the Lorentz norms $L^{p,q}$ were defined in Exercise 1.11.7. (*Hint*: repeat the proof of the Marcinkiewicz interpolation theorem, but partition the sum $\sum_{n,m}$ into regions of the form $\{n\alpha q_\theta - m p_\theta = k + O(1)\}$ for integer k. Obtain a bound for each summand which decreases geometrically as $k \to \pm\infty$.) Conclude that the hypotheses $p_0 \leq q_0, p_1 \leq q_1$ in the Marcinkiewicz interpolation theorem can be replaced by $p_\theta \leq q_\theta$. This Lorentz space version of the interpolation theorem is in some sense the "right" version of the theorem, but the Lorentz spaces are slightly more technical to deal with than the Lebesgue spaces, and the Lebesgue space version of Marcinkiewicz interpolation is largely sufficient for most applications.

Exercise 1.11.22. For $i = 1, 2$, let $X_i = (X_i, \mathcal{X}_i, \mu_i), Y_i = (Y_i, \mathcal{Y}_i, \nu_i)$ be σ-finite measure spaces, and let T_i be a linear operator from simple functions of finite measure support on X_i to measurable functions on Y_i (modulo almost everywhere equivalence, as always). Let $X = X_1 \times X_2$, $Y = Y_1 \times Y_2$ be the product spaces (with product σ-algebra

and product measure). Show that there exists a unique (modulo a.e. equivalence) linear operator T defined on linear combinations of indicator functions $1_{E_1 \times E_2}$ of product sets of sets $E_1 \subset X_1$, $E_2 \subset X_2$ of finite measure, such that

$$T1_{E_1 \times E_2}(y_1, y_2) := T_1 1_{E_1}(y_1) T_2 1_{E_2}(y_2)$$

for a.e. $(y_1, y_2) \in Y$; we refer to T as the *tensor product* of T_1 and T_2 and write $T = T_1 \otimes T_2$. Show that if T_1, T_2 are of strong-type (p, q) for some $1 \leq p, q < \infty$ with operator norms B_1, B_2 respectively, then T can be extended to a bounded linear operator on $L^p(X)$ to $L^q(Y)$ with operator norm exactly equal to $B_1 B_2$, thus

$$\|T_1 \otimes T_2\|_{L^p(X_1 \times X_2) \to L^q(Y_1 \times Y_2)} = \|T_1\|_{L^p(X_1) \to L^q(Y_1)} \|T_2\|_{L^p(X_2) \to L^q(Y_2)}.$$

(*Hint*: for the lower bound, show that $T_1 \otimes T_2(f_1 \otimes f_2) = (T_1 f_1) \otimes (T_2 f_2)$ for all simple functions f_1, f_2. For the upper bound, express $T_1 \times T_2$ as the composition of two other operators $T_1 \otimes I_1$ and $I_2 \otimes T_2$ for some identity operators I_1, I_2, and establish operator norm bounds on these two operators separately.) Use this and the tensor power trick to deduce the Riesz-Thorin theorem (in the special case when $1 \leq p_i \leq q_i < \infty$ for $i = 0, 1$, and $q_0 \neq q_1$) from the Marcinkiewicz interpolation theorem. Thus one can (with some effort) avoid the use of complex variable methods to prove the Riesz-Thorin theorem, at least in some cases.

Exercise 1.11.23 (Hölder's inequality for Lorentz spaces)**.** Let $f \in L^{p_1, r_1}(X)$ and $g \in L^{p_2, r_2}(X)$ for some $0 < p_1, p_2, r_1, r_2 \leq \infty$. Show that $fg \in L^{p_3, r_3}(X)$, where $1/p_3 = 1/p_1 + 1/p_2$ and $1/r_3 = 1/r_1 + 1/r_2$, with the estimate

$$\|fg\|_{L^{p_3, r_3}(X)} \leq C_{p_1, p_2, r_1, r_2} \|f\|_{L^{p_1, r_1}(X)} \|g\|_{L^{p_2, r_2}(X)}$$

for some constant C_{p_1, p_2, r_1, r_2}. (This estimate is due to O'Neil[**ON1963**].)

Remark 1.11.13. Just as interpolation of functions can be clarified by using step functions $f = A1_E$ as a test case, it is instructive to use rank one operators such as

$$Tf := A\langle f, 1_E \rangle 1_F = A(\int_E f \, d\mu) 1_F$$

where $E \subset X, F \subset Y$ are finite measure sets, as test cases for the real and complex interpolation methods. (After understanding the

rank one case, I then recommend looking at the rank two case, e.g. $Tf := A_1\langle f, 1_{E_1}\rangle 1_{F_1} + A_2\langle f, 1_{E_2}\rangle 1_{F_2}$, where E_2, F_2 could be very different in size from E_1, F_1.)

1.11.4. Some examples of interpolation. Now we apply the interpolation theorems to some classes of operators. An important such class is given by the *integral operators*

$$Tf(y) := \int_X K(x,y) f(x)\, d\mu(x)$$

from functions $f : X \to \mathbf{C}$ to functions $Tf : Y \to \mathbf{C}$, where $K : X \times Y \to \mathbf{C}$ is a fixed measurable function, known as the *kernel* of the integral operator T. Of course, this integral is not necessarily convergent, so we will also need to study the sublinear analogue

$$|T|f(y) := \int_X |K(x,y)||f(x)|\, d\mu(x)$$

which is well-defined (though it may be infinite).

The following useful lemma gives us strong-type bounds on $|T|$ and hence T, assuming certain L^p type bounds on the rows and columns of K.

Lemma 1.11.14 (Schur's test)**.** *Let $K : X \times Y \to \mathbf{C}$ be a measurable function obeying the bounds*

$$\|K(x,\cdot)\|_{L^{q_0}(Y)} \leq B_0$$

for almost every $x \in X$, and

$$\|K(\cdot,y)\|_{L^{p_1'}(X)} \leq B_1$$

for almost every $y \in Y$, where $1 \leq p_1, q_0 \leq \infty$ and $B_0, B_1 > 0$. Then for every $0 < \theta < 1$, $|T|$ and T are of strong-type (p_θ, q_θ), with $Tf(y)$ well-defined for all $f \in L^{p_\theta}(X)$ and almost every $y \in Y$, and furthermore

$$\|Tf\|_{L^{q_\theta}(Y)} \leq B_\theta \|f\|_{L^{p_\theta}(X)}.$$

Here we adopt the convention that $p_0 := 1$ and $q_1 := \infty$, thus $q_\theta = q_0/(1-\theta)$ and $p_\theta' = p_1'/\theta$.

Proof. The hypothesis $\|K(x,\cdot)\|_{L^{q_0}(Y)} \leq B_0$, combined with *Minkowski's integral inequality*, shows us that

$$\||T|f\|_{L^{q_0}(Y)} \leq B_0 \|f\|_{L^1(X)}$$

for all $f \in L^1(X)$; in particular, for such f, Tf is well-defined almost everywhere, and

$$\|Tf\|_{L^{q_0}(Y)} \leq B_0 \|f\|_{L^1(X)}.$$

Similarly, Hölder's inequality tells us that for $f \in L^{p_1}(X)$, Tf is well-defined everywhere, and

$$\|Tf\|_{L^\infty(Y)} \leq B_1 \|f\|_{L^{p_1}(X)}.$$

Applying the Riesz-Thorin theorem we conclude that

$$\|Tf\|_{L^{q_\theta}(Y)} \leq B_\theta \|f\|_{L^{p_\theta}(X)}$$

for all simple functions f with finite measure support; replacing K with $|K|$ we also see that

$$\||T|f\|_{L^{q_\theta}(Y)} \leq B_\theta \|f\|_{L^{p_\theta}(X)}$$

for all simple functions f with finite measure support, and thus (by monotone convergence, Theorem 1.1.21) for all $f \in L^{p_\theta}(X)$. The claim then follows. □

Example 1.11.15. Let $A = (a_{ij})_{1\leq i\leq n, 1\leq j\leq m}$ be a matrix such that the sum of the magnitudes of the entries in every row and column is at most B, i.e. $\sum_{i=1}^n |a_{ij}| \leq B$ for all j and $\sum_{j=1}^m |a_{ij}| \leq B$ for all i. Then one has the bound

$$\|Ax\|_{\ell^p_m} \leq B\|x\|_{\ell^p_n}$$

for all vectors $x \in \mathbf{C}^n$ and all $1 \leq p \leq \infty$. Note the extreme cases $p = 1$, $p = \infty$ can be seen directly; the remaining cases then follow from interpolation.

A useful special case arises when A is an *S-sparse* matrix, which means that at most S entries in any row or column are non-zero (e.g. *permutation matrices* are 1-sparse). We then conclude that the ℓ^p operator norm of A is at most $S \sup_{i,j} |a_{i,j}|$.

Exercise 1.11.24. Establish Schur's test by more direct means, taking advantage of the duality relationship

$$\|g\|_{L^p(Y)} := \sup\{|\int_Y gh| : \|h\|_{L^{p'}(Y)} \leq 1\}$$

for $1 \leq p \leq \infty$, as well as *Young's inequality* $xy \leq \frac{1}{r}x^r + \frac{1}{r'}x^{r'}$ for $1 < r < \infty$. (You may wish to first work out Example 1.11.15, say with $p = 2$, to figure out the logic.)

A useful corollary of Schur's test is *Young's convolution inequality* for the convolution $f * g$ of two functions $f : \mathbf{R}^n \to \mathbf{C}$, $g : \mathbf{R}^n \to \mathbf{C}$, defined as

$$f * g(x) := \int_{\mathbf{R}^n} f(y)g(x-y)\, dy$$

provided of course that the integrand is absolutely convergent.

Exercise 1.11.25 (Young's inequality)**.** Let $1 \leq p, q, r \leq \infty$ be such that $\frac{1}{p} + \frac{1}{q} = \frac{1}{r} + 1$. Show that if $f \in L^p(\mathbf{R}^n)$ and $g \in L^q(\mathbf{R}^n)$, then $f * g$ is well-defined almost everywhere and lies in $L^r(\mathbf{R}^n)$, and furthermore that

$$\|f * g\|_{L^r(\mathbf{R}^n)} \leq \|f\|_{L^p(\mathbf{R}^n)}\|g\|_{L^q(\mathbf{R}^n)}.$$

(*Hint*: Apply Schur's test to the kernel $K(x, y) := g(x - y)$.)

Remark 1.11.16. There is nothing special about $\mathbf{R}^n$ here; one could in fact use any locally compact group G with a bi-invariant *Haar measure*. On the other hand, if one specialises to $\mathbf{R}^n$, then it is possible to improve Young's inequality slightly, to

$$\|f * g\|_{L^r(\mathbf{R}^n)} \leq (A_p A_q A_{r'})^{n/2}\|f\|_{L^p(\mathbf{R}^n)}\|g\|_{L^q(\mathbf{R}^n)}.$$

where $A_p := p^{1/p}/(p')^{1/p'}$, a celebrated result of Beckner [**Be1975**]; the constant here is best possible, as can be seen by testing the inequality in the case when f, g are Gaussians.

Exercise 1.11.26. Let $1 \leq p \leq \infty$, and let $f \in L^p(\mathbf{R}^n)$, $g \in L^{p'}(\mathbf{R}^n)$. Young's inequality tells us that $f * g \in L^\infty(\mathbf{R}^n)$. Refine this further by showing that $f * g \in C_0(\mathbf{R}^n)$, i.e. $f * g$ is continuous and goes to zero at infinity. (*Hint*: first show this when $f, g \in C_c(\mathbf{R}^n)$, then use a limiting argument.)

We now give a variant of Schur's test that allows for weak estimates.

Lemma 1.11.17 (Weak-type Schur's test)**.** *Let $K : X \times Y \to \mathbf{C}$ be a measurable function obeying the bounds*

$$\|K(x,\cdot)\|_{L^{q_0,\infty}(Y)} \leq B_0$$

for almost every $x \in X$, and

$$\|K(\cdot,y)\|_{L^{p_1',\infty}(X)} \leq B_1$$

for almost every $y \in Y$, where $1 < p_1, q_0 < \infty$ and $B_0, B_1 > 0$ (note the endpoint exponents $1, \infty$ are now excluded). Then for every $0 < \theta < 1$, $|T|$ and T are of strong-type (p_θ, q_θ), with $Tf(y)$ well-defined for all $f \in L^{p_\theta}(X)$ and almost every $y \in Y$, and furthermore

$$\|Tf\|_{L^{q_\theta}(Y)} \leq C_{p_1,q_0,\theta} B_\theta \|f\|_{L^{p_\theta}(X)}.$$

Here we again adopt the convention that $p_0 := 1$ and $q_1 := \infty$.

Proof. From Exercise 1.11.11 we see that

$$\int_Y |K(x,y)| 1_E(y)\ d\nu(y) \lesssim B_0 \mu(E)^{1/q_0'}$$

for any measurable $E \subset Y$, where we use $A \lesssim B$ to denote $A \leq C_{p_1,q_0,\theta} B$ for some $C_{p_1,q_0,\theta}$ depending on the indicated parameters. By the Fubini-Tonelli theorem, we conclude that

$$\int_Y |T|f(y) 1_E(y)\ d\nu(y) \lesssim B_0 \mu(E)^{1/q_0'} \|f\|_{L^1(X)}$$

for any $f \in L^1(X)$; by Exercise 1.11.11 again we conclude that

$$\||T|f\|_{L^{q_0,\infty}(Y)} \lesssim B_0 \|f\|_{L^1(X)}$$

thus $|T|$ is of weak-type $(1, q_0)$. In a similar vein, from yet another application of Exercise 1.11.11 we see that

$$\||T|f\|_{L^\infty(Y)} \lesssim B_1 \mu(F)^{1/p_1}$$

whenever $0 \leq f \leq 1_F$ and $F \subset X$ has finite measure; thus $|T|$ is of restricted type (p_1, ∞). Applying Exercise 1.11.18 we conclude that $|T|$ is of strong type (p_θ, q_θ) (with operator norm $\lesssim B_\theta$), and the claim follows. □

This leads to a weak-type version of Young's inequality:

Exercise 1.11.27 (Weak-type Young's inequality)**.** Let $1 < p, q, r < \infty$ be such that $\frac{1}{p} + \frac{1}{q} = \frac{1}{r} + 1$. Show that if $f \in L^p(\mathbf{R}^n)$ and $g \in L^{q,\infty}(\mathbf{R}^n)$, then $f * g$ is well-defined almost everywhere and lies in $L^r(\mathbf{R}^n)$, and furthermore that

$$\|f * g\|_{L^r(\mathbf{R}^n)} \leq C_{p,q} \|f\|_{L^p(\mathbf{R}^n)} \|g\|_{L^{q,\infty}(\mathbf{R}^n)}.$$

for some constant $C_{p,q} > 0$.

Exercise 1.11.28. Refine the previous exercise by replacing $L^r(\mathbf{R}^n)$ with the Lorentz space $L^{r,p}(\mathbf{R}^n)$ throughout.

Recall that the function $1/|x|^\alpha$ will lie in $L^{n/\alpha,\infty}(\mathbf{R}^n)$ for $\alpha > 0$. We conclude

Corollary 1.11.18 (Hardy-Littlewood-Sobolev fractional integration inequality)**.** *Let $1 < p, r < \infty$ and $0 < \alpha < n$ be such that $\frac{1}{p} + \frac{\alpha}{n} = \frac{1}{r} + 1$. If $f \in L^p(\mathbf{R}^n)$, then the function $I_\alpha f$, defined as*

$$I_\alpha f(x) := \int_{\mathbf{R}^n} \frac{f(y)}{|x-y|^\alpha}\, dy$$

is well-defined almost everywhere and lies in $L^r(\mathbf{R}^n)$, and furthermore that

$$\|I_\alpha f\|_{L^r(\mathbf{R}^n)} \leq C_{p,\alpha,n} \|f\|_{L^p(\mathbf{R}^n)}$$

for some constant $C_{p,\alpha,n} > 0$.

This inequality is of importance in the theory of Sobolev spaces, which we will discuss in Section 1.14.

Exercise 1.11.29. Show that Corollary 1.11.18 can fail at the endpoints $p = 1$, $r = \infty$, or $\alpha = n$.

Notes. This lecture first appeared at `terrytao.wordpress.com/2009/03/30`. Thanks to PDEbeginner, Samir Chomsky, Spencer, Xiaochuan Liu and anonymous commenters for corrections.

1.12. The Fourier transform

In these notes we lay out the basic theory of the *Fourier transform*, which is of course the most fundamental tool in *harmonic analysis* and also of major importance in related fields (functional analysis,

complex analysis, PDE, number theory, additive combinatorics, representation theory, signal processing, etc.). The Fourier transform, in conjunction with the *Fourier inversion formula*, allows one to take essentially arbitrary (complex-valued) functions on a group G (or more generally, a space X that G acts on, e.g. a *homogeneous space* G/H), and decompose them as a (discrete or continuous) superposition of much more symmetric functions on the domain, such as *characters* $\chi : G \to S^1$; the precise superposition is given by *Fourier coefficients* $\hat{f}(\xi)$, which take values in some dual object such as the *Pontryagin dual* $\hat{G}$ of G. Characters behave in a very simple manner with respect to translation (indeed, they are eigenfunctions of the translation action), and so the Fourier transform tends to simplify any mathematical problem which enjoys a translation invariance symmetry (or an approximation to such a symmetry), and is somehow "linear" (i.e. it interacts nicely with superpositions). In particular, Fourier analytic methods are particularly useful for studying operations such as convolution $f, g \mapsto f * g$ and set-theoretic addition $A, B \mapsto A + B$, or the closely related problem of counting solutions to additive problems such as $x = a_1 + a_2 + a_3$ or $x = a_1 - a_2$, where a_1, a_2, a_3 are constrained to lie in specific sets A_1, A_2, A_3. The Fourier transform is also a particularly powerful tool for solving constant-coefficient linear ODE and PDE (because of the translation invariance), and can also approximately solve some variable-coefficient (or slightly non-linear) equations if the coefficients vary smoothly enough and the nonlinear terms are sufficiently tame.

The Fourier transform $\hat{f}(\xi)$ also provides an important new way of looking at a function $f(x)$, as it highlights the distribution of f in *frequency space* (the domain of the frequency variable ξ) rather than *physical space* (the domain of the physical variable x). A given property of f in the physical domain may be transformed to a rather different-looking property of $\hat{f}$ in the frequency domain. For instance:

- Smoothness of f in the physical domain corresponds to decay of $\hat{f}$ in the Fourier domain, and conversely. (More generally, fine scale properties of f tend to manifest themselves as coarse scale properties of $\hat{f}$, and conversely.)

- Convolution in the physical domain corresponds to pointwise multiplication in the Fourier domain, and conversely.
- Constant coefficient differential operators such as d/dx in the physical domain corresponds to multiplication by polynomials such as $2\pi i\xi$ in the Fourier domain, and conversely.
- More generally, translation invariant operators in the physical domain correspond to multiplication by symbols in the Fourier domain, and conversely.
- Rescaling in the physical domain by an invertible linear transformation corresponds to an inverse (adjoint) rescaling in the Fourier domain.
- Restriction to a subspace (or subgroup) in the physical domain corresponds to projection to the dual quotient space (or quotient group) in the Fourier domain, and conversely.
- Frequency modulation in the physical domain corresponds to translation in the frequency domain, and conversely.

(We will make these statements more precise below.)

On the other hand, some operations in the physical domain remain essentially unchanged in the Fourier domain. Most importantly, the L^2 norm (or *energy*) of a function f is the same as that of its Fourier transform, and more generally the inner product $\langle f, g\rangle$ of two functions f is the same as that of their Fourier transforms. Indeed, the Fourier transform is a unitary operator on L^2 (a fact which is variously known as the *Plancherel theorem* or the *Parseval identity*). This makes it easier to pass back and forth between the physical domain and frequency domain, so that one can combine techniques that are easy to execute in the physical domain with other techniques that are easy to execute in the frequency domain. (In fact, one can combine the physical and frequency domains together into a product domain known as *phase space*, and there are entire fields of mathematics (e.g. microlocal analysis, geometric quantisation, time-frequency analysis) devoted to performing analysis on these sorts of spaces directly, but this is beyond the scope of this course.)

In these notes, we briefly discuss the general theory of the Fourier transform, but will mainly focus on the two classical domains for

Fourier analysis: the torus $\mathbf{T}^d := (\mathbf{R}/\mathbf{Z})^d$, and the Euclidean space $\mathbf{R}^d$. For these domains one has the advantage of being able to perform very explicit algebraic calculations, involving concrete functions such as plane waves $x \mapsto e^{2\pi i x\cdot\xi}$ or Gaussians $x \mapsto A^{d/2}e^{-\pi A|x|^2}$.

1.12.1. Generalities. Let us begin with some generalities. An *abelian topological group* is an abelian group $G = (G, +)$ with a topological structure, such that the group operations of addition $+ : G \times G \to G$ and negation $- : G \to G$ are continuous. (One can of course also consider abelian multiplicative groups $G = (G, \cdot)$, but to fix the notation we shall restrict attention to additive groups.) For technical reasons (and in particular, in order to apply many of the results from the previous sections) it is convenient to restrict attention to abelian topological groups which are *locally compact Hausdorff* (LCH); these are known as *locally compact abelian (LCA) groups*.

Some basic examples of locally compact abelian groups are:

- Finite additive groups (with the discrete topology), such as cyclic groups $\mathbf{Z}/N\mathbf{Z}$.
- Finitely generated additive groups (with the discrete topology), such as the standard lattice $\mathbf{Z}^d$.
- Tori, such as the standard d-dimensional torus $\mathbf{T}^d := (\mathbf{R}/\mathbf{Z})^d$ with the standard topology.
- Euclidean spaces, such the standard d-dimensional Euclidean space $\mathbf{R}^d$ (with the standard topology, of course).
- The rationals $\mathbf{Q}$ are *not* locally compact with the usual topology; but if one uses the discrete topology instead, one recovers an LCA group.
- Another example of an LCA group, of importance in number theory, is the *adele ring* $\mathbb{A}$, discussed in Section 1.5 of *Poincaré's legacies, Vol. I.*

Thus we see that locally compact abelian groups can be either discrete or continuous, and either compact or non-compact; all four combinations of these cases are of importance. The topology of course generates a Borel σ-algebra in the usual fashion, as well as a space $C_c(G)$ of continuous, compactly supported complex-valued functions.

There is a translation action $x \mapsto \tau_x$ of G on $C_c(G)$, where for every $x \in G$, $\tau_x : C_c(G) \to C_c(G)$ is the translation operation

$$\tau_x f(y) := f(y - x).$$

LCA groups need not be σ-compact (think of the free abelian group on uncountably many generators, with the discrete topology), but one has the following useful substitute:

Exercise 1.12.1. Show that every LCA group G contains a σ-compact open subgroup H, and in particular is the disjoint union of σ-compact sets. (*Hint*: Take a compact symmetric neighbourhood K of the identity, and consider the group H generated by this neighbourhood.)

An important notion for us will be that of a *Haar measure*: a *Radon measure* μ on G which is *translation-invariant* (i.e. $\mu(E+x) = \mu(E)$ for all Borel sets $E \subset G$ and all $x \in G$, where $E + x := \{y + x : y \in E\}$ is the translation of E by x). From this and the definition of integration we see that integration $f \mapsto \int_G f \, d\mu$ against a Haar measure (an operation known as the *Haar integral*) is also translation-invariant, thus

$$\int_G f(y - x) \, d\mu(y) = \int_G f(y) \, d\mu(y) \tag{1.98}$$

or equivalently

$$\int_G \tau_x f \, d\mu = \int_G f \, d\mu \tag{1.99}$$

for all $f \in C_c(G)$ and $x \in G$. The trivial measure 0 is of course a Haar measure; all other Haar measures are called *non-trivial*.

Let us note some non-trivial Haar measures in the four basic examples of locally compact abelian groups:

- For a finite additive group G, one can take either *counting measure* $\#$ or normalised counting measure $\#/\#(G)$ as a Haar measure. (The former measure emphasises the discrete nature of G; the latter measure emphasises the compact nature of G.)
- For finitely generated additive groups such as $\mathbf{Z}^d$, counting measure $\#$ is a Haar measure.

- For the standard torus $(\mathbf{R}/\mathbf{Z})^d$, one can obtain a Haar measure by identifying this torus with $[0,1)^d$ in the usual manner and then taking Lebesgue measure on the latter space. This Haar measure is a probability measure.
- For the standard Euclidean space $\mathbf{R}^d$, Lebesgue measure is a Haar measure.

Of course, any non-negative constant multiple of a Haar measure is again a Haar measure. The converse is also true:

Exercise 1.12.2 (Uniqueness of Haar measure up to scalars)**.** Let μ, ν be two non-trivial Haar measures on a locally compact abelian group G. Show that μ, ν are scalar multiples of each other, i.e. there exists a constant $c > 0$ such that $\nu = c\mu$. (*Hint*: for any $f, g \in C_c(G)$, compute the quantity $\int_G \int_G g(y) f(x+y)\ d\mu(x) d\nu(y)$ in two different ways.)

The above argument also implies a useful symmetry property of Haar measures:

Exercise 1.12.3 (Haar measures are symmetric)**.** Let μ be a Haar measure on a locally compact abelian group G. Show that $\int_G f(-x)\ dx = \int_G f(x)\ dx$ for all $f \in C_c(G)$. (*Hint*: expand $\int_G \int_G f(y) f(x+y)\ d\mu(x) d\mu(y)$ in two different ways.) Conclude that Haar measures on LCA groups are *symmetric* in the sense that $\mu(-E) = \mu(E)$ for all measurable E, where $-E := \{-x : x \in E\}$ is the reflection of E.

Exercise 1.12.4 (Open sets have positive measure)**.** Let μ be a non-trivial Haar measure on a locally compact abelian group G. Show that $\mu(U) > 0$ for any non-empty open set U. Conclude that if $f \in C_c(G)$ is non-negative and not identically zero, then $\int_G f\ d\mu > 0$.

Exercise 1.12.5. If G is an LCA group with non-trivial Haar measure μ, show that $L^1(G)^*$ is identifiable with $L^\infty(G)$. (Unfortunately, G is not always σ-finite, and so the standard duality theorem from Section 1.3 does not directly apply. However, one can get around this using Exercise 1.12.1.)

It is a (not entirely trivial) theorem, due to André Weil, that all LCA groups have a non-trivial Haar measure. For discrete groups, one

can of course take counting measure as a Haar measure. For compact groups, the result is due to Haar, and one can argue as follows:

Exercise 1.12.6 (Existence of Haar measure, compact case)**.** Let G be a compact metrisable abelian group. For any real-valued $f \in C_c(G)$, and any Borel probability measure μ on G, define the *oscillation* $\mathrm{osc}_f(\mu)$ of μ with respect to f to be the quantity $\mathrm{osc}_f(\mu) := \sup_{y \in G} \int_G \tau_y f \, d\mu(x) - \inf_{y \in G} \int_G \tau_y f \, d\mu(x)$.

- (a) Show that a Borel probability measure μ is a Haar measure if and only if $\mathrm{osc}_f(\mu) = 0$ for all $f \in C_c(G)$.
- (b) If a sequence μ_n of Borel probability measures converges in the vague topology to another Borel probability measure μ, show that $\mathrm{osc}_f(\mu_n) \to \mathrm{osc}_f(\mu)$ for all $f \in C_c(G)$.
- (c) If μ is a Borel probability measure and $f \in C_c(G)$ is such that $\mathrm{osc}_f(\mu) > 0$, show that there exists a Borel probability measure μ' such that $\mathrm{osc}_f(\mu') < \mathrm{osc}_f(\mu)$ and $\mathrm{osc}_g(\mu') \leq \mathrm{osc}_g(\mu)$ for all $g \in C_c(G)$. (*Hint*: take μ' to be the an average of certain translations of μ.)
- (d) Given any finite number of functions $f_1, \ldots, f_n \in C_c(G)$, show that there exists a Borel probability measure μ such that $\mathrm{osc}_{f_i}(\mu) = 0$ for all $i = 1, \ldots, n$. (*Hint*: Use Prokhorov's theorem, see Corollary 1.10.22. Try the $n = 1$ case first.)
- (e) Show that there exists a unique Haar probability measure on G. (*Hint*: One can identify each probability measure μ with the element $(\int_G f \, d\mu)_{f \in C_c(G)}$ of the product space $\prod_{f \in C_c(G)} [-\sup_{x \in G} |f(x)|, \sup_{x \in G} |f(x)|]$, which is compact by Tychonoff's theorem. Now use (d) and the *finite intersection property*.)

(The argument can be adapted to the case when G is not metrisable, but one has to replace the sequential compactness given by Prokhorov's theorem with the topological compactness given by the Banach-Alaoglu theorem.)

For general LCA groups, the proof is more complicated:

Exercise 1.12.7 (Existence of Haar measure, general case)**.** Let G be an LCA group. Let $C_c(G)^+$ denote the space of non-negative

functions $f \in C_c(G)$ that are not identically zero. Given two $f, g \in C_c(G)^+$, define a *g-cover* of f to be an expression of the form $a_1\tau_{x_1}g + \ldots + a_n\tau_{x_n}g$ that pointwise dominates f, where $a_1, \ldots, a_n$ are non-negative numbers and $x_1, \ldots, x_n \in G$. Let $(f : g)$ denote the infimum of the quantity $a_1 + \ldots + a_n$ for all g-covers of f.

- (a) (Finiteness) Show that $0 < (f : g) < +\infty$ for all $f, g \in C_c(G)^+$.
- (b) Let μ is a Haar measure on G. Show that $\int_G f \; d\mu \leq (f : g)(\int_G g \; d\mu)$ for all $f, g \in C_c(G)^+$. Conversely, for every $f \in C_c(G)^+$ and $\varepsilon > 0$, show that there exists $g \in C_c(G)^+$ such that $\int_G f \; d\mu \geq (f : g)(\int_G g \; d\mu) - \varepsilon$. (*Hint*: f is uniformly continuous. Take g to be an approximation to the identity.) Thus Haar integrals are related to certain renormalised versions of the functionals $f \mapsto (f : g)$; this observation underlies the strategy for construction of Haar measure in the rest of this exercise.
- (c) (Transitivity) Show that $(f : h) \leq (f : g)(g : h)$ for all $f, g, h \in C_c(G)^+$.
- (d) (Translation invariance) Show that $(\tau_x f : g) = (f : g)$ for all $f, g \in C_c(G)^+$ and $x \in G$.
- (e) (Sublinearity) Show that $(f + g : h) \leq (f : h) + (g : h)$ and $(cf : g) = c(f : g)$ for all $f, g, h \in C_c(G)^+$ and $c > 0$.
- (f) (Approximate superadditivity) If $f, g \in C_c(G)^+$ and $\varepsilon > 0$, show that there exists a neighbourhood U of the identity such that $(f : h) + (g : h) \leq (1 + \varepsilon)(f + g : h)$ whenever $h \in C_c(G)^+$ is supported in U. (*Hint*: $f, g, f+g$ are all uniformly continuous. Take a h-cover of $f+g$ and multiply the weight a_i at x_i by weights such as $f(x_i)/(f(x_i) + g(x_i) - \varepsilon)$ and $g(x_i)/(f(x_i) + g(x_i) - \varepsilon)$.)

Next, fix a reference function $f_0 \in C_c(G)^+$, and define the functional $I_g : C_c(G)^+ \to \mathbf{R}^+$ for all $g \in C_c(G)^+$ by the formula $I_g(f) := (f : g)/(f_0 : g)$.

- (g) Show that for any fixed f, $I_g(f)$ ranges in the compact interval $[(f_0 : f)^{-1}, (f : f_0)]$; thus I_g can be viewed as an

element of the product space $\prod_{f \in C_c(G)^+} [(f_0 : f)^{-1}, (f : f_0)]$, which is compact by *Tychonoff's theorem*.

- (h) From (d), (e) we have the translation-invariance property $I_g(\tau_x f) = I_g(f)$, the homogeneity property $I_g(cf) = cI_g(f)$, and the sub-additivity property $I_g(f+f') \leq I_g(f) + I_g(f')$ for all $g, f, f' \in C_c(G)^+$, $x \in G$, and $c > 0$; we also have the normalisation $I_g(f_0) = 1$. Now show that for all $f_1, \ldots, f_n, f'_1, \ldots, f'_n \in C_c(G)^+$ and $\varepsilon > 0$, there exists $g \in C_c(G)^+$ such that $I_g(f_i + f'_i) \geq I_g(f_i) + I_g(f'_i) - \varepsilon$ for all $i = 1, \ldots, n$.
- (i) Show that there exists a unique Haar measure μ on G with $\mu(f_0) = 1$. (*Hint*: Use (h) and the *finite intersection property* to obtain a translation-invariant positive linear functional on $C_c(G)$, then use the *Riesz representation theorem*.)

Now we come to a fundamental notion, that of a *character*.

Definition 1.12.1 (Characters)**.** Let G be a LCA group. A *multiplicative character* χ is a continuous function $\chi : G \to S^1$ to the unit circle $S^1 := \{z \in \mathbf{C} : |z| = 1\}$ which is a homomorphism, i.e. $\chi(x+y) = \chi(x)\chi(y)$ for all $x, y \in G$. An *additive character* or *frequency* $\xi : x \mapsto \xi \cdot x$ is a continuous function $\xi : G \to \mathbf{R}/\mathbf{Z}$ which is a homomorphism, thus $\xi \cdot (x+y) = \xi \cdot x + \xi \cdot y$ for all $x, y \in G$. The set of all frequencies ξ is called the *Pontryagin dual* of G and is denoted $\hat{G}$; it is clearly an abelian group. A multiplicative character is called *non-trivial* if it is not the constant function 1; an additive character is called *non-trivial* if it is not the constant function 0.

Multiplicative characters and additive characters are clearly related: if $\xi \in \hat{G}$ is an additive character, then the function $x \mapsto e^{2\pi i \xi \cdot x}$ is a multiplicative character, and conversely every multiplicative character arises uniquely from an additive character in this fashion.

Exercise 1.12.8. Let G be an LCA group. We give $\hat{G}$ the topology of local uniform convergence on compact sets, thus the topology on $\hat{G}$ are generated by sets of the form $\{\xi \in \hat{G} : |\xi \cdot x - \xi_0 \cdot x| < \varepsilon \text{ for all } x \in K\}$ for compact $K \subset G$, $\xi_0 \in \hat{G}$, and $\varepsilon > 0$. Show that this turns $\hat{G}$

into an LCA group. (*Hint*: Show that for any neighbourhood U of the identity in G, the sets $\{\xi \in \hat{G} : \xi \cdot x \in [-\varepsilon, \varepsilon] \text{ for all } x \in U\}$ for $0 < \varepsilon < 1/4$ (say) are compact.) Furthermore, if G is discrete, show that $\hat{G}$ is compact.

The Pontryagin dual can be computed easily for various classical LCA groups:

Exercise 1.12.9. Let $d \geq 1$ be an integer.

(a) Show that the Pontryagin dual $\widehat{\mathbf{Z}^d}$ of $\mathbf{Z}^d$ is identifiable as an LCA group with $(\mathbf{R}/\mathbf{Z})^d$, by identifying each $\xi \in (\mathbf{R}/\mathbf{Z})^d$ with the frequency $x \mapsto \xi \cdot x$ given by the dot product.

(b) Show that the Pontryagin dual $\widehat{\mathbf{R}^d}$ of $\mathbf{R}^d$ is identifiable as an LCA group with $\mathbf{R}^d$, by identifying each $\xi \in \mathbf{R}^d$ with the frequency $x \mapsto \xi \cdot x$ given by the dot product.

(c) Show that the Pontryagin dual $\widehat{(\mathbf{R}/\mathbf{Z})^d}$ of $(\mathbf{R}/\mathbf{Z})^d$ is identifiable as an LCA group with $\mathbf{Z}^d$, by identifying each $\xi \in \mathbf{Z}^d$ with the frequency $x \mapsto \xi \cdot x$ given by the dot product.

(d) (Contravariant functoriality) If $\phi : G \to H$ is a continuous homomorphism between LCA groups, show that there is a continuous homomorphism $\phi^* : \hat{H} \to \hat{G}$ between their Pontryagin duals, defined by $\phi^*(\xi) \cdot x := \xi \cdot \phi(x)$ for $\xi \in \hat{H}$ and $x \in G$.

(e) If H is a closed subgroup of an LCA group G (and is thus also LCA), show that $\hat{H}$ is identifiable with $\hat{G}/H^\perp$, where $H^\perp$ is the space of all frequencies $\xi \in \hat{G}$ which annihilate H (i.e. $\xi \cdot x = 0$ for all $x \in H$).

(f) If G, H are LCA groups, show that $\widehat{G \times H}$ is identifiable as an LCA group with $\hat{G} \times \hat{H}$.

(g) Show that the Pontryagin dual of a finite abelian group G is identifiable with itself. (*Hint*: first do this for cyclic groups $\mathbf{Z}/N\mathbf{Z}$, identifying $\xi \in \mathbf{Z}/N\mathbf{Z}$ with the additive character $x \mapsto x\xi/N$), then use the *classification of finite abelian groups*.) Note that this identification is not unique.

Exercise 1.12.10. Let G be an LCA group with non-trivial Haar measure μ, and let $\chi : G \to S^1$ be a measurable function such that

$\chi(x)\chi(y) = \chi(x+y)$ for almost every $x, y \in G$. Show that χ is equal almost everywhere to a multiplicative character $\tilde{\chi}$ of G. (*Hint*: on the one hand, $\tau_x\chi = \chi(-x)\chi$ a.e. for almost every x. On the other hand, $\tau_x\chi$ depends continuously on x in, say, the local L^1 topology.)

In the remainder of this section, G is a fixed LCA group with a non-trivial Haar measure μ.

Given an absolutely integrable function $f \in L^1(G)$, we define the *Fourier transform* $\hat{f} : \hat{G} \to \mathbf{C}$ by the formula

$$\hat{f}(\xi) := \int_G f(x) e^{-2\pi i \xi \cdot x}\, d\mu(x).$$

This is clearly a linear transformation, with the obvious bound

$$\sup_{\xi \in \hat{G}} |\hat{f}(\xi)| \leq \|f\|_{L^1(G)}.$$

It converts translations into frequency modulations: indeed, one easily verifies that

$$\widehat{\tau_{x_0} f}(\xi) = e^{-2\pi i \xi \cdot x_0} \hat{f}(\xi) \tag{1.100}$$

for any $f \in L^1(G)$, $x_0 \in G$, and $\xi \in \hat{G}$. Conversely, it converts frequency modulations to translations: one has

$$\widehat{\chi_{\xi_0} f}(\xi) = \hat{f}(\xi - \xi_0) \tag{1.101}$$

for any $f \in L^1(G)$ and $\xi_0, \xi \in \hat{G}$, where χ_{ξ_0} is the multiplicative character $\chi_{\xi_0} : x \mapsto e^{2\pi i \xi_0 \cdot x}$.

Exercise 1.12.11 (Riemann-Lebesgue lemma)**.** If $f \in L^1(G)$, show that $\hat{f} : \hat{G} \to \mathbf{C}$ is continuous. Furthermore, show that $\hat{f}$ goes to zero at infinity in the sense that for every $\varepsilon > 0$ there exists a compact subset K of $\hat{G}$ such that $|\hat{f}(\xi)| \leq \varepsilon$ for $\xi \notin K$. (*Hint*: First show that there exists a neighbourhood U of the identity in G such that $\|\tau_x f - f\|_{L^1(G)} \leq \varepsilon^2$ (say) for all $x \in U$. Now take the Fourier transform of this fact.) Thus the Fourier transform maps $L^1(G)$ continuously to $C_0(\hat{G})$, the space of continuous functions on $\hat{G}$ which go to zero at infinity; the decay at infinity is known as the *Riemann-Lebesgue lemma*.

Exercise 1.12.12. Let G be an LCA group with non-trivial Haar measure μ. Show that the topology of $\hat{G}$ is the weakest topology such that $\hat{f}$ is continuous for every $f \in L^1(G)$.

Given two $f, g \in L^1(G)$, recall that the *convolution* $f * g : G \to \mathbf{C}$ is defined as

$$f * g(x) := \int_G f(y)g(x-y)\, d\mu(y).$$

From *Young's inequality* (Exercise 1.11.25) we know that $f * g$ is defined a.e., and lies in $L^1(G)$; indeed, we have

$$\|f * g\|_{L^1(G)} \leq \|f\|_{L^1(G)}\|g\|_{L^1(G)}.$$

Exercise 1.12.13. Show that the operation $f, g \mapsto f * g$ is a bilinear, continuous, commutative, and associative operation on $L^1(G)$. As a consequence, the Banach space $L^1(G)$ with the convolution operation as "multiplication" operation becomes a *commutative Banach algebra*. If we also define $f^*(x) := \overline{f(-x)}$ for all $f \in L^1(G)$, this turns $L^1(G)$ into a *B^*-algebra*Banach $*$-algebra.

For $f, g \in L^1(G)$, show that

$$\widehat{f * g}(\xi) = \hat{f}(\xi)\hat{g}(\xi) \tag{1.102}$$

for all $\xi \in \hat{G}$; thus the Fourier transform converts convolution to pointwise product.

Exercise 1.12.14. Let G, H be LCA groups with non-trivial Haar measures μ, ν respectively, and let $f \in L^1(G)$, $g \in L^1(H)$. Show that the tensor product $f \otimes g \in L^1(G \times H)$ (with product Haar measure $\mu \times \nu$) has a Fourier transform of $\hat{f} \otimes \hat{g}$, where we identify $\widehat{G \times H}$ with $\hat{G} \times \hat{H}$ as per Exercise 1.12.9(f). Informally, this exercise asserts that the Fourier transform commutes with tensor products. (Because of this fact, the tensor power trick (Section 1.9 of *Structure and Randomness*) is often available when proving results about the Fourier transform on general groups.)

Exercise 1.12.15 (Convolution and Fourier transform of measures)**.** If $\nu \in M(G)$ is a finite Radon measure on an LCA group G with non-trivial Haar measure μ, define the *Fourier-Stieltjes transform* $\hat{\nu} : \hat{G} \to \mathbf{C}$ by the formula $\hat{\nu}(\xi) := \int_G e^{-2\pi i \xi \cdot x}\, d\nu(x)$ (thus for instance $\hat{\mu_f} = \hat{f}$ for any $f \in L^1(G)$). Show that $\hat{\nu}$ is a bounded continuous function

on $\hat{G}$. Given any $f \in L^1(G)$, define the convolution $f * \nu : G \to \mathbf{C}$ to be the function

$$f * \nu(x) := \int_G f(x-y)\ d\nu(y)$$

and given any finite Radon measure ρ, let $\nu * \rho : G \to \mathbf{C}$ be the measure

$$\nu * \rho(E) := \int_G \int_G 1_E(x+y)\ d\nu(x) d\rho(y).$$

Show that $f * \nu \in L^1(G)$ and $\widehat{f * \nu}(\xi) = \hat{f}(\xi)\hat{\nu}(\xi)$ for all $\xi \in \hat{G}$, and similarly that $\nu * \rho$ is a finite measure and $\widehat{\nu * \rho}(\xi) = \hat{\nu}(\xi)\hat{\rho}(\xi)$ for all $\xi \in \hat{G}$. Thus the convolution and Fourier structure on $L^1(G)$ can be extended to the larger space $M(G)$ of finite Radon measures.

1.12.2. The Fourier transform on compact abelian groups. In this section we specialise the Fourier transform to the case when the locally compact group G is in fact compact, thus we now have a compact abelian group G with non-trivial Haar measure μ. This case includes that of finite groups, together with that of the tori $(\mathbf{R}/\mathbf{Z})^d$.

As μ is a Radon measure, compact groups G have finite measure. It is then convenient to normalise the Haar measure μ so that $\mu(G) = 1$, thus μ is now a probability measure. For the remainder of this section, we will assume that G is a compact abelian group and μ is its (unique) Haar probability measure, as given by Exercise 1.12.6.

A key advantage of working in the compact setting is that multiplicative characters $\chi : G \to S^1$ now lie in $L^2(G)$ and $L^1(G)$. In particular, they can be integrated:

Lemma 1.12.2. *Let χ be a multiplicative character. Then $\int_G \chi\ d\mu$ equals 1 when χ is trivial and 0 when χ is non-trivial. Equivalently, for $\xi \in \hat{G}$, we have $\int_G e^{2\pi i \xi \cdot x}\ d\mu = \delta_0(\xi)$, where δ is the* Kronecker delta function *at 0.*

Proof. The claim is clear when χ is trivial. When χ is non-trivial, there exists $x \in G$ such that $\chi(x) \neq 1$. If one then integrates the identity $\tau_x \chi = \chi(-x)\chi$ using (1.99) one obtains the claim. □

Exercise 1.12.16. Show that the Pontryagin dual $\hat{G}$ of a compact abelian group G is discrete (compare with Exercise 1.12.8).

Exercise 1.12.17. Show that the Fourier transform of the constant function 1 is the Kronecker delta function δ_0 at 0. More generally, for any $\xi_0 \in \hat{G}$, show that the Fourier transform of the multiplicative character $x \mapsto e^{2\pi i \xi_0 \cdot x}$ is the Kronecker delta function δ_{ξ_0} at ξ_0.

Since the pointwise product of two multiplicative characters is again a multiplicative character, and the conjugate of a multiplicative character is also a multiplicative character, we obtain

Corollary 1.12.3. *The space of multiplicative chararacters is an orthonormal set in the complex Hilbert space $L^2(G)$.*

Actually, one can say more:

Theorem 1.12.4 (Plancherel theorem for compact abelian groups)**.** *Let G be a compact abelian group with probability Haar measure μ. Then the space of multiplicative characters is an orthonormal* basis *for the complex Hilbert space $L^2(G)$.*

The full proof of this theorem requires the *spectral theorem* and is not given here, though see Exercise 1.12.43 below. However, we can work out some important special cases here.

- When G is a torus $G = \mathbf{T}^d = (\mathbf{R}/\mathbf{Z})^d$, the multiplicative characters $x \mapsto e^{2\pi i \xi \cdot x}$ separate points (given any two $x, y \in G$, there exists a character which takes different values at x and at y). The space of finite linear combinations of multiplicative characters (i.e. the space of *trigonometric polynomials*) is then an algebra closed under conjugation that separates points and contains the unit 1, and thus by the *Stone-Weierstrass theorem*, is dense in $C(G)$ in the uniform (and hence in L^2) topology, and is thus dense in $L^2(G)$ (in the L^2 topology) also.
- The same argument works when G is a cyclic group $\mathbf{Z}/N\mathbf{Z}$, using the multiplicative characters $x \mapsto e^{2\pi i \xi x/N}$ for $\xi \in \mathbf{Z}/N\mathbf{Z}$. As every finite abelian group is isomorphic to the product of cyclic groups, we also obtain the claim for finite abelian groups.
- Alternatively, when G is finite, one can argue by viewing the linear operators $\tau_x : C_c(G) \to C_c(G)$ as $|G| \times |G|$ *unitary*

matrices (in fact, they are *permutation matrices*) for each $x \in G$. The spectral theorem for unitary matrices allows each of these matrices to be diagonalised; as G is abelian, the matrices commute and so one can *simultaneously* diagonalise these matrices. It is not hard to see that each simultaneous eigenvector of these matrices is a multiple of a character, and so the characters span $L^2(G)$, yielding the claim. (The same argument will in fact work for arbitrary compact abelian groups, once we obtain the spectral theorem for unitary operators.)

If $f \in L^2(G)$, the inner product $\langle f, \chi_\xi \rangle_{L^2(G)}$ of f with any multiplicative character $\chi_\xi : x \mapsto e^{2\pi i \xi \cdot x}$ is just the Fourier coefficient $\hat{f}(\xi)$ of f at the corresponding frequency. Applying the general theory of orthonormal bases (see Section 1.4), we obtain the following consequences:

Corollary 1.12.5 (Plancherel theorem for compact abelian groups, again)**.** *Let G be a compact abelian group with probability Haar measure μ.*

- *(Parseval identity) For any $f \in L^2(G)$, we have $\|f\|_{L^2(G)}^2 = \sum_{\xi \in \hat{G}} |\hat{f}(\xi)|^2$.*
- *(Parseval identity, II) For any $f, g \in L^2(G)$, we have $\langle f, g \rangle_{L^2(G)} = \sum_{\xi \in \hat{G}} \hat{f}(\xi)\overline{\hat{g}(\xi)}$.*
- *(Unitarity) Thus the Fourier transform is a* unitary transformation *from $L^2(G)$ to $\ell^2(\hat{G})$.*
- *(Inversion formula) For any $f \in L^2(G)$, the series $x \mapsto \sum_{\xi \in \hat{G}} \hat{f}(\xi) e^{2\pi i \xi \cdot x}$ converges unconditionally in $L^2(G)$ to f.*
- *(Inversion formula, II) For any sequence $(c_\xi)_{\xi \in \hat{G}}$ in $\ell^2(\hat{G})$, the series $x \mapsto \sum_{\xi \in \hat{G}} c_\xi e^{2\pi i \xi \cdot x}$ converges unconditionally in $L^2(G)$ to a function f with c_ξ as its Fourier coefficients.*

We can record here a textbook application of the *Riesz-Thorin interpolation theorem* from Section 1.11. Observe that the Fourier transform map $\mathcal{F} : f \mapsto \hat{f}$ maps $L^2(G)$ to $\ell^2(\hat{G})$ with norm 1, and

also trivially maps $L^1(G)$ to $\ell^\infty(\hat{G})$ with norm 1. Applying the interpolation theorem, we conclude the *Hausdorff-Young inequality*

$$\|\hat{f}\|_{\ell^{p'}(\hat{G})} \leq \|f\|_{L^p(G)} \tag{1.103}$$

for all $1 \leq p \leq 2$ and all $f \in L^p(G)$; in particular, the Fourier transform maps $L^p(G)$ to $\ell^{p'}(\hat{G})$, where p' is the dual exponent of p, thus $1/p + 1/p' = 1$. It is remarkably difficult (though not impossible) to establish the inequality (1.103) without the aid of the Riesz-Thorin theorem. (For instance, one could use the Marcinkiewicz interpolation theorem combined with the tensor power trick.) The constant 1 cannot be improved, as can be seen by testing (1.103) with the function $f = 1$ and using Exercise 1.12.17. By combining (1.103) with Hölder's inequality, one concludes that

$$\|\hat{f}\|_{\ell^q(\hat{G})} \leq \|f\|_{L^p(G)} \tag{1.104}$$

whenever $2 \leq q \leq \infty$ and $\frac{1}{p} + \frac{1}{q} \leq 1$. These are the optimal hypotheses on p, q for which (1.104) holds, though we will not establish this fact here.

Exercise 1.12.18. If $f, g \in L^2(G)$, show that the Fourier transform of $fg \in L^1(G)$ is given by the formula

$$\widehat{fg}(\xi) = \sum_{\eta \in \hat{G}} \hat{f}(\eta)\hat{g}(\xi - \eta).$$

Thus multiplication is converted via the Fourier transform to convolution; compare this with (1.102).

Exercise 1.12.19 (Hardy-Littlewood majorant property)**.** Let $p \geq 2$ be an even integer. If $f, g \in L^p(G)$ are such that $|\hat{f}(\xi)| \leq \hat{g}(\xi)$ for all $\xi \in \hat{G}$ (in particular, $\hat{g}$ is non-negative), show that $\|f\|_{L^p(G)} \leq \|g\|_{L^p(G)}$. (*Hint*: use Exercise 1.12.18 and the Plancherel identity.) The claim fails for all other values of p, a result of Fournier[**Fo1974**].

Exercise 1.12.20. In this exercise and the next two, we will work on the torus $\mathbf{T} = \mathbf{R}/\mathbf{Z}$ with the probability Haar measure μ. The Pontryagin dual $\hat{\mathbf{T}}$ is identified with $\mathbf{Z}$ in the usual manner, thus $\hat{f}(n) = \int_{\mathbf{R}/\mathbf{Z}} f(x)e^{-2\pi inx}\,dx$ for all $f \in L^1(\mathbf{T})$. For every integer $N > 0$ and $f \in L^1(\mathbf{T})$, define the *partial Fourier series* $S_N f$ to be

the expression

$$S_N f(x) := \sum_{n=-N}^{N} \hat{f}(n) e^{2\pi i n x}.$$

- Show that $S_N f = f * D_N$, where D_N is the *Dirichlet kernel* $D_N(x) := \frac{\sin((N+1/2)x)}{\sin x/2}$.
- Show that $\|D_N\|_{L^1(\mathbf{T})} \geq c \log N$ for some absolute constant $c > 0$. Conclude that the operator norm of S_N on $C(\mathbf{T})$ (with the uniform norm) is at least $c \log N$.
- Conclude that there exists a continuous function f such that the partial Fourier series $S_N f$ does not converge uniformly. (*Hint*: use the uniform boundedness principle.) This is despite the fact that $S_N f$ must converge to f in L^2 norm, by the Plancherel theorem. (Another example of non-uniform convergence of $S_N f$ is given by the *Gibbs phenomenon*.)

Exercise 1.12.21. We continue the notational conventions of the preceding exercise. For every integer $N > 0$ and $f \in L^1(\mathbf{T})$, define the *Césaro-summed partial Fourier series* $C_N f$ to be the expression

$$C_N f(x) := \frac{1}{N} \sum_{n=0}^{N-1} D_n f(x).$$

- Show that $C_N f = f * F_N$, where F_N is the *Fejér kernel* $F_N(x) := \frac{1}{n}(\frac{\sin(nx/2)}{\sin(x/2)})^2$.
- Show that $\|F_N\|_{L^1(\mathbf{T})} = 1$. (*Hint*: what is the Fourier coefficient of F_N at zero?)
- Show that $C_N f$ converges uniformly to f for every $f \in C(\mathbf{T})$. (Thus we see that Césaro averaging improves the convergence properties of Fourier series.)

Exercise 1.12.22. *Carleson's inequality* asserts that for any $f \in L^2(\mathbf{T})$, one has the weak-type inequality

$$\| \sup_{N>0} |D_N f(x)| \|_{L^{2,\infty}(\mathbf{T})} \leq C \|f\|_{L^2(\mathbf{T})}$$

for some absolute constant C. Assuming this (deep) inequality, establish *Carleson's theorem* that for any $f \in L^2(\mathbf{T})$, the partial Fourier series $D_N f(x)$ converge for almost every x to $f(x)$. (Conversely, a

general principle of Stein[**St1961**], analogous to the uniform boundedness principle, allows one to deduce Carleson's inequality from Carleson's theorem. A later result of Hunt[**Hu1968**] extends Carleson's theorem to $L^p(\mathbf{T})$ for any $p > 1$, but a famous example of Kolmogorov shows that almost everywhere convergence can fail for $L^1(\mathbf{T})$ functions; in fact the series may diverge pointwise everywhere.)

1.12.3. The Fourier transform on Euclidean spaces. We now turn to the Fourier transform on the Euclidean space $\mathbf{R}^d$, where $d \geq 1$ is a fixed integer. From Exercise 1.12.9 we can identify the Pontryagin dual of $\mathbf{R}^d$ with itself, and then the Fourier transform $\hat{f} : \mathbf{R}^d \to \mathbf{C}$ of a function $f \in L^1(\mathbf{R}^d)$ is given by the formula

$$\hat{f}(\xi) := \int_{\mathbf{R}^d} f(x) e^{-2\pi i \xi \cdot x}\, dx. \tag{1.105}$$

Remark 1.12.6. One needs the Euclidean inner product structure on $\mathbf{R}^d$ in order to identify $\hat{\mathbf{R}^d}$ with $\mathbf{R}^d$. Without this structure, it is more natural to identify $\hat{\mathbf{R}^d}$ with the dual space $(\mathbf{R}^d)^*$ of $\mathbf{R}^d$. (In the language of physics, one should interpret frequency as a *covector* rather than a vector.) However, we will not need to consider such subtleties here. In other areas of mathematics than harmonic analysis, the normalisation of the Fourier transform (particularly with regard to the positioning of the sign $-$ and the factor 2π) is sometimes slightly different from that presented here. For instance, in PDE, the factor of 2π is often omitted from the exponent in order to slightly simplify the behaviour of differential operators under the Fourier transform (at the cost of introducing factors of 2π in various identities, such as the Plancherel formula or inversion formula).

In Exercise 1.12.11 we saw that if f was in $L^1(\mathbf{R}^d)$, then $\hat{f}$ was continuous and decayed to zero at infinity. One can improve both the regularity and decay on $\hat{f}$ by strengthening the hypotheses on f. We need two basic facts:

Exercise 1.12.23 (Decay transforms to regularity)**.** Let $1 \leq j \leq d$, and suppose that $f, x_j f$ both lie in $L^1(\mathbf{R}^d)$, where x_j is the j^{th} coordinate function. Show that $\hat{f}$ is continuously differentiable in the

ξ_j variable, with

$$\frac{\partial}{\partial \xi_j} \hat{f}(\xi) = -2\pi i \widehat{x_j f}(\xi).$$

(*Hint*: The main difficulty is to justify differentiation under the integral sign. Use the fact that the function $x \mapsto e^{ix}$ has a derivative of magnitude 1, and is hence Lipschitz by the fundamental theorem of calculus. Alternatively, one can show first that $\hat{f}(\xi)$ is the indefinite integral of $-2\pi i \widehat{x_j f}$ and then use the fundamental theorem of calculus.)

Exercise 1.12.24 (Regularity transforms to decay)**.** Let $1 \leq j \leq d$, and suppose that $f \in L^1(\mathbf{R}^d)$ has a derivative $\frac{\partial f}{\partial x_j}$ in $L^1(\mathbf{R}^d)$, for which one has the fundamental theorem of calculus

$$f(x_1, \ldots, x_n) = \int_{-\infty}^{x_j} \frac{\partial f}{\partial x_j}(x_1, \ldots, x_{j-1}, t, x_{j+1}, \ldots, x_n)\ dt$$

for almost every $x_1, \ldots, x_n$. (This is equivalent to f being absolutely continuous in x_j for almost every $x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n$.) Show that

$$\widehat{\frac{\partial f}{\partial x_j}}(\xi) = 2\pi i \xi_j \hat{f}(\xi).$$

In particular, conclude that $|\xi_j| \hat{f}(\xi)$ goes to zero as $|\xi| \to \infty$.

Remark 1.12.7. Exercise 1.12.24 shows that *Fourier transforms diagonalise differentiation*: (constant-coefficient) differential operators such as $\frac{\partial}{\partial x_j}$, when viewed in frequency space, become nothing more than multiplication operators $\hat{f}(\xi) \mapsto 2\pi i \xi_j \hat{f}(\xi)$. (Multiplication operators are the continuous analogue of *diagonal matrices*.) It is because of this fact that the Fourier transform is extremely useful in PDE, particularly in constant-coefficient linear PDE, or perturbations thereof.

It is now convenient to work with a class of functions which has an infinite amount of both regularity and decay.

Definition 1.12.8 (Schwartz class)**.** A *rapidly decreasing function* is a measurable function $f : \mathbf{R}^d \to \mathbf{C}$ such that $|x|^n f(x)$ is bounded for every non-negative integer n. A *Schwartz function* is a smooth function $f : \mathbf{R}^d \to \mathbf{C}$ such that all derivatives $\partial_{x_1}^{n_1} \ldots \partial_{x_d}^{n_d} f$ are rapidly decreasing. The space of all Schwartz functions is denoted $\mathcal{S}(\mathbf{R}^d)$.

Example 1.12.9. Any smooth, compactly supported function $f : \mathbf{R}^d \to \mathbf{C}$ is a Schwartz function. The gaussian functions

$$f(x) = Ae^{2\pi i\theta}e^{2\pi i\xi_0\cdot x}e^{-\pi|x-x_0|^2/R^2} \tag{1.106}$$

for $A \in \mathbf{R}$, $\theta \in \mathbf{R}/\mathbf{Z}$, $x_0, \xi_0 \in \mathbf{R}^d$ are also Schwartz functions.

Exercise 1.12.25. Show that the seminorms

$$\|f\|_{k,n} := \sup_{x\in\mathbf{R}^n} |x|^n|\nabla^k f(x)|$$

for $k, n \geq 0$, where we think of $\nabla^k f(x)$ as a d^k-dimensional vector (or, if one wishes, a rank k d-dimensional tensor), give $\mathcal{S}(\mathbf{R}^d)$ the structure of a *Fréchet space*. In particular, $\mathcal{S}(\mathbf{R}^d)$ is a topological vector space.

Clearly, every Schwartz function is both smooth and rapidly decreasing. The following exercise explores the converse:

Exercise 1.12.26.

- Give an example to show that not all smooth, rapidly decreasing functions are Schwartz.
- Show that if f is a smooth, rapidly decreasing function, and all derivatives of f are bounded, then f is Schwartz. (*Hint*: use Taylor's theorem with remainder.)

One of the reasons why the Schwartz space is convenient to work with is that it is closed under a wide variety of operations. For instance, the derivative of a Schwartz function is again a Schwartz function, and that the product of a Schwartz function with a polynomial is again a Schwartz function. Here are some further such closure properties:

Exercise 1.12.27. Show that the product of two Schwartz functions is again a Schwartz function. Moreover, show that the product map $f, g \mapsto fg$ is continuous from $\mathcal{S}(\mathbf{R}^d) \times \mathcal{S}(\mathbf{R}^d)$ to $\mathcal{S}(\mathbf{R}^d)$.

Exercise 1.12.28. Show that the convolution of two Schwartz functions is again a Schwartz function. Moreover, show that the convolution map $f, g \mapsto f * g$ is continuous from $\mathcal{S}(\mathbf{R}^d) \times \mathcal{S}(\mathbf{R}^d)$ to $\mathcal{S}(\mathbf{R}^d)$.

Exercise 1.12.29. Show that the Fourier transform of a Schwartz function is again a Schwartz function. Moreover, show that the Fourier transform map $\mathcal{F} : f \mapsto \hat{f}$ is continuous from $\mathcal{S}(\mathbf{R}^d)$ to $\mathcal{S}(\mathbf{R}^d)$.

The other important property of the Schwartz class is that it is dense in many other spaces:

Exercise 1.12.30. Show that $\mathcal{S}(\mathbf{R}^d)$ is dense in $L^p(\mathbf{R}^d)$ for every $1 \leq p < \infty$, and is also dense in $C_0(\mathbf{R}^d)$ (with the uniform topology). (*Hint*: one can either use the Stone-Weierstrass theorem, or convolutions with approximations to the identity.)

Because of this density property, it becomes possible to establish various estimates and identities in spaces of rough functions (e.g. L^p functions) by first establishing these estimates on Schwartz functions (where it is easy to justify operations such as differentiation under the integral sign) and then taking limits.

Having defined the Fourier transform $\mathcal{F} : \mathcal{S}(\mathbf{R}^d) \to \mathcal{S}(\mathbf{R}^d)$, we now introduce the *adjoint Fourier transform* $\mathcal{F}^* : \mathcal{S}(\mathbf{R}^d) \to \mathcal{S}(\mathbf{R}^d)$ by the formula

$$\mathcal{F}^* F(x) := \int_{\mathbf{R}^d} e^{2\pi i \xi \cdot x} F(\xi)\ d\xi$$

(note the sign change from (1.105)). We will shortly demonstrate that the adjoint Fourier transform is also the inverse Fourier transform: $\mathcal{F}^* = \mathcal{F}^{-1}$.

From the identity

$$\mathcal{F}^* f = \overline{\mathcal{F}\overline{f}} \tag{1.107}$$

we see that $\mathcal{F}^*$ obeys much the same propeties as $\mathcal{F}$; for instance, it is also continuous from $\mathcal{S}(\mathbf{R}^d)$ to $\mathcal{S}(\mathbf{R}^d)$. It is also the adjoint to $\mathcal{F}$ in the sense that

$$\langle \mathcal{F} f, g \rangle_{L^2(\mathbf{R}^d)} = \langle f, \mathcal{F}^* g \rangle_{L^2(\mathbf{R}^d)}$$

for all $f, g \in \mathcal{S}(\mathbf{R}^d)$.

Now we show that $\mathcal{F}^*$ inverts $\mathcal{F}$. We begin with an easy preliminary result:

Exercise 1.12.31. For any $f, g \in \mathcal{S}(\mathbf{R}^d)$, establish the identity $\mathcal{F}^*\mathcal{F}(f * g) = f * \mathcal{F}^*\mathcal{F} g$.

Next, we perform a computation:

Exercise 1.12.32 (Fourier transform of Gaussians)**.** Let $r > 0$. Show that the Fourier transform of the gaussian function $g_r(x) := r^{-d} e^{-\pi |x|^2/r^2}$ is $\hat{g}_r(\xi) = e^{-\pi r^2 |\xi|^2}$. (*Hint*: Reduce to the case $d = 1$ and $r = 1$, then complete the square and use contour integration and the classical identity $\int_{-\infty}^{\infty} e^{-\pi x^2}\ dx = 1$.) Conclude that $\mathcal{F}^* \mathcal{F} g_r = g_r$.

Exercise 1.12.33. With g_r as in the previous exercise, show that $f * g_r$ converges in the Schwartz space topology to f as $r \to 0$ for all $f \in \mathcal{S}(\mathbf{R}^d)$. (*Hint*: First show convergence in the uniform topology, then use the identities $\frac{\partial}{\partial x_j}(f * g) = (\frac{\partial}{\partial x_j} f) * g$ and $x_j (f * g) = (x_j f) * g + f(x_j g)$ for $f, g \in \mathcal{S}(\mathbf{R}^d)$.)

From Exercises 1.12.31, 1.12.32 we see that

$$\mathcal{F}^* \mathcal{F}(f * g_r) = f * g_r$$

for all $r > 0$ and $f \in \mathcal{S}(\mathbf{R}^d)$. Taking limits as $r \to 0$ using Exercises 1.12.29, 1.12.33 we conclude that

$$\mathcal{F}^* \mathcal{F} f = f$$

for all $f \in \mathcal{S}(\mathbf{R}^d)$, or in other words we have the Fourier inversion formula

$$f(x) = \int_{\mathbf{R}^d} \hat{f}(\xi) e^{2\pi i \xi \cdot x}\ d\xi \tag{1.108}$$

for all $x \in \mathbf{R}^d$. From (1.107) we also have

$$\mathcal{F} \mathcal{F}^* f = f.$$

Taking inner products with another Schwartz function g, we obtain *Parseval's identity*

$$\langle \mathcal{F} f, \mathcal{F} g \rangle_{L^2(\mathbf{R}^d)} = \langle f, g \rangle_{L^2(\mathbf{R}^d)}$$

for all $f, g \in \mathcal{S}(\mathbf{R}^d)$, and similarly for $\mathcal{F}^*$. In particular, we obtain *Plancherel's identity*

$$\|\mathcal{F} f\|_{L^2(\mathbf{R}^d)} = \|f\|_{L^2(\mathbf{R}^d)} = \|\mathcal{F}^* f\|_{L^2(\mathbf{R}^d)}$$

for all $f \in \mathcal{S}(\mathbf{R}^d)$. We conclude that

Theorem 1.12.10 (Plancherel's theorem for $\mathbf{R}^d$)**.** *The Fourier transform operator $\mathcal{F} : \mathcal{S} \to \mathcal{S}$ can be uniquely extended to a unitary transformation $\mathcal{F} : L^2(\mathbf{R}^d) \to L^2(\mathbf{R}^d)$.*

Exercise 1.12.34. Show that the Fourier transform on $L^2(\mathbf{R}^d)$ given by Plancherel's theorem agrees with the Fourier transform on $L^1(\mathbf{R}^d)$ given by (1.105) on the common domain $L^2(\mathbf{R}^d) \cap L^1(\mathbf{R}^d)$. Thus we may define $\hat{f}$ for $f \in L^1(\mathbf{R}^d)$ or $f \in L^2(\mathbf{R}^d)$ (or even $f \in L^1(\mathbf{R}^d) + L^2(\mathbf{R}^d)$ without any ambiguity (other than the usual identification of any two functions that agree almost everywhere).

Note that it is certainly possible for a function f to lie in $L^2(\mathbf{R}^d)$ but not in $L^1(\mathbf{R}^d)$ (e.g. the function $(1+|x|)^{-d}$). In such cases, the integrand in (1.105) is not absolutely integrable, and so this formula does not define the Fourier transform of f directly. Nevertheless, one can recover the Fourier transform via a limiting version of (1.105):

Exercise 1.12.35. Let $f \in L^2(\mathbf{R}^d)$. Show that the partial Fourier integrals $\xi \mapsto \int_{|x|\leq R} f(x) e^{-2\pi i \xi \cdot x}\, dx$ converge in $L^2(\mathbf{R}^d)$ to $\hat{f}$ as $R \to \infty$.

Remark 1.12.11. It is a famous open question whether the partial Fourier integrals of an $L^2(\mathbf{R}^d)$ function also converge pointwise almost everywhere for $d \geq 2$. For $d = 1$, this is essentially the celebrated theorem of Carleson mentioned in Exercise 1.12.22.

Exercise 1.12.36 (Heisenberg uncertainty principle)**.** Let $d = 1$. Define the *position operator* $X : \mathcal{S}(\mathbf{R}) \to \mathcal{S}(\mathbf{R})$ and *momentum operator* $D : \mathcal{S}(\mathbf{R}) \to \mathcal{S}(\mathbf{R})$ by the formulae

$$Xf(x) := xf(x); \quad Df(x) := \frac{-1}{2\pi i} \frac{d}{dx} f(x).$$

Establish the identities

$$\mathcal{F}D = X\mathcal{F}; \quad \mathcal{F}X = -\mathcal{F}D; \quad DX - XD = \frac{-1}{2\pi i} \tag{1.109}$$

and the formal self-adjointness relationships

$$\langle Xf, g\rangle_{L^2(\mathbf{R})} = \langle f, Xg\rangle_{L^2(\mathbf{R})}; \quad \langle Df, g\rangle_{L^2(\mathbf{R})} = \langle f, Dg\rangle_{L^2(\mathbf{R})}$$

and then establish the inequality

$$\|Xf\|_{L^2(\mathbf{R})} \|Df\|_{L^2(\mathbf{R})} \geq \frac{1}{4\pi} \|f\|^2_{L^2(\mathbf{R})}.$$

(*Hint*: start with the obvious inequality $\langle (aX+ibD)f, (aX+ibD)f\rangle_{L^2(\mathbf{R})} \geq 0$ for real numbers a, b, and optimise in a and b.) If $\|f\|_{L^2(\mathbf{R})} = 1$, deduce the *Heisenberg uncertainty principle*

$$[\int_{\mathbf{R}} (\xi - \xi_0)|\hat{f}(\xi)|^2 \, d\xi]^{1/2} [\int_{\mathbf{R}} (x - x_0)|f(x)|^2 \, dx]^{1/2} \geq \frac{1}{4\pi}$$

for any $x_0, \xi_0 \in \mathbf{R}$. (*Hint*: one can use the translation and modulation symmetries (1.100), (1.101) of the Fourier transform to reduce to the case $x_0 = \xi_0 = 0$.) Classify precisely the f, x_0, ξ_0 for which equality occurs.

Remark 1.12.12. For $x_0, \xi_0 \in \mathbf{R}^d$ and $R > 0$, define the *gaussian wave packet* $g_{x_0,\xi_0,R}$ by the formula

$$g_{x_0,\xi_0,R}(x) := 2^{d/2} R^{-d/2} e^{2\pi i \xi_0 \cdot x} e^{-\pi |x-x_0|^2/R^2}.$$

These wave packets are normalised to have L^2 norm one, and their Fourier transform is given by

$$\hat{g}_{x_0,\xi_0,R} = e^{2\pi i \xi_0 \cdot x_0} g_{\xi_0,-x_0,1/R}. \tag{1.110}$$

Informally, $g_{x_0,\xi_0,R}$ is localised to the region $x = x_0 + O(R)$ in physical space, and to the region $\xi = \xi_0 + O(1/R)$ in frequency space; observe that this is consistent with the uncertainty principle. These packets "almost diagonalise" the position and momentum operators X, D in the sense that (taking $d = 1$ for simplicity)

$$X g_{x_0,\xi_0,R} \approx x_0 g_{x_0,\xi_0,R}; \quad D g_{x_0,\xi_0,R} \approx \xi_0 g_{x_0,\xi_0,R}$$

(where the errors terms are morally of the form $O(R g_{x_0,\xi_0,R})$ and $O(R^{-1} g_{x_0,\xi_0,R})$ respectively). Of course, the non-commutativity of D and X as evidenced by the last equation in (1.109) shows that exact diagonalisation is impossible. Nevertheless it is useful, at an intuitive level at least, to view these wave-packets as a sort of (overdetermined) basis for $L^2(\mathbf{R})$ that approximately diagonalises X and D (as well as other formal combinations $a(X, D)$ of these operators, such as differential operators or *pseudodifferential operators*). Meanwhile, the Fourier transform morally maps the point (x_0, ξ_0) in phase space to $(\xi_0, -x_0)$, as evidenced by (1.110) or (1.109); it is the model example of the more general class of *Fourier integral operators*, which morally move points in phase space around by *canonical transformations*. The study of these types of objects (which are of importance in

linear PDE) is known as *microlocal analysis*, and is beyond the scope of this course.

The proof of the Hausdorff-Young inequality (1.103) carries over to the Euclidean space setting, and gives

$$\|\hat{f}\|_{L^{p'}(\mathbf{R}^d)} \leq \|f\|_{L^p(\mathbf{R}^d)} \tag{1.111}$$

for all $1 \leq p \leq 2$ and all $f \in L^p(\mathbf{R}^d)$; in particular the Fourier transform is bounded from $L^p(\mathbf{R}^d)$ to $L^{p'}(\mathbf{R}^d)$. The constant of 1 on the right-hand side of (1.111) turns out to not be optimal in the Euclidean setting, in contrast to the compact setting; the sharp constant is in fact $(p^{1/p}/(p')^{1/p'})^{d/2}$, a result of Beckner[**Be1975**]. (The fact that this constant cannot be improved can be seen by using the gaussians from Exercise 1.12.32.)

Exercise 1.12.37 (Entropy uncertainty principle)**.** For any $f \in \mathcal{S}(\mathbf{R}^d)$ with $\|f\|_{L^2(\mathbf{R}^d)} = 1$, show that

$$-\int_{\mathbf{R}^d} |f(x)|^2 \log \frac{1}{|f(x)|^2}\, dx - \int_{\mathbf{R}^d} |\hat{f}(\xi)|^2 \log \frac{1}{|\hat{f}(\xi)|^2}\, d\xi \geq 0.$$

(*Hint*: differentiate (!) (1.104) in p at $p = 2$, where one has equality in (1.104).) Using Beckner's improvement to (1.103), improve the right-hand side to the optimal value of $d \log(2e)$.

Exercise 1.12.38 (Fourier transform under linear changes of variable)**.** Let $L : \mathbf{R}^d \to \mathbf{R}^d$ be an invertible linear transformation. If $f \in \mathcal{S}(\mathbf{R}^d)$ and $f_L(x) := f(Lx)$, show that the Fourier transform of f_L is given by the formula

$$\hat{f}_L(\xi) = \frac{1}{|\det L|} \hat{f}((L^*)^{-1}\xi)$$

where $L^* : \mathbf{R}^d \to \mathbf{R}^d$ is the adjoint operator to L. Verify that this transformation is consistent with (1.104), and indeed shows that the exponent p' on the left-hand side cannot be replaced by any other exponent. (One can also establish this latter claim by dimensional analysis.)

Remark 1.12.13. As a corollary of Exercise 1.12.38, observe that if $f \in \mathcal{S}(\mathbf{R}^d)$ is spherically symmetric (thus $f = f \circ L$ for all rotation matrices L) then $\hat{f}$ is spherically symmetric also.

Exercise 1.12.39 (Fourier transform intertwines restriction and projection)**.** Let $1 \leq r \leq d$, and let $f \in \mathcal{S}(\mathbf{R}^d)$. We express $\mathbf{R}^d$ as $\mathbf{R}^r \times \mathbf{R}^{d-r}$ in the obvious manner.

- (Restriction becomes projection) If $g \in \mathcal{S}(\mathbf{R}^r)$ is the restriction $g(x) := f(x,0)$ of f to $\mathbf{R}^r \equiv \mathbf{R}^r \times \{0\}$, show that $\hat{g}(\xi) = \int_{\mathbf{R}^{d-r}} \hat{f}(\xi,\eta)\ d\eta$ for all $\xi \in \mathbf{R}^r$.
- (Projection becomes restriction) If $h \in \mathcal{S}(\mathbf{R}^r)$ is the projection $h(x) := \int_{\mathbf{R}^{d-r}} f(x,y)\ dy$ of f to $\mathbf{R}^r \equiv \mathbf{R}^d/\mathbf{R}^{d-r}$, show that $\hat{h}(\xi) = \hat{f}(\xi,0)$ for all $\xi \in \mathbf{R}^r$.

Exercise 1.12.40 (Fourier transform on large tori)**.** Let $L > 0$, and let $(\mathbf{R}/L\mathbf{Z})^d$ be the torus of length L with Lebesgue measure dx (thus the total measure of this torus is L^d. We identify the Pontryagin dual of this torus with $\frac{1}{L}\cdot\mathbf{Z}^d$ in the usual manner, thus we have the Fourier coefficients

$$\hat{f}(\xi) := \int_{(\mathbf{R}/L\mathbf{Z})^d} f(x) e^{-2\pi i \xi\cdot x}\ dx$$

for all $f \in L^1((\mathbf{R}/L\mathbf{Z})^d)$ and $\xi \in \frac{1}{L}\cdot\mathbf{Z}^d$.

- Show that for any $f \in L^2((\mathbf{R}/L\mathbf{Z})^d)$, the Fourier series $\frac{1}{L^d}\sum_{\xi\in\frac{1}{L}\cdot\mathbf{Z}^d} \hat{f}(\xi) e^{2\pi i\xi\cdot x}$ converges unconditionally in $L^2((\mathbf{R}/L\mathbf{Z})^d)$.
- Use this to give an alternate proof of the Fourier inversion formula (1.108) in the case where f is smooth and compactly supported.

Exercise 1.12.41 (Poisson summation formula)**.** Let $f \in \mathcal{S}(\mathbf{R}^d)$. Show that the function $F : (\mathbf{R}/\mathbf{Z})^d \to \mathbf{C}$ defined by $F(x+\mathbf{Z}^d) := \sum_{n\in\mathbf{Z}^d} f(x+n)$ has Fourier transform $\hat{F}(\xi) = \hat{f}(\xi)$ for all $\xi \in \mathbf{Z}^d \subset \mathbf{R}^d$ (note the two different Fourier transforms in play here). Conclude the *Poisson summation formula*

$$\sum_{n\in\mathbf{Z}^d} f(n) = \sum_{m\in\mathbf{Z}^d} \hat{f}(m).$$

Exercise 1.12.42. Let $f : \mathbf{R}^d \to \mathbf{C}$ be a compactly supported, absolutely integrable function. Show that the function $\hat{f}$ is real-analytic. Conclude that it is not possible to find a non-trivial $f \in L^1(\mathbf{R}^d)$ such that f and $\hat{f}$ are both compactly supported.

1.12.4. The Fourier transform on general groups (optional). The field of *abstract harmonic analysis* is concerned, among other things, with extensions of the above theory to more general groups, for instance arbitrary LCA groups. One of the ways to proceed is via Gelfand theory, which for instance can be used to show that the Fourier transform is at least injective:

Exercise 1.12.43 (Fourier analysis via Gelfand theory)**.** (Optional) In this exercise we use the Gelfand theory of commutative Banach *-algebras (see Section 1.10.4) to establish some basic facts of Fourier analysis in general groups. Let G be an LCA group. We view $L^1(G)$ as a commutative Banach *-algebra $L^1(G)$ (see Exercise 1.12.13).

(a) If $f \in L^1(G)$ is such that $\liminf_{n\to\infty} \|f^{*n}\|_{L^1(G)}^{1/n} > 0$, where $f^{*n} = f * \ldots * f$ is the convolution of n copies of f, show that there exists a non-zero complex number z such that the map $g \mapsto f * g - zg$ is not invertible on $L^1(G)$. (*Hint*: If $L^1(G)$ contains a unit, one can use Exercise 1.10.36; otherwise, adjoin a unit.)

(b) If f and z are as in (a), show that there exists a character $\lambda : L^1(G) \to \mathbf{C}$ (in the sense of Banach *-algebras, see Definition 1.10.25) such that $f * g - zg$ lies in the kernel of λ for all $g \in L^1(G)$. Conclude in particular that $\lambda(f)$ is non-zero.

(c) If $\lambda : L^1(G) \to \mathbf{C}$ is a character, show that there exists a multiplicative character $\chi : G \to S^1$ such that $\lambda(f) = \langle f, \chi\rangle$ for all $f \in L^1(G)$. (You will need Exercise 1.12.5 and Exercise 1.12.10.)

(d) For any $f \in L^1(G)$ and $g \in L^2(G)$, show that $|f*g*g^*(0)| \leq |f * f^* * g * g^*(0)|^{1/2} |g * g^*(0)|^{1/2}$, where 0 is the group identity and $f^*(x) := \overline{f(-x)}$ is the conjugate of f. (*Hint*: the inner product $\langle f_1, f_2\rangle_g := f_1 * f_2^* * g * g^*(0)$ is positive semi-definite.)

(e) Show that if $f \in L^1(G)$ is not identically zero, then there exists $\xi \in \hat{G}$ such that $\hat{f}(\xi) \neq 0$. (*Hint*: first find $g \in L^2(G)$ such that $f * g * g^*(0) \neq 0$ and $g * g^*(0) \neq 0$, and conclude using (d) repeatedly that $\liminf_{n\to\infty} \|(f * f^*)^{*n}\|_{L^1(G)}^{1/n} > 0$.

Then use (a), (b), (c).) Conclude that the Fourier transform is injective on $L^1(G)$. (The image of $L^1(G)$ under the Fourier transform is then a Banach *-algebra known as the *Wiener algebra*, and is denoted $A(\hat{G})$.)

(f) Prove Theorem 1.12.4.

It is possible to use arguments similar to those in Exercise 1.12.43 to characterise positive measures on $\hat{G}$ in terms of continuous functions on G, leading to *Bochner's theorem*:

Theorem 1.12.14 (Bochner's theorem)**.** *Let $\phi \in C(G)$ be a continuous function on an LCA group G. Then the following are equivalent:*

(a) $\sum_{n=1}^N \sum_{m=1}^N c_n \overline{c_m} \phi(x_n - x_m) \geq 0$ *for all* $x_1, \ldots, x_N \in G$ *and* $c_1, \ldots, c_N \in \mathbf{C}$.

(b) *There exists a non-negative finite Radon measure ν on $\hat{G}$ such that $\phi(x) = \int_{\hat{G}} e^{2\pi i \xi \cdot x}\ d\nu(\xi)$.*

Functions obeying either (a) or (b) are known as *positive-definite functions*. The space of such functions is denoted $B(G)$.

Exercise 1.12.44. Show that (b) implies (a) in Bochner's theorem. (The converse implication is significantly harder, reprising much of the machinery in Exercise 1.12.43, but with ϕ taking the place of $g * g^*$: see [**Ru1962**] for details.)

Using Bochner's theorem, it is possible to show

Theorem 1.12.15 (Plancherel's theorem for LCA groups)**.** *Let G be an LCA group with non-trivial Haar measure μ. Then there exists a non-trivial Haar measure ν on $\hat{G}$ such that the Fourier transform on $L^1(G) \cap L^2(G)$ can be extended continuously to a unitary transformation from $L^2(G)$ to $L^2(\hat{G})$. In particular we have the Plancherel identity*

$$\int_G |f(x)|^2\ d\mu(x) = \int_{\hat{G}} |\hat{f}(\xi)|^2\ d\nu(\xi)$$

for all $f \in L^2(G)$, and the Parseval identity

$$\int_G f(x)\overline{g(x)}\ d\mu(x) = \int_{\hat{G}} \hat{f}(\xi)\overline{\hat{g}(\xi)}\ d\nu(\xi)$$

for all $f, g \in L^2(G)$. Furthermore, the inversion formula

$$f(x) = \int_{\hat{G}} \hat{f}(\xi) e^{2\pi i \xi \cdot x} \, d\nu(\xi)$$

is valid for f in a dense subclass of $L^2(G)$ (in particular, it is valid for $f \in L^1(G) \cap B(G)$).

Again, see [**Ru1962**] for details. A related result is that of *Pontryagin duality*: if $\hat{G}$ is the Pontryagin dual of an LCA group G, then G is the Pontryagin dual of $\hat{G}$. (Certainly, every element $x \in G$ defines a character $\hat{x} : \xi \mapsto \xi \cdot x$ on $\hat{G}$, thus embedding G into $\hat{\hat{G}}$ via the *Gelfand transform* (see Section 1.10.4); the non-trivial fact is that this embedding is in fact surjective.) One can use Pontryagin duality to convert various properties of LCA groups into other properties on LCA groups. For instance, we have already seen that $\hat{G}$ is compact (resp. discrete) if G is discrete (resp. compact); with Pontryagin duality, the implications can now also be reversed. As another example, one can show that $\hat{G}$ is connected (resp. torsion-free) if and only if G is torsion-free (resp. connected). We will not prove these assertions here.

It is natural to ask what happens for non-abelian locally compact groups $G = (G, \cdot)$. One can still build non-trivial Haar measures (the proof sketched out in Exercise 1.12.7 extends without difficulty to the non-abelian setting), though one must now distinguish between left-invariant and right-invariant Haar measures. (The two notions are equivalent for some classes of groups, notably compact groups, but not in general. Groups for which the two notions of Haar measures coincide are called *unimodular*.) However, when G is non-abelian then there are not enough multiplicative characters $\chi : G \to S^1$ to have a satisfactory Fourier analysis. (Indeed, such characters must annihilate the commutator group $[G, G]$, and it is entirely possible for this commutator group to be all of G, e.g. if G is *simple* and non-abelian.) Instead, one must generalise the notion of a multiplicative character to that of a *unitary representation* $\rho : G \to U(H)$ from G to the group of unitary transformations on a complex Hilbert space H; thus the Fourier coefficients $\hat{f}(\rho)$ of a function will now be operators on thisl Hilbert space H, rather than complex numbers. When G

is a compact group, it turns out to be possible to restrict attention to finite-dimensional representations (thus one can replace $U(H)$ by the matrix group $U(n)$ for some n). The analogue of the Pontryagin dual $\hat{G}$ is then the collection of (irreducible) finite-dimensional unitary representations of G, up to isomorphism. There is an analogue of the Plancherel theorem in this setting, closely related to the *Peter-Weyl theorem* in representation theory. We will not discuss these topics here, but refer the reader instead to any representation theory text.

The situation for non-compact non-abelian groups (e.g. $SL_2(\mathbf{R})$) is significantly more subtle, as one must now consider infinite-dimensional representations as well as finite-dimensional ones, and the inversion formula can become quite non-trivial (one has to decide what "weight" each representation should be assigned in that formula). At this point it seems unprofitable to work in the category of locally compact groups, and specialise to a more structured class of groups, e.g. algebraic groups. The representation theory of such groups is a massive subject and well beyond the scope of this course.

1.12.5. Relatives of the Fourier transform (optional). There are a number of other Fourier-like transforms used in mathematics, which we will briefly survey here. Firstly, there are some rather trivial modifications one can make to the definition of Fourier transform, for instance by replacing the complex exponential $e^{2\pi ix}$ by trigonometric functions such as $\sin(x)$ and $\cos(x)$, or moving around the various factors of 2π, i, -1, etc. in the definition. In this spirit, we have the *Laplace transform*

$$\mathcal{L}f(t) := \int_0^\infty f(s)e^{-st}\,ds \tag{1.112}$$

of a measurable function $f : [0, +\infty) \to \mathbf{R}$ with some reasonable growth at infinity, where $t > 0$. Roughly speaking, the Laplace transform is "the Fourier transform without the i" (cf. *Wick rotation*), and so has the (mild) advantage of being definable in the realm of real-valued functions rather than complex-valued functions. It is particularly well suited for studying ODE on the half-line $[0, +\infty)$ (e.g. initial value problems for a finite-dimensional system). The Laplace transform and Fourier transform can be unified by allowing the t parameter in (1.112) to vary in the right-half plane $\{t \in \mathbf{C} : \mathrm{Re}(t) \geq 0\}$.

When the Fourier transform is applied to a spherically symmetric function $f(x) := F(|x|)$ on $\mathbf{R}^d$, then the Fourier transform is also spherically symmetric, given by the formula $\hat{f}(\xi) = G(|\xi|)$ where G is the *Fourier-Bessel transform* (or *Hankel transform*)

$$G(r) := 2\pi r^{-(d-2)/2} \int_0^\infty F(s) J_{(d-2)/2}(2\pi rs) s^{d/2}\, ds$$

where J_ν is the *Bessel function of the first kind* with index ν. In practice, one can then analyse the Fourier-analytic behaviour of spherically symmetric functions in terms of one-dimensional Fourier-like integrals by using various asymptotic expansions of the Bessel function.

There is a relationship between the d-dimensional Fourier transform and the one-dimensional Fourier transform, provided by the *Radon transform*, defined for $f \in \mathcal{S}(\mathbf{R}^d)$ (say) by the formula

$$\mathcal{R}f(\omega, t) := \int_{x\cdot\omega=t} f$$

where $\omega \in S^{d-1}$, $t \in \mathbf{R}$, and the integration is with respect to $d-1$-dimensional measure. Indeed one checks that the d-dimensional Fourier transform of f at $r\omega$ for some $r > 0$ and $\omega \in S^{d-1}$ is nothing more than the one-dimensional Fourier coefficient of the function $t \mapsto \mathcal{R}f(\omega, t)$ at r. The Radon transform is often used in scattering theory and related areas of analysis, geometry, and physics.

In analytic number theory, a multiplicative version of the Fourier-Laplace transform is often used, namely the *Mellin transform*

$$\mathcal{M}f(s) := \int_0^\infty x^s f(x) \frac{dx}{x}.$$

(Note that $\frac{dx}{x}$ is a Haar measure for the *multiplicative* group $\mathbf{R}^+ = (0, +\infty)$.) To see the relation with the Fourier-Laplace transform, write $f(x) = F(\log x)$, then the Mellin transform becomes

$$\mathcal{M}f(s) = \int_{\mathbf{R}} e^{st} f(t)\, dt.$$

Many functions of importance in analytic number theory, such as the *Gamma function* or the *zeta function*, can be expressed neatly in terms of Mellin transforms.

In electrical engineering and signal processing, the *z-transform* is often used, transforming a sequence $c = (c_n)_{n=-\infty}^{\infty}$ of complex numbers to a formal Laurent series

$$\mathcal{Z}c(z) := \sum_{n=-\infty}^{\infty} c_n z^n$$

(some authors use z^{-n} instead of z^n here). If one makes the substitution $z = e^{2\pi i n x}$ then this becomes a (formal) Fourier series expansion on the unit circle. If the sequence c_n is restricted to only be non-zero for non-negative n, and does not grow too quickly as $n \to \infty$, then the z-transform becomes holomorphic on the unit disk, thus providing a link between Fourier analysis and complex analysis. For instance, the standard formula

$$c_n = \frac{1}{2\pi i} \int_{|z|=1} \frac{f(z)}{z^{n+1}}\ dz$$

for the Taylor coefficients of a holomorphic function $f(z) = \sum_{n=0}^{\infty} c_n z^n$ at the origin can be viewed as a version of the Fourier inversion formula for the torus $\mathbf{R}/\mathbf{Z}$. Just as the Fourier or Laplace transforms are useful for analysing differential equations in continuous settings, the z-transform is useful for analysing difference equations in discrete settings. The z-transform is of course also very similar to the method of *generating functions* in combinatorics and probability.

In probability theory one also considers the *characteristic function* $\mathbf{E}(e^{itX})$ of a real-valued random variable X; this is essentially the Fourier transform of the probability distribution of X. Just as the Fourier transform is useful for understanding convolutions $f * g$, the characteristic function is useful for understanding sums $X_1 + X_2$ of independent random variables.

We have briefly touched upon the role of Gelfand theory in the general theory of the Fourier transform. Indeed, one can view the Fourier transform as the special case of the *Gelfand transform* for Banach *-algebras, which we already discussed in Section 1.10.4.

The *Fast Fourier Transform* (FFT) is not, strictly speaking, a variant of the Fourier transform, but rather is an efficient algorithm

for *computing* the Fourier transform

$$\hat{f}(\xi) = \frac{1}{N} \sum_{n=0}^{N-1} f(x) e^{-2\pi i \xi x/N}$$

on a cyclic group $\mathbf{Z}/N\mathbf{Z} \equiv \{0, \ldots, N-1\}$, when N is large but composite. Note that a brute force computation of this transform for all N values of ξ would require about $O(N^2)$ addition and multiplication operations. The FFT algorithm, in contrast, takes only $O(N \log N)$ operations, and is based on reducing the FFT for a large N to FFT for smaller N. For instance, suppose N is even, say $N = 2M$, then observe that

$$\hat{f}(\xi) = \frac{1}{2}(\hat{f}_0(\xi) + e^{-2\pi i \xi/N} \hat{f}_1(\xi))$$

where $f_0, f_1 : \mathbf{Z}/M\mathbf{Z} \to \mathbf{C}$ are the functions $f_j(x) := f(2x+j)$. Thus one can obtain the Fourier transform of the length N vector f from the Fourier transforms of the two length M vectors f_0, f_1 after about $O(N)$ operations. Iterating this we see that we can indeed compute $\hat{f}$ in $O(N \log N)$ operations, at least in the model case when N is a power of two; the general case has a similar but more complicated analysis.

In many situations (particularly in ergodic theory), it is desirable not to perform Fourier analysis on a group G directly, but instead on another space X that G acts on. Suppose for instance that G is a compact abelian group, with probability Haar measure dg, which acts in a measure-preserving (and measurable) fashion on a probability space (X, μ). Then one can decompose any $f \in L^2(X)$ into Fourier components $f = \sum_{\xi \in \hat{G}} f_\xi$, where $f_\xi(x) := \int_G e^{-2\pi i \xi \cdot g} f(gx)\, dg$, where the series is unconditionally convergent in $L^2(X)$. The reason for doing this is that each of the f_ξ behaves in a simple way with respect to the group action, indeed one has $f_\xi(gx) = e^{2\pi i \xi \cdot g} f_\xi(x)$ for (almost) all $g \in G, x \in X$. This decomposition is closely related to the decomposition in representation theory of a given representation into irreducible components. Perhaps the most basic example of this type of operation is the decomposition of a function $f : \mathbf{R} \to \mathbf{R}$ into even and odd components $\frac{f(x)+f(-x)}{2}$, $\frac{f(x)-f(-x)}{2}$; here the underlying group is $\mathbf{Z}/2\mathbf{Z}$, which acts on $\mathbf{R}$ by reflections, $gx := (-1)^g x$.

The operation of converting a square matrix $A = (a_{ij})_{1 \leq i,j \leq n}$ of numbers into eigenvalues $\lambda_1, \ldots, \lambda_n$ or singular values $\sigma_1, \ldots, \sigma_n$ can be viewed as a sort of non-commutative generalisation of the Fourier transform. (Note that the eigenvalues of a *circulant matrix* are essentially the Fourier coefficients of the first row of that matrix.) For instance, the identity $\sum_{i=1}^n \sum_{j=1}^n |a_{ij}|^2 = \sum_{k=1}^n \sigma_k^2$ can be viewed as a variant of the Plancherel identity. More generally, there are close relationships between spectral theory and Fourier analysis (as one can already see from the connection to Gelfand theory). For instance, in $\mathbf{R}^d$ and $\mathbf{T}^d$, one can view Fourier analysis as the spectral theory of the gradient operator ∇ (note that the characters $e^{2\pi i \xi \cdot x}$ are joint eigenfunctions of ∇). As the gradient operator is closely related to the Laplacian Δ, it is not surprising that Fourier analysis is also closely related to the spectral theory of the Laplacian, and in particular to various operators built using the Laplacian (e.g. resolvents, heat kernels, wave operators, Schrödinger operators, Littlewood-Paley projections, etc.). Indeed, the spectral theory of the Laplacian can serve as a partial substitute for the Fourier transform in situations in which there is not enough symmetry to exploit Fourier-analytic techniques (e.g. on a manifold with no translation symmetries).

Finally, there is an analogue of the Fourier duality relationship between an LCA group G and its Pontryagin dual $\hat{G}$ in algebraic geometry, known as the *Fourier-Mukai transform*, which relates an *abelian variety* X to its *dual* $\hat{X}$, and transforms *coherent sheaves* on the former to coherent sheaves on the latter. This transform obeys many of the algebraic identities that the Fourier transform does, although it does not seem to have much of the analytic structure.

Notes. This lecture first appeared at `terrytao.wordpress.com/2009/04/06`. Thanks to Hunter, Marco Frasca, Max Baroi, PDEbeginner, timur, Xiaochuan Liu, and anonymous commenters for corrections.

1.13. Distributions

In set theory, a function $f : X \to Y$ is defined as an object that *evaluates* every input x to exactly one output $f(x)$. However, in various branches of mathematics, it has become convenient to generalise this

classical concept of a function to a more abstract one. For instance, in *operator algebras*, *quantum mechanics*, or *non-commutative geometry*, one often replaces commutative algebras of (real or complex-valued) functions on some space X, such as $C(X)$ or $L^\infty(X)$, with a more general - and possibly non-commutative - algebra (e.g. a *C^*-algebra* or a *von Neumann algebra*). Elements in this more abstract algebra are no longer definable as functions in the classical sense of assigning a single value $f(x)$ to every point $x \in X$, but one can still define other operations on these "*generalised functions*" (e.g. one can multiply or take inner products between two such objects).

Generalisations of functions are also very useful in analysis. In our study of L^p spaces, we have already seen one such generalisation, namely the concept of a function defined up to almost everywhere equivalence. Such a function f (or more precisely, an equivalence class of classical functions) cannot be evaluated at any given point x, if that point has measure zero. However, it is still possible to perform algebraic operations on such functions (e.g. multiplying or adding two functions together), and one can also integrate such functions on measurable sets (provided, of course, that the function has some suitable integrability condition). We also know that the L^p spaces can usually be described via duality, as the dual space of $L^{p'}$ (except in some endpoint cases, namely when $p = \infty$, or when $p = 1$ and the underlying space is not σ-finite).

We have also seen (via the *Lebesgue-Radon-Nikodym theorem*) that locally integrable functions $f \in L^1_{\mathrm{loc}}(\mathbf{R})$ on, say, the real line $\mathbf{R}$, can be identified with locally finite absolutely continuous measures m_f on the line, by multiplying Lebesgue measure m by the function f. So another way to generalise the concept of a function is to consider arbitrary locally finite *Radon measures* μ (not necessarily absolutely continuous), such as the *Dirac measure* δ_0. With this concept of "generalised function", one can still add and subtract two measures μ, ν, and integrate any measure μ against a (bounded) measurable set E to obtain a number $\mu(E)$, but one cannot evaluate a measure μ (or more precisely, the Radon-Nikodym derivative $d\mu/dm$ of that measure) at a single point x, and one also cannot multiply two measures together to obtain another measure. From the Riesz representation

theorem, we also know that the space of (finite) Radon measures can be described via duality, as linear functionals on $C_c(\mathbf{R})$.

There is an even larger class of generalised functions that is very useful, particularly in linear PDE, namely the space of *distributions*, say on a Euclidean space $\mathbf{R}^d$. In contrast to Radon measures μ, which can be defined by how they "pair up" against continuous, compactly supported test functions $f \in C_c(\mathbf{R}^d)$ to create numbers $\langle f, \mu \rangle := \int_{\mathbf{R}^d} f \, d\overline{\mu}$, a distribution λ is defined by how it pairs up against a *smooth* compactly supported function $f \in C_c^\infty(\mathbf{R}^d)$ to create a number $\langle f, \lambda \rangle$. As the space $C_c^\infty(\mathbf{R}^d)$ of smooth compactly supported functions is smaller than (but dense in) the space $C_c(\mathbf{R}^d)$ of continuous compactly supported functions (and has a stronger topology), the space of distributions is larger than that of measures. But the space $C_c^\infty(\mathbf{R}^d)$ is closed under more operations than $C_c(\mathbf{R}^d)$, and in particular is closed under differential operators (with smooth coefficients). Because of this, the space of distributions is similarly closed under such operations; in particular, one can differentiate a distribution and get another distribution, which is something that is not always possible with measures or L^p functions. But as measures or functions can be interpreted as distributions, this leads to the notion of a *weak derivative* for such objects, which makes sense (but only as a distribution) even for functions that are not classically differentiable. Thus the theory of distributions can allow one to rigorously manipulate rough functions "as if" they were smooth, although one must still be careful as some operations on distributions are not well-defined, most notably the operation of multiplying two distributions together. Nevertheless one can use this theory to justify many formal computations involving derivatives, integrals, etc. (including several computations used routinely in physics) that would be difficult to formalise rigorously in a purely classical framework.

If one shrinks the space of distributions slightly, to the space of *tempered distributions* (which is formed by enlarging dual class $C_c^\infty(\mathbf{R}^d)$ to the *Schwartz class* $\mathcal{S}(\mathbf{R}^d)$), then one obtains closure under another important operation, namely the *Fourier transform*. This allows one to define various Fourier-analytic operations (e.g. *pseudo-differential operators*) on such distributions.

Of course, at the end of the day, one is usually not all that interested in distributions in their own right, but would like to be able to use them as a tool to study more classical objects, such as smooth functions. Fortunately, one can recover facts about smooth functions from facts about the (far rougher) space of distributions in a number of ways. For instance, if one convolves a distribution with a smooth, compactly supported function, one gets back a smooth function. This is a particularly useful fact in the theory of constant-coefficient linear partial differential equations such as $Lu = f$, as it allows one to recover a smooth solution u from smooth, compactly supported data f by convolving f with a specific distribution G, known as the *fundamental solution* of L. We will give some examples of this later in this section.

It is this unusual and useful combination of both being able to pass from classical functions to generalised functions (e.g. by differentiation) and then back from generalised functions to classical functions (e.g. by convolution) that sets the theory of distributions apart from other competing theories of generalised functions, in particular allowing one to justify many formal calculations in PDE and Fourier analysis rigorously with relatively little additional effort. On the other hand, being defined by linear duality, the theory of distributions becomes somewhat less useful when one moves to more nonlinear problems, such as nonlinear PDE. However, they still serve an important supporting role in such problems as a "ambient space" of functions, inside of which one carves out more useful function spaces, such as Sobolev spaces, which we will discuss in the next set of notes.

1.13.1. Smooth functions with compact support. In the rest of the notes we will work on a fixed Euclidean space $\mathbf{R}^d$. (One can also define distributions on other domains related to $\mathbf{R}^d$, such as open subsets of $\mathbf{R}^d$, or d-dimensional manifolds, but for simplicity we shall restrict attention to Euclidean spaces in these notes.)

A *test function* is any smooth, compactly supported function $f : \mathbf{R}^d \to \mathbf{C}$; the space of such functions is denoted $C_c^\infty(\mathbf{R}^d)$. (In some texts, this space is denoted $C_0^\infty(\mathbf{R}^d)$ instead.)

From *analytic continuation* one sees that there are no real-analytic test functions other than the zero function. Despite this negative result, test functions actually exist in abundance:

Exercise 1.13.1.

(i) Show that there exists at least one test function that is not identically zero. (*Hint*: it suffices to do this for $d = 1$. One starting point is to use the fact that the function $f : \mathbf{R} \to \mathbf{R}$ defined by $f(x) := e^{-1/x}$ for $x > 0$ and $f(x) := 0$ otherwise is smooth, even at the origin 0.)

(ii) Show that if $f \in C_c^\infty(\mathbf{R}^d)$ and $g : \mathbf{R}^d \to \mathbf{R}$ is absolutely integrable and compactly supported, then the convolution $f * g$ is also in $C_c^\infty(\mathbf{R}^d)$. (*Hint*: first show that $f * g$ is continuously differentiable with $\nabla(f * g) = (\nabla f) * g$.)

(iii) (C^∞ Urysohn lemma) Let K be a compact subset of $\mathbf{R}^d$, and let U be an open neighbourhood of K. Show that there exists a function $f : C_c^\infty(\mathbf{R}^d)$ supported in U which equals 1 on K. (*Hint*: use the ordinary Urysohn lemma to find a function in $C_c(\mathbf{R}^d)$ that equals 1 on a neighbourhood of K and is supported in a compact subset of U, then convolve this function by a suitable test function.)

(iv) Show that $C_c^\infty(\mathbf{R}^d)$ is dense in $C_0(\mathbf{R}^d)$ (in the uniform topology), and dense in $L^p(\mathbf{R}^d)$ (with the L^p topology) for all $0 < p < \infty$.

The space $C_c^\infty(\mathbf{R}^d)$ is clearly a vector space. Now we place a (very strong!) topology on it. We first observe that $C_c^\infty(\mathbf{R}^d) = \bigcup_K C_c^\infty(K)$, where K ranges over all compact subsets of $\mathbf{R}^d$ and $C_c^\infty(K)$ consists of those functions $f \in C_c^\infty(\mathbf{R}^d)$ which are supported in K. Each $C_c^\infty(K)$ will be given a topology (called the *smooth topology*) generated by the norms

$$\|f\|_{C^k} := \sup_{x \in \mathbf{R}^d} \sum_{j=0}^{k} |\nabla^j f(x)|$$

for $k = 0, 1, \ldots$, where we view $\nabla^j f(x)$ as a d^j-dimensional vector (or, if one wishes, a d-dimensional rank j tensor); thus a sequence $f_n \in C_c^\infty(K)$ converges to a limit $f \in C_c^\infty(K)$ if and only if $\nabla^j f_n$

converges uniformly to $\nabla^j f$ for all $j = 0, 1, \ldots$. (This gives $C_c^\infty(K)$ the structure of a *Fréchet space*, though we will not use this fact here.)

We make the trivial remark that if $K \subset K'$ are compact sets, then $C_c^\infty(K)$ is a subspace of $C_c^\infty(K')$, and the topology on the former space is the restriction of the topology of the latter space. Because of this, we are able to give $C_c^\infty(\mathbf{R}^d)$ the *final topology* induced by the topologies on the $C_c^\infty(K)$, defined as the strongest topology on $C_c^\infty(\mathbf{R}^d)$ which restricts to the topologies on $C_c^\infty(K)$ for each K. Equivalently, a set is open in $C_c^\infty(\mathbf{R}^d)$ if and only if its restriction to $C_c^\infty(K)$ is open for every compact K.

Exercise 1.13.2. Let f_n be a sequence in $C_c^\infty(\mathbf{R}^d)$, and let f be another function in $C_c^\infty(\mathbf{R}^d)$. Show that f_n converges in the topology of $C_c^\infty(\mathbf{R}^d)$ to f if and only if there exists a compact set K such that f_n, f are all supported in K, and f_n converges to f in the smooth topology of $C_c^\infty(K)$.

Exercise 1.13.3.

(i) Show that the topology of $C_c^\infty(K)$ is *first countable* for every compact K.

(ii) Show that the topology of $C_c^\infty(\mathbf{R}^d)$ is *not* first countable. (*Hint*: given any countable sequence of open neighbourhoods of 0, build a new open neighbourhood that does not contain any of the previous ones, using the σ-compact nature of $\mathbf{R}^d$.)

(iii) Despite this, show that an element $f \in C_c^\infty(\mathbf{R}^d)$ is an adherent point of a set $E \subset C_c^\infty(\mathbf{R}^d)$ if and only if there is a sequence $f_n \in E$ that converges to f. (*Hint*: argue by contradiction.) Conclude in particular that a subset of $C_c^\infty(\mathbf{R}^d)$ is closed if and only if it is sequentially closed. Thus while first countability fails for $C_c^\infty(\mathbf{R}^d)$, we have a serviceable substitute for this property.

There are plenty of continuous operations on $C_c^\infty(\mathbf{R}^d)$:

Exercise 1.13.4.

(i) Let K be a compact set. Show that a linear map $T : C_c^\infty(K) \to X$ into a normed vector space X is continuous if and only if there exists $k \geq 0$ and $C > 0$ such that $\|Tf\|_X \leq C\|f\|_{C^k}$ for all $f \in C_c^\infty(K)$.

(ii) Let K, K' be compact sets. Show that a linear map $T : C_c^\infty(K) \to C_c^\infty(K')$ is continuous if and only if for every $k \geq 0$ there exists $k' \geq 0$ and a constant $C_k > 0$ such that $\|Tf\|_{C^k} \leq C_k\|f\|_{C^{k'}}$ for all $f \in C_c^\infty(K)$.

(iii) Show that a map $T : C_c^\infty(\mathbf{R}^d) \to X$ to a topological space is continuous if and only if for every compact set $K \subset \mathbf{R}^d$, T maps $C_c^\infty(K)$ continuously to X.

(iv) Show that the inclusion map from $C_c^\infty(\mathbf{R}^d)$ to $L^p(\mathbf{R}^d)$ is continuous for every $0 < p \leq \infty$.

(v) Show that a map $T : C_c^\infty(\mathbf{R}^d) \to C_c^\infty(\mathbf{R}^d)$ is continuous if and only if for every compact set $K \subset \mathbf{R}^d$ there exists a compact set K' such that T maps $C_c^\infty(K)$ continuously to $C_c^\infty(K')$.

(vi) Show that every linear differential operator with smooth coefficients is a continuous operation on $C_c^\infty(\mathbf{R}^d)$.

(vii) Show that convolution with any absolutely integrable, compactly supported function is a continuous operation on $C_c^\infty(\mathbf{R}^d)$.

(viii) Show that $C_c^\infty(\mathbf{R}^d)$ is a *topological vector space*.

(ix) Show that the product operation $f, g \mapsto fg$ is continuous from $C_c^\infty(\mathbf{R}^d) \times C_c^\infty(\mathbf{R}^d)$ to $C_c^\infty(\mathbf{R}^d)$.

A sequence $\phi_n \in C_c(\mathbf{R}^d)$ of continuous, compactly supported functions is said to be an *approximation to the identity* if the ϕ_n are non-negative, have total mass $\int_{\mathbf{R}^n} \phi_n$ equal to 1, and converge uniformly to zero away from the origin, thus $\sup_{|x|\geq r} |\phi_n(x)| \to 0$ for all $r > 0$. One can generate such a sequence by starting with a single non-negative continuous compactly supported function ϕ of total mass 1, and then setting $\phi_n(x) := n^d\phi(nx)$; many other constructions are possible also.

One has the following useful fact:

Exercise 1.13.5. Let $\phi_n \in C_c^\infty(\mathbf{R}^d)$ be a sequence of approximations to the identity.

(i) If $f \in C(\mathbf{R}^d)$ is continuous, show that $f * \phi_n$ converges uniformly on compact sets to f.

(ii) If $f \in L^p(\mathbf{R}^d)$ for some $1 \leq p < \infty$, show that $f * \phi_n$ converges in $L^p(\mathbf{R}^d)$ to f. (*Hint*: use (i), the density of $C_0(\mathbf{R}^d)$ in $L^p(\mathbf{R}^d)$, and Young's inequality, Exercise 1.11.25.)

(iii) If $f \in C_c^\infty(\mathbf{R}^d)$, show that $f * \phi_n$ converges in $C_c^\infty(\mathbf{R}^d)$ to f. (*Hint*: use the identity $\nabla(f * \phi_n) = (\nabla f) * \phi_n$, cf. Exercise 1.13.1(ii).)

Exercise 1.13.6. Show that $C_c^\infty(\mathbf{R}^d)$ is separable. (*Hint*: it suffices to show that $C_c^\infty(K)$ is separable for each compact K. There are several ways to accomplish this. One is to begin with the Stone-Weierstrass theorem, which will give a countable set which is dense in the uniform topology, then use the fundamental theorem of calculus to strengthen the topology. Another is to use Exercise 1.13.5 and then discretise the convolution. Another is to embed K into a torus and use Fourier series, noting that the Fourier coefficients $\hat{f}$ of a smooth function $f : \mathbf{T}^d \to \mathbf{C}$ decay faster than any power of $|n|$.)

1.13.2. Distributions. Now we can define the concept of a *distribution*.

Definition 1.13.1 (Distribution)**.** A *distribution* on $\mathbf{R}^d$ is a continuous linear functional $\lambda : f \mapsto \langle f, \lambda \rangle$ from $C_c^\infty(\mathbf{R}^d)$ to $\mathbf{C}$. The space of such distributions is denoted $C_c^\infty(\mathbf{R}^d)^*$, and is given the *weak-* topology*. In particular, a sequence of distributions λ_n converges (in the sense of distributions) to a limit λ if one has $\langle f, \lambda_n \rangle \to \langle f, \lambda \rangle$ for all $f \in C_c^\infty(\mathbf{R}^d)$.

A technical point: we endow the space $C_c^\infty(\mathbf{R}^d)^*$ with the *conjugate* complex structure. Thus, if $\lambda \in C_c^\infty(\mathbf{R}^d)^*$, and c is a complex number, then $c\lambda$ is the distribution that maps a test function f to $\overline{c}\langle f, \lambda \rangle$ rather than $c\langle f, \lambda \rangle$; thus $\langle f, c\lambda \rangle = \overline{c}\langle f, \lambda \rangle$. This is to keep the analogy between the evaluation of a distribution against a function, and the usual Hermitian inner product $\langle f, g \rangle = \int_{\mathbf{R}^d} f\overline{g}$ of two test functions.

From Exercise 1.13.4, we see that a linear functional $\lambda : C_c^\infty(\mathbf{R}^d) \to \mathbf{C}$ is a distribution if, for every compact set $K \subset \mathbf{R}^d$, there exists $k \geq 0$ and $C > 0$ such that

$$|\langle f, \lambda \rangle| \leq C \|f\|_{C^k} \tag{1.113}$$

for all $f \in C_c^\infty(K)$.

Exercise 1.13.7. Show that $C_c^\infty(\mathbf{R}^d)^*$ is a Hausdorff topological vector space.

We note two basic examples of distributions:

- Any locally integrable function $g \in L^1_{\mathrm{loc}}(\mathbf{R}^d)$ can be viewed as a distribution, by writing $\langle f, g \rangle := \int_{\mathbf{R}^d} f(x)\overline{g(x)}\, dx$ for all test functions f.
- Any complex Radon measure μ can be viewed as a distribution, by writing $\langle f, \mu \rangle := \int_{\mathbf{R}^d} f(x)\, d\overline{\mu}$, where $\overline{\mu}$ is the complex conjugate of μ (thus $\overline{\mu}(E) := \overline{\mu(E)}$). (Note that this example generalises the preceding one, which corresponds to the case when μ is absolutely continuous with respect to Lebesgue measure.) Thus, for instance, the Dirac measure δ at the origin is a distribution, with $\langle f, \delta \rangle = f(0)$ for all test functions f.

Exercise 1.13.8. Show that the above identifications of locally integrable functions or complex Radon measures with distributions are injective. (*Hint*: use Exercise 1.13.1(iv).)

From the above exercise, we may view locally integrable functions and locally finite measures as a special type of distribution. In particular, $C_c^\infty(\mathbf{R}^d)$ and $L^p(\mathbf{R}^d)$ are now contained in $C_c^\infty(\mathbf{R}^d)^*$ for all $1 \leq p \leq \infty$.

Exercise 1.13.9. Show that if a sequence of locally integrable functions converge in L^1_{loc} to a limit, then they also converge in the sense of distributions; similarly, if a sequence of complex Radon measures converge in the vague topology to a limit, then they also converge in the sense of distributions.

Thus we see that convergence in the sense of distributions is among the weakest of the notions of convergence used in analysis;

however, from the Hausdorff property, distributional limits are still *unique*.

Exercise 1.13.10. If ϕ_n is a sequence of approximations to the identity, show that ϕ_n converges in the sense of distributions to the Dirac distribution δ.

More exotic examples of distributions can be given:

Exercise 1.13.11 (Derivative of the delta function)**.** Let $d = 1$. Show that the functional $\delta' : f \mapsto -f'(0)$ for all test functions f is a distribution which does not arise from either a locally integrable function or a Radon measure. (Note how it is important here that f is smooth (and in particular differentiable, and not merely continuous.) The presence of the minus sign will be explained shortly.

Exercise 1.13.12 (Principal value of $1/x$)**.** Let $d = 1$. Show that the functional $\mathrm{p.\,v.}\, 1/x$ defined by the formula

$$\langle f, \mathrm{p.\,v.}\, \frac{1}{x} \rangle := \lim_{\varepsilon \to 0} \int_{|x|>\varepsilon} \frac{f(x)}{x}\, dx$$

is a distribution which does not arise from either a locally integrable function or a Radon measure. (Note that $1/x$ is not a locally integrable function!)

Exercise 1.13.13 (Distributional interpretations of $1/|x|$)**.** Let $d = 1$. For any $r > 0$, show that the functional λ_r defined by the formula

$$\langle f, \lambda_r \rangle := \int_{|x|<r} \frac{f(x) - f(0)}{|x|}\, dx + \int_{|x|\geq r} \frac{f(x)}{|x|}\, dx$$

is a distribution that does not arise from either a locally integrable function or a Radon measure. Note that any two such functionals $\lambda_r, \lambda_{r'}$ differ by a constant multiple of the Dirac delta distribution.

Exercise 1.13.14. A distribution λ is said to be *real* if $\langle f, \lambda \rangle$ is real for every real-valued test function f. Show that every distribution λ can be uniquely expressed as $\mathrm{Re}(\lambda) + i\,\mathrm{Im}(\lambda)$ for some real distributions $\mathrm{Re}(\lambda), \mathrm{Im}(\lambda)$.

Exercise 1.13.15. A distribution λ is said to be *non-negative* if $\langle f, \lambda \rangle$ is non-negative for every non-negative test function f. Show

that a distribution is non-negative if and only if it is a non-negative Radon measure. (*Hint*: use the Riesz representation theorem and Exercise 1.13.1(iv).) Note that this implies that the analogue of the *Jordan decomposition* fails for distributions; any distribution which is not a Radon measure will not be the difference of non-negative distributions.

We will now extend various operations on locally integrable functions or Radon measures to distributions by arguing by analogy. (Shortly we will give a more formal approach, based on density.)

We begin with the operation of multiplying a distribution λ by a smooth function $h : \mathbf{R}^d \to \mathbf{C}$. Observe that

$$\langle f, gh \rangle = \langle f\overline{h}, g \rangle$$

for all test functions f, g, h. Inspired by this formula, we define the product $\lambda h = h\lambda$ of a distribution with a smooth function by setting

$$\langle f, \lambda h \rangle := \langle f\overline{h}, \lambda \rangle$$

for all test functions f. It is easy to see (e.g. using Exercise 1.13.4(vi)) that this defines a distribution λh, and that this operation is compatible with existing definitions of products between a locally integrable function (or Radon measure) with a smooth function. It is important that h is smooth (and not merely, say, continuous) because one needs the product of a test function f with $\overline{h}$ to still be a test function.

Exercise 1.13.16. Let $d = 1$. Establish the identity

$$\delta f = f(0)\delta$$

for any smooth function f. In particular,

$$\delta x = 0$$

where we abuse notation slightly and write x for the identity function $x \mapsto x$. Conversely, if λ is a distribution such that

$$\lambda x = 0,$$

show that λ is a constant multiple of δ. (*Hint*: Use the identity $f(x) = f(0) + x \int_0^1 f'(tx)\, dt$ to write $f(x)$ as the sum of $f(0)\psi$ and x times a test function for any test function f, where ψ is a fixed test function equalling 1 at the origin.)

Remark 1.13.2. Even though distributions are not, strictly speaking, functions, it is often useful heuristically to view them as such, thus for instance one might write a distributional identity such as $\delta x = 0$ suggestively as $\delta(x)x = 0$. Another useful (and rigorous) way to view such identities is to write distributions such as δ as a limit of approximations to the identity ψ_n, and show that the relevant identity becomes true in the limit; thus, for instance, to show that $\delta x = 0$, one can show that $\psi_n x \to 0$ in the sense of distributions as $n \to \infty$. (In fact, $\psi_n x$ converges to zero in the L^1 norm.)

Exercise 1.13.17. Let $d = 1$. With the distribution $\mathrm{p.v.}\, \frac{1}{x}$ from Exercise 1.13.12, show that $(\mathrm{p.v.}\, \frac{1}{x})x$ is equal to 1. With the distributions λ_r from Exercise 1.13.13, show that $\lambda_r x = \mathrm{sgn}$, where sgn is the *signum function*.

A distribution λ is said to be *supported* in a closed set K in $\langle f, \lambda\rangle = 0$ for all f that vanish on an open neighbourhood of K. The intersection of all K that λ is supported on is denoted $\mathrm{supp}(\lambda)$ and is referred to as the *support* of the distribution; this is the smallest closed set that λ is supported on. Thus, for instance, the Dirac delta function is supported on $\{0\}$, as are all derivatives of that function. (Note here that it is important that f vanish on a *neighbourhood* of K, rather than merely vanishing on K itself; for instance, in one dimension, there certainly exist test functions f that vanish at 0 but nevertheless have a non-zero inner product with δ'.)

Exercise 1.13.18. Show that every distribution is the limit of a sequence of compactly supported distributions (using the weak-* topology, of course). (*Hint*: Approximate a distribution λ by the truncated distributions $\lambda\eta_n$ for some smooth cutoff functions η_n constructed using Exercise 1.13.1(iii).)

In a similar spirit, we can convolve a distribution λ by an absolutely integrable, compactly supported function $h \in L^1(\mathbf{R}^d)$. From Fubini's theorem we observe the formula

$$\langle f, g * h\rangle = \langle f * \tilde{h}, g\rangle$$

for all test functions f, g, h, where $\tilde{h}(x) := \overline{h(-x)}$. Inspired by this formula, we define the convolution $\lambda * h = h * \lambda$ of a distribution

with an absolutely integrable, compactly supported function by the formula

$$\langle f, \lambda * h \rangle := \langle f * \tilde{h}, \lambda \rangle \tag{1.114}$$

for all test functions f. This gives a well-defined distribution λh (thanks to Exercise 1.13.4(vii)) which is compatible with previous notions of convolution.

Example 1.13.3. One has $\delta * f = f * \delta = f$ for all test functions f. In one dimension, we have $\delta' * f = f'$ (why?), thus differentiation can be viewed as convolution with a distribution.

A remarkable fact about convolutions of two functions $f * g$ is that they inherit the regularity of the *smoother* of the two factors f, g (in contrast to products fg, which tend to inherit the regularity of the *rougher* of the two factors). (This disparity can be also be seen by contrasting the identity $\nabla(f * g) = (\nabla f) * g = f * (\nabla g)$ with the identity $\nabla(fg) = (\nabla f)g + f(\nabla g)$.) In the case of convolving distributions with test functions, this phenomenon is manifested as follows:

Lemma 1.13.4. *Let $\lambda \in C_c^\infty(\mathbf{R}^d)^*$ be a distribution, and let $h \in C_c^\infty(\mathbf{R}^d)$ be a test function. Then $\lambda * h$ is equal to a smooth function.*

Proof. If λ were itself a smooth function, then one could easily verify the identity

$$\lambda * h(x) = \overline{\langle h_x, \lambda \rangle} \tag{1.115}$$

where $h_x(y) := \overline{h}(x - y)$. As h is a test function, it is easy to see that h_x varies smoothly in x in any C^k norm (indeed, it has Taylor expansions to any order in such norms) and so the right-hand side is a smooth function of x. So it suffices to verify the identity (1.115). As distributions are defined against test functions f, it suffices to show that

$$\langle f, \lambda * h \rangle = \int_{\mathbf{R}^d} f(x) \langle h_x, \lambda \rangle \, dx.$$

On the other hand, we have from (1.114) that

$$\langle f, \lambda * h \rangle = \langle f * \tilde{h}, \lambda \rangle = \langle \int_{\mathbf{R}^d} f(x) h_x \, dx, \lambda \rangle.$$

So the only issue is to justify the interchange of integral and inner product:

$$\int_{\mathbf{R}^d} f(x)\langle h_x, \lambda\rangle \; dx = \langle \int_{\mathbf{R}^d} f(x) h_x \; dx, \lambda\rangle.$$

Certainly, (from the compact support of f) any Riemann sum can be interchanged with the inner product:

$$\sum_n f(x_n)\langle h_{x_n}, \lambda\rangle \Delta x = \langle \sum_n f(x_n) h_{x_n} \Delta x, \lambda\rangle,$$

where x_n ranges over some lattice and Δx is the volume of the fundamental domain. A modification of the argument that shows convergence of the Riemann integral for smooth, compactly supported functions then works here and allows one to take limits; we omit the details. □

This has an important corollary:

Lemma 1.13.5. *Every distribution is the limit of a sequence of test functions. In particular, $C_c^\infty(\mathbf{R}^d)$ is dense in $C_c^\infty(\mathbf{R}^d)^*$.*

Proof. By Exercise 1.13.18, it suffices to verify this for compactly supported distributions λ. We let ϕ_n be a sequence of approximations to the identity. By Exercise 1.13.5(iii) and (1.114), we see that $\lambda * \phi_n$ converges in the sense of distributions to λ. By Lemma 1.13.4, $\lambda * \phi_n$ is a smooth function; as λ and ϕ_n are both compactly supported, $\lambda * \phi_n$ is compactly supported also. The claim follows. □

Because of this lemma, we can formalise the previous procedure of extending operations that were previously defined on test functions, to distributions, provided that these operations were continuous in distributional topologies. However, we shall continue to proceed by analogy as it requires fewer verifications in order to motivate the definition.

Exercise 1.13.19. Another consequence of Lemma 1.13.4 is that it allows one to extend the definition (1.114) of convolution to the case when h is not an integrable function of compact support, but is instead merely a distribution of compact support. Adopting this convention, show that convolution of distributions of compact support

is both commutative and associative. (*Hint*: this can either be done directly, or by carefully taking limits using Lemma 1.13.5.)

The next operation we will introduce is that of differentiation. An integration by parts reveals the identity

$$\langle f, \frac{\partial}{\partial x_j} g \rangle = -\langle \frac{\partial}{\partial x_j} f, g \rangle$$

for any test functions f, g and $j = 1, \ldots, d$. Inspired by this, we define the (distributional) partial derivative $\frac{\partial}{\partial x_j}\lambda$ of a distribution λ by the formula

$$\langle f, \frac{\partial}{\partial x_j} \lambda \rangle := -\langle \frac{\partial}{\partial x_j} f, \lambda \rangle.$$

This can be verified to still be a distribution, and by Exercise 1.13.4(vi), the operation of differentiation is a continuous one on distributions. More generally, given any linear differential operator P with smooth coefficients, one can define $P\lambda$ for a distribution λ by the formula

$$\langle f, P\lambda \rangle := \langle P^* f, \lambda \rangle$$

where P^* is the adjoint differential operator P, which can be defined implicitly by the formula

$$\langle f, Pg \rangle = \langle P^* f, g \rangle$$

for test functions f, g, or more explicitly by replacing all coefficients with complex conjugates, replacing each partial derivative $\frac{\partial}{\partial x_j}$ with its negative, and reversing the order of operations (thus for instance the adjoint of the first-order operator $a(x)\frac{d}{dx} : f \mapsto af'$ would be $-\frac{d}{dx}a(x) : f \mapsto -(af)'$).

Example 1.13.6. The distribution δ' defined in Exercise 1.13.11 is the derivative $\frac{d}{dx}\delta$ of δ, as defined by the above formula.

Many of the identities one is used to in classical calculus extend to the distributional setting (as one would already expect from Lemma 1.13.5). For instance:

Exercise 1.13.20 (Product rule)**.** Let $\lambda \in C_c^\infty(\mathbf{R}^d)^*$ be a distribution, and let $f : \mathbf{R}^d \to \mathbf{C}$ be smooth. Show that

$$\frac{\partial}{\partial x_j}(\lambda f) = (\frac{\partial}{\partial x_j}\lambda) f + \lambda(\frac{\partial}{\partial x_j} f)$$

for all $j = 1, \ldots, d$.

Exercise 1.13.21. Let $d = 1$. Show that $\delta' x = -\delta$ in three different ways:

- Directly from the definitions;
- using the product rule;
- Writing δ as the limit of approximations ψ_n to the identity.

Exercise 1.13.22. Let $d = 1$.

(i) Show that if λ is a distribution and $n \geq 1$ is an integer, then $\lambda x^n = 0$ if and only if is a linear combination of δ and its first $n-1$ derivatives $\delta', \delta'', \ldots, \delta^{(n-1)}$.

(ii) Show that a distribution λ is supported on $\{0\}$ if and only if it is a linear combination of δ and finitely many of its derivatives.

(iii) Generalise (ii) to the case of general dimension d (where of course one now uses partial derivatives instead of derivatives).

Exercise 1.13.23. Let $d = 1$.

- Show that the derivative of the *Heaviside function* $1_{[0,+\infty)}$ is equal to δ.
- Show that the derivative of the signum function $\mathrm{sgn}(x)$ is equal to 2δ.
- Show that the derivative of the locally integrable function $\log|x|$ is equal to $\mathrm{p.v.}\, \frac{1}{x}$.
- Show that the derivative of the locally integrable function $\log|x|\,\mathrm{sgn}(x)$ is equal to the distribution λ_1 from Exercise 1.13.13.
- Show that the derivative of the locally integrable function $|x|$ is the locally integrable function $\mathrm{sgn}(x)$.

If a locally integrable function has a distributional derivative which is also a locally integrable function, we refer to the latter as the *weak derivative* of the former. Thus, for instance, the weak derivative of $|x|$ is $\mathrm{sgn}(x)$ (as one would expect), but $\mathrm{sgn}(x)$ does not have a

weak derivative (despite being (classically) differentiable almost everywhere), because the distributional derivative 2δ of this function is not itself a locally integrable function. Thus weak derivatives differ in some respects from their classical counterparts, though of course the two concepts agree for smooth functions.

Exercise 1.13.24. Let $d \geq 1$. Show that for any $1 \leq i, j \leq d$, and any distribution $\lambda \in C_c^\infty(\mathbf{R}^d)^*$, we have $\frac{\partial}{\partial x_i}\frac{\partial}{\partial x_j}\lambda = \frac{\partial}{\partial x_j}\frac{\partial}{\partial x_i}\lambda$, thus weak derivatives commute with each other. (This is in contrast to classical derivatives, which can fail to commute for non-smooth functions; for instance, $\frac{\partial}{\partial x}\frac{\partial}{\partial y}\frac{xy^3}{x^2+y^2} \neq \frac{\partial}{\partial y}\frac{\partial}{\partial x}\frac{xy^3}{x^2+y^2}$ at the origin $(x,y) = 0$, despite both derivatives being defined. More generally, weak derivatives tend to be less pathological than classical derivatives, but of course the downside is that weak derivatives do not always have a classical interpretation as a limit of a Newton quotient.)

Exercise 1.13.25. Let $d = 1$, and let $k \geq 0$ be an integer. Let us say that a compactly supported distribution $\lambda \in C_c^\infty(\mathbf{R})^*$ *has of order at most* k if the functional $f \mapsto \langle f, \lambda\rangle$ is continuous in the C^k norm. Thus, for instance, δ has order at most 0, and δ' has order at most 1, and every compactly supported distribution is of order at most k for some sufficiently large k.

- Show that if λ is a compactly supported distribution of order at most 0, then it is a compactly supported Radon measure.
- Show that if λ is a compactly supported distribution of order at most k, then λ' has order at most $k+1$.
- Conversely, if λ is a compactly supported distribution of order $k+1$, then we can write $\lambda = \rho' + \nu$ for some compactly supported distributions of order k. (*Hint*: one has to "dualise" the fundamental theorem of calculus, and then apply smooth cutoffs to recover compact support.)
- Show that every compactly supported distribution can be expressed as a finite linear combination of (distributional) derivatives of compactly supported Radon measures.
- Show that every compactly supported distribution can be expressed as a finite linear combination of (distributional) derivatives of functions in $C_0^k(\mathbf{R})$, for any fixed k.

We now set out some other operations on distributions. If we define the translation $\tau_x f$ of a test function f by a shift $x \in \mathbf{R}^d$ by the formula $\tau_x f(y) := f(y-x)$, then we have

$$\langle f, \tau_x g \rangle = \langle \tau_{-x} f, g \rangle$$

for all test functions f, g, so it is natural to define the translation $\tau_x \lambda$ of a distribution λ by the formula

$$\langle f, \tau_x \lambda \rangle := \langle \tau_{-x} f, \lambda \rangle.$$

Next, we consider linear changes of variable.

Exercise 1.13.26 (Linear changes of variable)**.** Let $d \geq 1$, and let $L : \mathbf{R}^d \to \mathbf{R}^d$ be a linear transformation. Given a distribution $\lambda \in C_c^\infty(\mathbf{R}^d)^*$, let $\lambda \circ L$ be the distribution given by the formula

$$\langle f, \lambda \circ L \rangle := \frac{1}{|\det L|} \langle f \circ L^{-1}, \lambda \rangle$$

for all test functions f. (How would one motivate this formula?)

- Show that $\delta \circ L = \frac{1}{|\det L|} \delta$ for all linear transformations L.
- If $d = 1$, show that $\mathrm{p.v.}\, \frac{1}{x} \cdot L = \frac{1}{|\det L|} \mathrm{p.v.}\, \frac{1}{x}$ for all linear transformations L.
- Conversely, if $d = 1$ and λ is a distribution such that $\lambda \cdot L = \frac{1}{|\det L|} \lambda$ for all linear transformations L. (*Hint*: first show that there exists a constant c such that $\langle f, \lambda \rangle = c \int_0^\infty \frac{f(x)}{x}\, dx$ whenever f is a bump function supported in $(0, +\infty)$. To show this, approximate f by the function
$$\int_{-\infty}^\infty f(e^t x) \psi_n(t)\, dt = \int_0^\infty \frac{f(y)}{y} \psi_n(\log \frac{x}{y}) 1_{x>0}\, dy$$
for ψ_n an approximation to the identity.)

Remark 1.13.7. One can also compose distributions with *diffeomorphisms*. However, things become much more delicate if the map one is composing with contains stationary points; for instance, in one dimension, one cannot meaningfully make sense of $\delta(x^2)$ (the composition of the Dirac delta distribution with $x \mapsto x^2$); this can be seen by first noting that for an approximation ψ_n to the identity, $\psi_n(x^2)$ does not converge to a limit in the distributional sense.

Exercise 1.13.27 (Tensor product of distributions)**.** Let $d, d' \geq 1$ be integers. If $\lambda \in C_c^\infty(\mathbf{R}^d)^*$ and $\rho \in C_c^\infty(\mathbf{R}^{d'})^*$ are distributions, show that there is a unique distribution $\lambda \otimes \rho \in C_c^\infty(\mathbf{R}^{d+d'})^*$ with the property that

$$\langle f \otimes g, \lambda \otimes \rho \rangle = \langle f, \lambda \rangle \langle g, \rho \rangle \tag{1.116}$$

for all test functions $f \in C_c^\infty(\mathbf{R}^d)$, $g \in C_c^\infty(\mathbf{R}^{d'})$, where $f \otimes g : C_c^\infty(\mathbf{R}^{d+d'})$ is the tensor product $f \otimes g(x, x') := f(x)g(x')$ of f and g. (*Hint*: like many other constructions of tensor products, this is rather intricate. One way is to start by fixing two cutoff functions ψ, ψ' on $\mathbf{R}^d, \mathbf{R}^{d'}$ respectively, and define $\lambda \otimes \rho$ on modulated test functions $e^{2\pi i \xi \cdot x} e^{2\pi i \xi' \cdot x} \psi(x)\psi'(x')$ for various frequencies ξ, ξ', and then use Fourier series to define $\lambda \otimes \rho$ on $F(x, x')\psi(x)\psi'(x')$ for smooth F. Then show that these definitions of $\lambda \otimes \rho$ are compatible for different choices of ψ, ψ' and can be glued together to form a distribution; finally, go back and verify (1.116).)

We close this section with one caveat. Despite the many operations that one can perform on distributions, there are two types of operations which cannot, in general, be defined on arbitrary distributions (at least while remaining in the class of distributions):

- Nonlinear operations (e.g. taking the absolute value of a distribution); or
- Multiplying a distribution by anything rougher than a smooth function.

Thus, for instance, there is no meaningful way to interpret the square δ^2 of the Dirac delta function as a distribution. This is perhaps easiest to see using an approximation ψ_n to the identity: ψ_n converges to δ in the sense of distributions, but ψ_n^2 does not converge to anything (the integral against a test function that does not vanish at the origin will go to infinity as $n \to \infty$). For similar reasons, one cannot meaningfully interpret the absolute value $|\delta'|$ of the derivative of the delta function. (One also cannot multiply δ by $\mathrm{sgn}(x)$ - why?)

Exercise 1.13.28. Let X be a normed vector space which contains $C_c^\infty(\mathbf{R}^d)$ as a dense subspace (and such that the inclusion of $C_c^\infty(\mathbf{R}^d)$ to X is continuous). The adjoint (or transpose) of this inclusion map

is then an injection from X^* to the space of distributions $C_c^\infty(\mathbf{R}^d)^*$; thus X^* can be viewed as a subspace of the space of distributions.

- Show that the closed unit ball in X^* is also closed in the space of distributions.
- Conclude that any distributional limit of a bounded sequence in $L^p(\mathbf{R}^d)$ for $1 < p \leq \infty$, is still in $L^p(\mathbf{R}^d)$.
- Show that the previous claim fails for $L^1(\mathbf{R}^d)$, but holds for the space $M(\mathbf{R}^d)$ of finite measures.

1.13.3. Tempered distributions. The list of operations one can define on distributions has one major omission - the *Fourier transform* $\mathcal{F}$. Unfortunately, one cannot easily define the Fourier transform for all distributions. One can see this as follows. From *Plancherel's theorem* one has the identity

$$\langle f, \mathcal{F}g \rangle = \langle \mathcal{F}^* f, g \rangle$$

for test functions f, g, so one would like to define the Fourier transform $\mathcal{F}\lambda = \hat{\lambda}$ of a distribution λ by the formula

$$\langle f, \mathcal{F}\lambda \rangle := \langle \mathcal{F}^* f, \lambda \rangle. \tag{1.117}$$

Unfortunately this does not quite work, because the adjoint Fourier transform $\mathcal{F}^*$ of a test function is not a test function, but is instead just a Schwartz function. (Indeed, by Exercise 1.12.42, it is not possible to find a non-trivial test function whose Fourier transform is again a test function.) To address this, we need to work with a slightly smaller space than that of all distributions, namely those of *tempered* distributions:

Definition 1.13.8 (Tempered distributions)**.** A tempered distribution is a continuous linear functional $\lambda : f \mapsto \langle f, \lambda \rangle$ on the Schwartz space $\mathcal{S}(\mathbf{R}^d)$ (with the topology given by Exercise 1.12.25), i.e. an element of $\mathcal{S}(\mathbf{R}^d)^*$.

Since $C_c^\infty(\mathbf{R}^d)$ embeds continuously into $\mathcal{S}(\mathbf{R}^d)$ (with a dense image), we see that the space of tempered distributions can be embedded into the space of distributions. However, not every distribution is tempered:

Example 1.13.9. The distribution e^x is not tempered. Indeed, if ψ is a bump function, observe that the sequence of functions $e^{-n}\psi(x-n)$ converges to zero in the Schwartz space topology, but $\langle e^{-n}\psi(x-n), e^x\rangle$ does not go to zero, and so this distribution does not correspond to a tempered distribution.

On the other hand, distributions which avoid this sort of exponential growth, and instead only grow polynomially, tend to be tempered:

Exercise 1.13.29. Show that any Radon measure μ which is of *polynomial growth* in the sense that $|\mu|(B(0,R)) \leq CR^k$ for all $R \geq 1$ and some constants $C, k > 0$, where $B(0,R)$ is the ball of radius R centred at the origin in $\mathbf{R}^d$, is tempered.

Remark 1.13.10. As a zeroth approximation, one can roughly think of "tempered" as being synonymous with "polynomial growth". However, this is not strictly true: for instance, the (weak) derivative of a function of polynomial growth will still be tempered, but need not be of polynomial growth (for instance, the derivative $e^x \cos(e^x)$ of $\sin(e^x)$ is a tempered distribution, despite having exponential growth). While one can eventually describe which distributions are tempered by measuring their "growth" in both physical space and in frequency space, we will not do so here.

Most of the operations that preserve the space of distributions, also preserve the space of tempered distributions. For instance:

Exercise 1.13.30.

- Show that any derivative of a tempered distribution is again a tempered distribution.
- Show that and any convolution of a tempered distribution with a compactly supported distribution is again a tempered distribution.
- Show that if f is a measurable function which is *rapidly decreasing* in the sense that $|x|^k f(x)$ is an $L^\infty(\mathbf{R}^d)$ function for each $k = 0, 1, 2, \ldots$, then a convolution of a tempered distribution with f can be defined, and is again a tempered distribution.
- Show that if f is a smooth function such that f and all its derivatives have *at most polynomial growth* (thus for each

$j \geq 0$ there exists $C, k \geq 0$ such that $|\nabla^j f(x)| \leq C(1+|x|)^k$ for all $x \in \mathbf{R}^d$) then the product of a tempered distribution with f is again a tempered distribution. Give a counterexample to show that this statement fails if the polynomial growth hypotheses are dropped.

- Show that the translate of a tempered distribution is again a tempered distribution.

But we can now add a new operation to this list using (1.117): as the Fourier transform $\mathcal{F}$ maps Schwartz functions continuously to Schwartz functions, it also continuously maps the space of tempered distributions to itself. One can also define the inverse Fourier transform $\mathcal{F}^* = \mathcal{F}^{-1}$ on tempered distributions in a similar manner.

It is not difficult to extend many of the properties of the Fourier transform from Schwartz functions to distributions. For instance:

Exercise 1.13.31. Let $\lambda \in \mathcal{S}(\mathbf{R}^d)^*$ be a tempered distribution, and let $f \in \mathcal{S}(\mathbf{R}^d)$ be a Schwartz function.

- (Inversion formula) Show that $\mathcal{F}^*\mathcal{F}\lambda = \mathcal{F}\mathcal{F}^*\lambda = \lambda$.
- (Multiplication intertwines with convolution) Show that $\mathcal{F}(\lambda f) = (\mathcal{F}\lambda) * (\mathcal{F}f)$ and $\mathcal{F}(\lambda * f) = (\mathcal{F}\lambda)(\mathcal{F}f)$.
- (Translation intertwines with modulation) For any $x_0 \in \mathbf{R}^d$, show that $\mathcal{F}(\tau_{x_0}\lambda) = e_{-x_0}\mathcal{F}\lambda$, where $e_{-x_0}(\xi) := e^{-2\pi i \xi \cdot x_0}$. Similarly, show that for any $\xi_0 \in \mathbf{R}^d$, one has $\mathcal{F}(e_{\xi_0}\lambda) = \tau_{\xi_0}\mathcal{F}\lambda$.
- (Linear transformations) For any invertible linear transformation $L : \mathbf{R}^d \to \mathbf{R}^d$, show that $\mathcal{F}(\lambda \circ L) = \frac{1}{|\det L|}(\mathcal{F}\lambda) \circ (L^*)^{-1}$.
- (Differentiation intertwines with polynomial multiplication) For any $1 \leq j \leq d$, show that $\mathcal{F}(\frac{\partial}{\partial x_j}\lambda) = 2\pi i \xi_j \mathcal{F}\lambda$, where x_j and ξ_j is the j^{th} coordinate function in physical space and frequency space respectively, and similarly $\mathcal{F}(-2\pi i x_j \lambda) = \frac{\partial}{\partial \xi_j}\mathcal{F}\lambda$.

Exercise 1.13.32. Let $d \geq 1$.

- (Inversion formula) Show that $\mathcal{F}\delta = 1$ and $\mathcal{F}1 = \delta$.

- (Orthogonality) Let V be a subspace of $\mathbf{R}^d$, and let μ be Lebesgue measure on V. Show that $\mathcal{F}\mu$ is Lebesgue measure on the orthogonal complement $V^\perp$ of V. (Note that this generalises the previous exercise.)
- (Poisson summation formula) Let $\sum_{k\in\mathbf{Z}^d} \tau_k\delta$ be the distribution
$$\langle f, \sum_{k\in\mathbf{Z}^d} \tau_k\delta\rangle := \sum_{k\in\mathbf{Z}^d} f(k).$$
Show that this is a tempered distribution which is equal to its own Fourier transform.

One can use these properties of tempered distributions to start solving constant-coefficient PDE. We first illustrate this by an ODE example, showing how the formal symbolic calculus for solving such ODE that you may have seen as an undergraduate, can now be (sometimes) justified using tempered distributions.

Exercise 1.13.33. Let $d = 1$, let a, b be real numbers, and let D be the operator $D = \frac{d}{dx}$.

- If $a \neq b$, use the Fourier transform to show that all tempered distribution solutions to the ODE $(D-ia)(D-ib)\lambda = 0$ are of the form $\lambda = Ae^{iax} + Be^{ibx}$ for some constants A, B.
- If $a = b$, show that all tempered distribution solutions to the ODE $(D-ia)(D-ib)\lambda = 0$ are of the form $\lambda = Ae^{iax} + Bxe^{iax}$ for some constants A, B.

Remark 1.13.11. More generally, one can solve any homogeneous constant-coefficient ODE using tempered distributions and the Fourier transform so long as the roots of the *characteristic polynomial* are purely imaginary. In all other cases, solutions can grow exponentially as $x \to +\infty$ or $x \to -\infty$ and so are not tempered. There are other theories of generalised functions that can handle these objects (e.g. *hyperfunctions*) but we will not discuss them here.

Now we turn to PDE. To illustrate the method, let us focus on solving *Poisson's equation*

$$\Delta u = f \tag{1.118}$$

in $\mathbf{R}^d$, where f is a Schwartz function and u is a distribution, where $\Delta = \sum_{j=1}^d \frac{\partial^2}{\partial x_j^2}$ is the *Laplacian*. (In some texts, particularly those using spectral analysis, the Laplacian is occasionally defined instead as $-\sum_{j=1}^d \frac{\partial^2}{\partial x_j^2}$, to make it positive semi-definite, but we will eschew that sign convention here, though of course the theory is only changed in a trivial fashion if one adopts it.)

We first settle the question of uniqueness:

Exercise 1.13.34. Let $d \geq 1$. Using the Fourier transform, show that the only tempered distributions $\lambda \in \mathcal{S}(\mathbf{R}^d)^*$ which are harmonic (by which we mean that $\Delta\lambda = 0$ in the sense of distributions) are the harmonic polynomials. (*Hint*: use Exercise 1.13.22.) Note that this generalises *Liouville's theorem*. There are of course many other harmonic functions than the harmonic polynomials, e.g. $e^x \cos(y)$, but such functions are not tempered distributions.

From the above exercise, we know that the solution u to (1.118), if tempered, is defined up to harmonic polynomials. To find a solution, we observe that it is enough to find a *fundamental solution*, i.e. a tempered distribution K solving the equation

$$\Delta K = \delta.$$

Indeed, if one then convolves this equation with the Schwartz function f, and uses the identity $(\Delta K) * f = \Delta(K * f)$ (which can either be seen directly, or by using Exercise 1.13.31), we see that $u = K * f$ will be a tempered distribution solution to (1.118) (and all the other solutions will equal this solution plus a harmonic polynomial). So, it is enough to locate a fundamental solution K. We can take Fourier transforms and rewrite this equation as

$$-4\pi^2|\xi|^2 \hat{K}(\xi) = 1$$

(here we are treating the tempered distribution $\hat{K}$ as a function to emphasise that the dependent variable is now ξ). It is then natural to propose to solve this equation as

$$\hat{K}(\xi) = \frac{1}{-4\pi^2|\xi|^2}, \tag{1.119}$$

though this may not be the unique solution (for instance, one is free to modify K by a multiple of the Dirac delta function, cf. Exercise 1.13.16).

A short computation in polar coordinates shows that $\frac{1}{-4\pi^2|\xi|^2}$ is locally integrable in dimensions $d \geq 3$, so the right-hand side of (1.119) makes sense. To then compute K explicitly, we have from the distributional inversion formula that

$$K = \frac{-1}{4\pi^2} \mathcal{F}^* |\xi|^{-2}$$

so we now need to figure out what the Fourier transform of a negative power of $|x|$ (or the adjoint Fourier transform of a negative power of $|\xi|$) is.

Let us work formally at first, and consider the problem of computing the Fourier transform of the function $|x|^{-\alpha}$ in $\mathbf{R}^d$ for some exponent α. A direct attack, based on evaluating the (formal) Fourier integral

$$\widehat{|x|^{-\alpha}}(\xi) = \int_{\mathbf{R}^d} |x|^{-\alpha} e^{-2\pi i \xi \cdot x}\ dx \tag{1.120}$$

does not seem to make much sense (the integral is not absolutely integrable), although a change of variables (or dimensional analysis) heuristic can at least lead to the prediction that the integral (1.120) should be some multiple of $|\xi|^{\alpha-d}$. But which multiple should it be? To continue the formal calculation, we can write the non-integrable function $|x|^{-\alpha}$ as an average of integrable functions whose Fourier transforms are already known. There are many such functions that one could use here, but it is natural to use Gaussians, as they have a particularly pleasant Fourier transform, namely

$$\widehat{e^{-\pi t^2 |x|^2}}(\xi) = t^d e^{-\pi |\xi|^2/t^2}$$

for $t > 0$ (see Exercise 1.12.32). To get from Gaussians to $|x|^{-\alpha}$, one can observe that $|x|^{-\alpha}$ is invariant under the scaling $f(x) \mapsto t^\alpha f(tx)$ for $t > 0$. Thus, it is natural to average the standard Gaussian $e^{-\pi|x|^2}$ with respect to this scaling, thus producing the function $t^\alpha e^{-\pi t^2 |x|^2}$, then integrate with respect to the multiplicative Haar measure $\frac{dt}{t}$. A

straightforward change of variables then gives the identity

$$\int_0^\infty t^\alpha e^{-\pi t^2|x|^2} \frac{dt}{t} = \frac{1}{2}\pi^{-\alpha/2}|x|^{-\alpha}\Gamma(\alpha/2)$$

where

$$\Gamma(s) := \int_0^\infty t^s e^{-t} \frac{dt}{t}$$

is the *Gamma function*. If we formally take Fourier transforms of this identity, we obtain

$$\int_0^\infty t^\alpha t^{-d} e^{-\pi|x|^2/t^2} \frac{dt}{t} = \frac{1}{2}\pi^{-\alpha/2}\widehat{|x|^{-\alpha}}(\xi)\Gamma(\alpha/2).$$

Another change of variables shows that

$$\int_0^\infty t^\alpha t^{-d} e^{-\pi|x|^2/t^2} \frac{dt}{t} = \frac{1}{2}\pi^{-(d-\alpha)/2}|\xi|^{-(d-\alpha)}\Gamma((d-\alpha)/2)$$

and so we conclude (formally) that

$$\widehat{|x|^{-\alpha}}(\xi) = \frac{\pi^{-(d-\alpha)/2}\Gamma((d-\alpha)/2)}{\pi^{-\alpha/2}\Gamma(\alpha/2)}|\xi|^{-(d-\alpha)} \tag{1.121}$$

thus solving the problem of what the constant multiple of $|\xi|^{-(d-\alpha)}$ should be.

Exercise 1.13.35. Give a rigorous proof of (1.121) for $0 < \alpha < d$ (when both sides are locally integrable) in the sense of distributions. (*Hint*: basically, one needs to test the entire formal argument against an arbitrary Schwartz function.) The identity (1.121) can in fact be continued meromorphically in α, but the interpretation of distributions such as $|x|^{-\alpha}$ when $|x|^{-\alpha}$ is not locally integrable is somewhat complicated (cf. Exercise 1.13.12) and will not be discussed here.

Specialising back to the current situation with $d = 3, \alpha = 2$, and using the standard identities

$$\Gamma(n) = (n-1)!; \quad \Gamma(\frac{1}{2}) = \sqrt{\pi}$$

we see that

$$\widehat{\frac{1}{|x|^2}}(\xi) = \pi|\xi|^{-1}$$

and similarly

$$\mathcal{F}^* \frac{1}{|\xi|^2} = \pi|x|^{-1}$$

and so from (1.119) we see that one choice of the fundamental solution K is the *Newton potential*

$$K = \frac{-1}{4\pi|x|},$$

leading to an explicit (and rigorously derived) solution

$$u(x) := f * K(x) = -\frac{1}{4\pi} \int_{\mathbf{R}^3} \frac{f(y)}{|x-y|}\, dy \tag{1.122}$$

to the Poisson equation (1.118) in $d = 3$ for Schwartz functions f. (This is not quite the only fundamental solution K available; one can add a harmonic polynomial to K, which will end up adding a harmonic polynomial to u, since the convolution of a harmonic polynomial with a Schwartz function is easily seen to still be harmonic.)

Exercise 1.13.36. Without using the theory of distributions, give an alternate (and still rigorous) proof that the function u defined in (1.122) solves (1.118) in $d = 3$.

Exercise 1.13.37.

- Show that for any $d \geq 3$, a fundamental solution K to the Poisson equation is given by the locally integrable function

$$K(x) = \frac{1}{d(d-2)\omega_d} \frac{1}{|x|^{d-2}},$$

  where $\omega_d = \pi^{d/2}/\Gamma(\frac{d}{2}+1)$ is the volume of the unit ball in d dimensions.
- Show that for $d = 1$, a fundamental solution is given by the locally integrable function $K(x) = |x|/2$.
- Show that for $d = 2$, a fundamental solution is given by the locally integrable function $K(x) = \frac{1}{2\pi} \log |x|$.

This we see that for the Poisson equation, $d = 2$ is a "critical" dimension, requiring a logarithmic correction to the usual formula.

Similar methods can solve other constant coefficient linear PDE. We give some standard examples in the exercises below.

Exercise 1.13.38. Let $d \geq 1$. Show that a smooth solution $u : \mathbf{R}^+ \times \mathbf{R}^d \to \mathbf{C}$ to the *heat equation* $\partial_t u = \Delta u$ with initial data

$u(0,x) = f(x)$ for some Schwartz function f is given by $u(t) = f * K_t$ for $t > 0$, where K_t is the *heat kernel*

$$K_t(x) = \frac{1}{(4\pi t)^{d/2}} e^{-|x-y|^2/4t}.$$

(This solution is unique assuming certain smoothness and decay conditions at infinity, but we will not pursue this issue here.)

Exercise 1.13.39. Let $d \geq 1$. Show that a smooth solution $u : \mathbf{R} \times \mathbf{R}^d \to \mathbf{C}$ to the *Schrödinger equation* $\partial_t u = i\Delta u$ with initial data $u(0,x) = f(x)$ for some Schwartz function f is given by $u(t) = f * K_t$ for $t \neq 0$, where K_t is the *Schrödinger kernel*[12]

$$K_t(x) = \frac{1}{(4\pi i t)^{d/2}} e^{i|x-y|^2/4t}$$

and we use the standard branch of the complex logarithm (with cut on the negative real axis) to define $(4\pi i t)^{d/2}$. (*Hint*: You may wish to investigate the Fourier transform of $e^{-z|\xi|^2}$, where z is a complex number with positive real part, and then let z approach the imaginary axis.)

Exercise 1.13.40. Let $d = 3$. Show that a smooth solution $u : \mathbf{R} \times \mathbf{R}^3 \to \mathbf{C}$ to the *wave equation* $-\partial_{tt} u + \Delta u$ with initial data $u(0,x) = f(x), \partial_t u(0,x) = g(x)$ for some Schwartz functions f is given by the formula

$$u(t) = f * \partial_t K_t + g * K_t$$

for $t \neq 0$, where K_t is the distribution

$$\langle f, K_t \rangle := \frac{t}{4\pi} \int_{S^2} f(t\omega)\, d\omega$$

where ω is Lebesgue measure on the sphere S^2, and the derivative $\partial_t K_t$ is defined in the Newtonian sense $\lim_{dt \to 0} \frac{K_{t+dt} - K_t}{dt}$, with the limit taken in the sense of distributions.

[12] The close similarity here with the heat kernel is a manifestation of *Wick rotation* in action. However, from an analytical viewpoint, the two kernels are very different. For instance, the convergence of $f * K_t$ to f as $t \to 0$ follows in the heat kernel case by the theory of approximations to the identity, whereas the convergence in the Schrödinger case is much more subtle, and is best seen via Fourier analysis.

Remark 1.13.12. The theory of (tempered) distributions is also highly effective for studying variable coefficient linear PDE, especially if the coefficients are fairly smooth, and particularly if one is primarily interested in the singularities of solutions to such PDE and how they propagate; here the Fourier transform must be augmented with more general transforms of this type, such as *Fourier integral operators*. A classic reference for this topic is [**Ho1990**]. For nonlinear PDE, subspaces of the space of distributions, such as *Sobolev spaces*, tend to be more useful; we will discuss these in the next section.

Notes. This lecture first appeared at `terrytao.wordpress.com/2009/04/19`. Thanks to Dale Roberts, Max Baroi, and an anonymous commenter for corrections.

1.14. Sobolev spaces

As discussed in previous sections, a function space norm can be viewed as a means to rigorously quantify various statistics of a function $f : X \to \mathbf{C}$. For instance, the "height" and "width" can be quantified via the $L^p(X,\mu)$ norms (and their relatives, such as the Lorentz norms $\|f\|_{L^{p,q}(X,\mu)}$). Indeed, if f is a step function $f = A1_E$, then the L^p norm of f is a combination $\|f\|_{L^p(X,\mu)} = |A|\mu(E)^{1/p}$ of the height (or amplitude) A and the width $\mu(E)$.

However, there are more features of a function f of interest than just its width and height. When the domain X is a Euclidean space $\mathbf{R}^d$ (or domains related to Euclidean spaces, such as open subsets of $\mathbf{R}^d$, or manifolds), then another important feature of such functions (especially in PDE) is the *regularity* of a function, as well as the related concept of the *frequency scale* of a function. These terms are not rigorously defined; but roughly speaking, regularity measures how smooth a function is (or how many times one can differentiate the function before it ceases to be a function), while the frequency scale of a function measures how quickly the function oscillates (and would be inversely proportional to the wavelength). One can illustrate this informal concept with some examples:

- Let $\phi \in C_c^\infty(\mathbf{R})$ be a test function that equals 1 near the origin, and N be a large number. Then the function $f(x) :=$

$\phi(x)\sin(Nx)$ oscillates at a wavelength of about $1/N$, and a frequency scale of about N. While f is, strictly speaking, a smooth function, it becomes increasingly less smooth in the limit $N \to \infty$; for instance, the derivative $f'(x) = \phi'(x)\sin(Nx) + N\phi(x)\cos(Nx)$ grows at a roughly linear rate as $N \to \infty$, and the higher derivatives grow at even faster rates. So this function does not really have any regularity in the limit $N \to \infty$. Note however that the height and width of this function is bounded uniformly in N; so regularity and frequency scale are independent of height and width.

- Continuing the previous example, now consider the function $g(x) := N^{-s}\phi(x)\sin(Nx)$, where $s \geq 0$ is some parameter. This function also has a frequency scale of about N. But now it has a certain amount of regularity, even in the limit $N \to \infty$; indeed, one easily checks that the k^{th} derivative of g stays bounded in N as long as $k \leq s$. So one could view this function as having "s degrees of regularity" in the limit $N \to \infty$.
- In a similar vein, the function $N^{-s}\phi(Nx)$ also has a frequency scale of about N, and can be viewed as having s degrees of regularity in the limit $N \to \infty$.
- The function $\phi(x)|x|^s 1_{x>0}$ also has about s degrees of regularity, in the sense that it can be differentiated up to s times before becoming unbounded. By performing a dyadic decomposition of the x variable, one can also decompose this function into components $\psi(2^n x)|x|^s$ for $n \geq 0$, where $\psi(x) := (\phi(x) - \phi(2x))1_{x>0}$ is a bump function supported away from the origin; each such component has frequency scale about 2^n and s degrees of regularity. Thus we see that the original function $\phi(x)|x|^s 1_{x>0}$ has a range of frequency scales, ranging from about 1 all the way to $+\infty$.
- One can of course concoct higher-dimensional analogues of these examples. For instance, the localised plane wave $\phi(x)\sin(\xi \cdot x)$ in $\mathbf{R}^d$, where $\phi \in C_c^\infty(\mathbf{R}^d)$ is a test function, would have a frequency scale of about $|\xi|$.

There are a variety of function space norms that can be used to capture frequency scale (or regularity) in addition to height and width. The most common and well-known examples of such spaces are the *Sobolev space norms* $\|f\|_{W^{s,p}(\mathbf{R}^d)}$, although there are a number of other norms with similar features (such as *Hölder norms*, *Besov norms*, and *Triebel-Lizorkin norms*). Very roughly speaking, the $W^{s,p}$ norm is like the L^p norm, but with "s additional degrees of regularity". For instance, in one dimension, the function $A\phi(x/R)\sin(Nx)$, where ϕ is a fixed test function and R, N are large, will have a $W^{s,p}$ norm of about $|A|R^{1/p}N^s$, thus combining the "height" $|A|$, the "width" R, and the "frequency scale" N of this function together. (Compare this with the L^p norm of the same function, which is about $|A|R^{1/p}$.)

To a large extent, the theory of the Sobolev spaces $W^{s,p}(\mathbf{R}^d)$ resembles their Lebesgue counterparts $L^p(\mathbf{R}^d)$ (which are as the special case of Sobolev spaces when $s = 0$), but with the additional benefit of being able to interact very nicely with (weak) derivatives: a first derivative $\frac{\partial f}{\partial x_j}$ of a function in an L^p space usually leaves all Lebesgue spaces, but a first derivative of a function in the Sobolev space $W^{s,p}$ will end up in another Sobolev space $W^{s-1,p}$. This compatibility with the differentiation operation begins to explain why Sobolev spaces are so useful in the theory of partial *differential* equations. Furthermore, the regularity parameter s in Sobolev spaces is not restricted to be a natural number; it can be any real number, and one can use *fractional* derivative or integration operators to move from one regularity to another. Despite the fact that most partial differential equations involve differential operators of integer order, fractional spaces are still of importance; for instance it often turns out that the Sobolev spaces which are *critical* (scale-invariant) for a certain PDE are of fractional order.

The *uncertainty principle* in Fourier analysis places a constraint between the width and frequency scale of a function; roughly speaking (and in one dimension for simplicity), the product of the two quantities has to be bounded away from zero (or to put it another way, a wave is always at least as wide as its wavelength). This constraint can

be quantified as the very useful *Sobolev embedding theorem*, which allows one to trade regularity for integrability: a function in a Sobolev space $W^{s,p}$ will automatically lie in a number of other Sobolev spaces $W^{\tilde{s},\tilde{p}}$ with $\tilde{s} < s$ and $\tilde{p} > p$; in particular, one can often embed Sobolev spaces into Lebesgue spaces. The trade is not reversible: one cannot start with a function with a lot of integrability and no regularity, and expect to recover regularity in a space of lower integrability. (One can already see this with the most basic example of Sobolev embedding, coming from the fundamental theorem of calculus. If a (continuously differentiable) function $f : \mathbf{R} \to \mathbf{R}$ has f' in $L^1(\mathbf{R})$, then we of course have $f \in L^\infty(\mathbf{R})$; but the converse is far from true.)

Plancherel's theorem reveals that Fourier-analytic tools are particularly powerful when applied to L^2 spaces. Because of this, the Fourier transform is very effective at dealing with the L^2-based Sobolev spaces $W^{s,2}(\mathbf{R}^d)$, often abbreviated $H^s(\mathbf{R}^d)$. Indeed, using the fact that the Fourier transform converts regularity to decay, we will see that the $H^s(\mathbf{R}^d)$ spaces are nothing more than Fourier transforms of weighted L^2 spaces, and in particular enjoy a Hilbert space structure. These Sobolev spaces, and in particular the *energy space* $H^1(\mathbf{R}^d)$, are of particular importance in any PDE that involves some sort of energy functional (this includes large classes of elliptic, parabolic, dispersive, and wave equations, and especially those equations connected to physics and/or geometry).

We will not fully develop the theory of Sobolev spaces here, as this would require the theory of *singular integrals*, which is beyond the scope of this course. There are of course many references for further reading, such as [**St1970**].

1.14.1. Hölder spaces. Throughout these notes, $d \geq 1$ is a fixed dimension.

Before we study Sobolev spaces, let us first look at the more elementary theory of *Hölder spaces* $C^{k,\alpha}(\mathbf{R}^d)$, which resemble Sobolev spaces but with the aspect of width removed (thus Hölder norms only measure a combination of height and frequency scale). One can define these spaces on many domains (for instance, the $C^{0,\alpha}$ norm can be

defined on any metric space) but we shall largely restrict attention to Euclidean spaces $\mathbf{R}^d$ for sake of concreteness.

We first recall the $C^k(\mathbf{R}^d)$ spaces, which we have already been implicitly using in previous lectures. The space $C^0(\mathbf{R}^d) = BC(\mathbf{R}^d)$ is the space of bounded continuous functions $f : \mathbf{R}^d \to \mathbf{C}$ on $\mathbf{R}^d$, with norm

$$\|f\|_{C^0(\mathbf{R}^d)} := \sup_{x \in \mathbf{R}^d} |f(x)| = \|f\|_{L^\infty(\mathbf{R}^d)}.$$

This norm gives C^0 the structure of a Banach space. More generally, one can then define the spaces $C^k(\mathbf{R}^d)$ for any non-negative integer k as the space of all functions which are k times continuously differentiable, with all derivatives of order k bounded, and whose norm is given by the formula

$$\|f\|_{C^k(\mathbf{R}^d)} := \sum_{j=0}^k \sup_{x \in \mathbf{R}^d} |\nabla^j f(x)| = \sum_{j=0}^k \|\nabla^j f\|_{L^\infty(\mathbf{R}^d)},$$

where we view $\nabla^j f$ as a rank j, dimension d tensor with complex coefficients (or equivalently, as a vector of dimension d^j with complex coefficients), thus

$$|\nabla^j f(x)| = (\sum_{i_1,\ldots,i_j=1,\ldots d} |\frac{\partial^j}{\partial_{x_{i_1}} \ldots \partial_{x_{i_j}}} f(x)|^2)^{1/2}.$$

(One does not have to use the ℓ^2 norm here, actually; since all norms on a finite-dimensional space are equivalent, any other means of taking norms here will lead to an equivalent definition of the C^k norm. More generally, all the norms discussed here tend to have several definitions which are equivalent up to constants, and in most cases the exact choice of norm one uses is just a matter of personal taste.)

Remark 1.14.1. In some texts, $C^k(\mathbf{R}^d)$ is used to denote the functions which are k times continuously differentiable, but whose derivatives up to k^{th} order are allowed to be unbounded, so for instance e^x would lie in $C^k(\mathbf{R})$ for every k under this definition. Here, we will refer to such functions (with unbounded derivatives) as lying in $C^k_{\text{loc}}(\mathbf{R}^d)$ (i.e. they are locally in C^k), rather than $C^k(\mathbf{R}^d)$. Similarly, we make a distinction between $C^\infty_{\text{loc}}(\mathbf{R}^d) = \bigcap_{k=1}^\infty C^k_{\text{loc}}(\mathbf{R}^d)$ (smooth functions, with no bounds on derivatives) and $C^\infty(\mathbf{R}^d) = \bigcap_{k=1}^\infty C^k(\mathbf{R}^d)$ (smooth

functions, all of whose derivatives are bounded). Thus, for instance, e^x lies in $C^\infty_{\mathrm{loc}}(\mathbf{R})$ but not $C^\infty(\mathbf{R})$.

Exercise 1.14.1. Show that $C^k(\mathbf{R}^d)$ is a Banach space.

Exercise 1.14.2. Show that for every $d \geq 1$ and $k \geq 0$, the $C^k(\mathbf{R}^d)$ norm is equivalent to the modified norm

$$\|f\|_{\tilde{C}^k(\mathbf{R}^d)} := \|f\|_{L^\infty(\mathbf{R}^d)} + \|\nabla^k f\|_{L^\infty(\mathbf{R}^d)}$$

in the sense that there exists a constant C (depending on k and d) such that

$$C^{-1}\|f\|_{C^k(\mathbf{R}^d)} \leq \|f\|_{\tilde{C}^k(\mathbf{R}^d)} \leq \|f\|_{C^k(\mathbf{R}^d)}$$

for all $f \in C^k(\mathbf{R}^d)$. (*Hint*: use Taylor series with remainder.) Thus when defining the C^k norms, one does not really need to bound all the intermediate derivatives $\nabla^j f$ for $0 < j < k$; the two extreme terms $j = 0, j = k$ suffice. (This is part of a more general interpolation phenomenon; the extreme terms in a sum often already suffice to control the intermediate terms.)

Exercise 1.14.3. Let $\phi \in C^\infty_c(\mathbf{R}^d)$ be a bump function, and $k \geq 0$. Show that if $\xi \in \mathbf{R}^d$ with $|\xi| \geq 1$, $R \geq 1/|\xi|$, and $A > 0$, then the function $A\phi(x/R)\sin(\xi \cdot x)$ has a C^k norm of at most $CA|\xi|^k$, where C is a constant depending only on ϕ, d and k. Thus we see how the C^∞_c norm relates to the height A, width R^d, and frequency scale N of the function, and in particular how the width R is largely irrelevant. What happens when the condition $R \geq 1/|\xi|$ is dropped?

We clearly have the inclusions

$$C^0(\mathbf{R}^d) \supset C^1(\mathbf{R}^d) \supset C^2(\mathbf{R}^d) \supset \ldots$$

and for any constant-coefficient partial differential operator

$$L = \sum_{i_1,\ldots,i_d \geq 0: i_1+\ldots+i_d \leq m} c_{i_1,\ldots,i_d} \frac{\partial^{i_1+\ldots+i_d}}{\partial_{x_1^{i_1}} \ldots \partial_{x_d^{i_d}}}$$

of some order $m \geq 0$, it is easy to see that L is a bounded linear operator from $C^{k+m}(\mathbf{R}^d)$ to $C^k(\mathbf{R}^d)$ for any $k \geq 0$.

The Hölder spaces $C^{k,\alpha}(\mathbf{R}^d)$ are designed to "fill up the gaps" between the discrete spectrum $C^k(\mathbf{R}^d)$ of the continuously differentiable spaces. For $k = 0$ and $0 \leq \alpha \leq 1$, these spaces are defined as

the subspace of functions $f \in C^0(\mathbf{R}^d)$ whose norm

$$\|f\|_{C^{0,\alpha}(\mathbf{R}^d)} := \|f\|_{C^0(\mathbf{R}^d)} + \sup_{x,y \in \mathbf{R}^d : x \neq y} \frac{|f(x) - f(y)|}{|x-y|^\alpha}$$

is finite. To put it another way, $f \in C^{0,\alpha}(\mathbf{R}^d)$ if f is bounded and continuous, and furthermore obeys the *Hölder continuity* bound

$$|f(x) - f(y)| \leq C|x-y|^\alpha$$

for some constant $C > 0$ and all $x, y \in \mathbf{R}^d$.

The space $C^{0,0}(\mathbf{R}^d)$ is easily seen to be just $C^0(\mathbf{R}^d)$ (with an equivalent norm). At the other extreme, $C^{0,1}(\mathbf{R}^d)$ is the class of *Lipschitz* functions, and is also denoted $\mathrm{Lip}(\mathbf{R}^d)$ (and the $C^{0,1}$ norm is also known as the *Lipschitz norm*).

Exercise 1.14.4. Show that $C^{0,\alpha}(\mathbf{R}^d)$ is a Banach space for every $0 \leq \alpha \leq 1$.

Exercise 1.14.5. Show that $C^{0,\alpha}(\mathbf{R}^d) \supset C^{0,\beta}(\mathbf{R}^d)$ for every $0 \leq \alpha \leq \beta \leq 1$, and that the inclusion map is continuous.

Exercise 1.14.6. If $\alpha > 1$, show that the $C^{0,\alpha}(\mathbf{R}^d)$ norm of a function f is finite if and only if f is constant. This explains why we generally restrict the Hölder index α to be less than or equal to 1.

Exercise 1.14.7. Show that $C^1(\mathbf{R}^d)$ is a proper subspace of $C^{0,1}(\mathbf{R}^d)$, and that the restriction of the $C^{0,1}(\mathbf{R}^d)$ norm to $C^1(\mathbf{R}^d)$ is equivalent to the C^1 norm. (The relationship between $C^1(\mathbf{R}^d)$ and $C^{0,1}(\mathbf{R}^d)$ is in fact closely analogous to that between $C^0(\mathbf{R}^d)$ and $L^\infty(\mathbf{R}^d)$, as can be seen from the fundamental theorem of calculus.)

Exercise 1.14.8. Let $f \in (C_c^\infty(\mathbf{R}))^*$ be a distribution. Show that $f \in C^{0,1}(\mathbf{R})$ if and only if $f \in L^\infty(\mathbf{R})$, and the distributional derivative f' of f also lies in $L^\infty(\mathbf{R})$. Furthermore, for $f \in C^{0,1}(\mathbf{R})$, show that $\|f\|_{C^{0,1}(\mathbf{R})}$ is comparable to $\|f\|_{L^\infty(\mathbf{R})} + \|f'\|_{L^\infty(\mathbf{R})}$.

We can then define the $C^{k,\alpha}(\mathbf{R}^d)$ spaces for natural numbers $k \geq 0$ and $0 \leq \alpha \leq 1$ to be the subspace of $C^k(\mathbf{R}^d)$ whose norm

$$\|f\|_{C^{k,\alpha}(\mathbf{R}^d)} := \sum_{j=0}^k \|\nabla^j f\|_{C^{0,\alpha}(\mathbf{R}^d)}$$

is finite. (As before, there are a variety of ways to define the $C^{0,\alpha}$ norm of the tensor-valued quantity $\nabla^j f$, but they are all equivalent to each other.)

Exercise 1.14.9. Show that $C^{k,\alpha}(\mathbf{R}^d)$ is a Banach space which contains $C^{k+1}(\mathbf{R}^d)$, and is contained in turn in $C^k(\mathbf{R}^d)$.

As before, $C^{k,0}(\mathbf{R}^d)$ is equal to $C^k(\mathbf{R}^d)$, and $C^{k,\alpha}(\mathbf{R}^d)$ is contained in $C^{k,\beta}(\mathbf{R}^d)$. The space $C^{k,1}(\mathbf{R}^d)$ is slightly larger than C^{k+1}, but is fairly close to it, thus providing a near-continuum of spaces between the sequence of spaces $C^k(\mathbf{R}^d)$. The following examples illustrates this:

Exercise 1.14.10. Let $\phi \in C_c^\infty(\mathbf{R})$ be a test function, let $k \geq 0$ be a natural number, and let $0 \leq \alpha \leq 1$.

- Show that the function $|x|^s \phi(x)$ lies in $C^{k,\alpha}(\mathbf{R})$ whenever $s \geq k + \alpha$.
- Conversely, if s is not an integer, $\phi(0) \neq 0$, and $s < k + \alpha$, show that $|x|^s \phi(x)$ does *not* lie in $C^{k,\alpha}(\mathbf{R})$.
- Show that $|x|^{k+1}\phi(x)1_{x>0}$ lies in $C^{k,1}(\mathbf{R})$, but not in $C^{k+1}(\mathbf{R})$.

This example illustrates that the quantity $k+\alpha$ can be viewed as measuring the total amount of regularity held by functions in $C^{k,\alpha}(\mathbf{R})$: k full derivatives, plus an additional α amount of Hölder continuity.

Exercise 1.14.11. Let $\phi \in C_c^\infty(\mathbf{R}^d)$ be a test function, let $k \geq 0$ be a natural number, and let $0 \leq \alpha \leq 1$. Show that for $\xi \in \mathbf{R}^d$ with $|\xi| \geq 1$, the function $\phi(x)\sin(\xi \cdot x)$ has a $C^{k,\alpha}(\mathbf{R})$ norm of at most $C|\xi|^{k+\alpha}$, for some C depending on ϕ, d, k, α.

By construction, it is clear that continuously differential operators L of order m will map $C^{k+m,\alpha}(\mathbf{R}^d)$ continuously to $C^{k,\alpha}(\mathbf{R}^d)$.

Now we consider what happens with products.

Exercise 1.14.12. Let $k, l \geq 0$ be natural numbers, and $0 \leq \alpha, \beta \leq 1$.

- If $f \in C^k(\mathbf{R}^d)$ and $g \in C^l(\mathbf{R}^d)$, show that $fg \in C^{\min(k,l)}(\mathbf{R}^d)$, and that the multiplication map is continuous from $C^k(\mathbf{R}^d) \times C^l(\mathbf{R}^d)$ to $C^{\min(k,l)}(\mathbf{R}^d)$. (*Hint*: reduce to the case $k = l$ and use induction.)

- If $f \in C^{k,\alpha}(\mathbf{R}^d)$ and $g \in C^{l,\beta}(\mathbf{R}^d)$, and $k + \alpha \leq l + \beta$, show that $fg \in C^{k,\alpha}(\mathbf{R}^d)$, and that the multiplication map is continuous from $C^{k,\alpha}(\mathbf{R}^d) \times C^{l,\beta}(\mathbf{R}^d)$ to $C^{k,\alpha}(\mathbf{R}^d)$.

It is easy to see that the regularity in these results cannot be improved (just take $g = 1$). This illustrates a general principle, namely that a pointwise product fg tends to acquire the *lower* of the regularities of the two factors f, g.

As one consequence of this exercise, we see that any *variable-coefficient* differential operator L of order m with $C^\infty(\mathbf{R})$ coefficients will map $C^{m+k,\alpha}(\mathbf{R}^d)$ to $C^{k,\alpha}(\mathbf{R}^d)$ for any $k \geq 0$ and $0 \leq \alpha \leq 1$.

We now briefly remark on Hölder spaces on open domains Ω in Euclidean space $\mathbf{R}^d$. Here, a new subtlety emerges; instead of having just one space $C^{k,\alpha}$ for each choice of exponents k, α, one actually has a range of spaces to choose from, depending on what kind of behaviour one wants to impose at the boundary of the domain. At one extreme, one has the space $C^{k,\alpha}(\Omega)$, defined as the space of k times continuously differentiable functions $f : \Omega \to \mathbf{C}$ whose Hölder norm

$$\|f\|_{C^{k,\alpha}(\Omega)} := \sum_{j=0}^{k} \sup_{x \in \Omega} |\nabla^j f(x)| + \sup_{x,y \in \Omega: x \neq y} \frac{|\nabla^j f(x) - \nabla^j f(y)|}{|x-y|^\alpha}$$

is finite; this is the "maximal" choice for the $C^{k,\alpha}(\Omega)$. At the other extreme, one has the space $C_0^{k,\alpha}(\Omega)$, defined as the closure of the compactly supported functions in $C^{k,\alpha}(\Omega)$. This space is smaller than $C^{k,\alpha}(\Omega)$; for instance, functions in $C_0^{0,\alpha}((0,1))$ must converge to zero at the endpoints $0, 1$, while functions in $C^{k,\alpha}((0,1))$ do not need to do so. An intermediate space is $C^{k,\alpha}(\mathbf{R}^d)\restriction_\Omega$, defined as the space of restrictions of functions in $C^{k,\alpha}(\mathbf{R}^d)$ to Ω. For instance, the restriction of $|x|\psi(x)$ to $\mathbf{R}\backslash\{0\}$, where ψ is a cutoff function non-vanishing at the origin, lies in $C^{1,0}(\mathbf{R}\backslash\{0\})$, but is not in $C^{1,0}(\mathbf{R})\restriction_{\mathbf{R}\backslash\{0\}}$ or $C_0^{1,0}(\mathbf{R}\backslash\{0\})$ (note that $|x|\psi(x)$ itself is not in $C^{1,0}(\mathbf{R})$, as it is not continuously differentiable at the origin). It is possible to clarify the exact relationships between the various flavours of Hölder spaces on domains (and similarly for the Sobolev spaces discussed below), but we will not discuss these topics here.

Exercise 1.14.13. Show that $C_c^\infty(\mathbf{R}^d)$ is a dense subset of $C^{k,\alpha}(\mathbf{R}^d)$ for any $k \geq 0$ and $0 \leq \alpha \leq 1$. (*Hint*: To approximate a $C^{k,\alpha}$ function by a C_c^∞ one, first smoothly truncate the function at a large spatial scale to be compactly supported, then convolve with a smooth, compactly supported approximation to the identity.)

Hölder spaces are particularly useful in elliptic PDE, because tools such as the maximum principle lend themselves well to the suprema that appear inside the definition of the $C^{k,\alpha}$ norms; see [**GiTr1998**] for a thorough treatment. For simple examples of elliptic PDE, such as the Poisson equation $\Delta u = f$, one can also use the explicit fundamental solution, through lengthy but straightforward computations. We give a typical example here:

Exercise 1.14.14 (Schauder estimate)**.** Let $0 < \alpha < 1$, and let $f \in C^{0,\alpha}(\mathbf{R}^3)$ be a function supported on the unit ball $B(0,1)$. Let u be the unique bounded solution to the Poisson equation $\Delta u = f$ (where $\Delta = \sum_{j=1}^3 \frac{\partial^2}{\partial x_j^2}$ is the Laplacian), given by convolution with the Newton kernel:

$$u(x) := \frac{1}{4\pi} \int_{\mathbf{R}^3} \frac{f(y)}{|x-y|}\, dy.$$

(i) Show that $u \in C^0(\mathbf{R}^3)$.

(ii) Show that $u \in C^1(\mathbf{R}^3)$, and rigorously establish the formula

$$\frac{\partial u}{\partial x_j}(x) = -\frac{1}{4\pi} \int_{\mathbf{R}^3} (x_j - y_j) \frac{f(y)}{|x-y|^3}\, dy$$

for $j = 1, 2, 3$.

(iii) Show that $u \in C^2(\mathbf{R}^3)$, and rigorously establish the formula

$$\frac{\partial^2 u}{\partial x_i \partial x_j}(x) = \frac{1}{4\pi} \lim_{\varepsilon \to 0} \int_{|x-y| \geq \varepsilon} [\frac{3(x_i - y_i)(x_j - y_j)}{|x-y|^5} - \frac{\delta_{ij}}{|x-y|^3}] f(y)\, dy$$

for $i, j = 1, 2, 3$, where δ_{ij} is the *Kronecker delta*. (*Hint*: first establish this in the two model cases when $f(x) = 0$, and when f is constant near x.)

(iv) Show that $u \in C^{2,\alpha}(\mathbf{R}^3)$, and establish the *Schauder estimate*

$$\|u\|_{C^{2,\alpha}(\mathbf{R}^3)} \leq C_\alpha \|f\|_{C^{0,\alpha}(\mathbf{R}^3)}$$

where C_α depends only on α.

(v) Show that the Schauder estimate fails when $\alpha = 0$. Using this, conclude that there eixsts $f \in C^0(\mathbf{R}^3)$ supported in the unit ball such that the function u defined above fails to be in $C^2(\mathbf{R}^3)$. (*Hint*: use the closed graph theorem, Theorem 1.7.19.) This failure helps explain why it is necessary to introduce Hölder spaces into elliptic theory in the first place (as opposed to the more intuitive C^k spaces).

Remark 1.14.2. Roughly speaking, the Schauder estimate asserts that if Δu has $C^{0,\alpha}$ regularity, then all other second derivatives of u have $C^{0,\alpha}$ regularity as well. This phenomenon - that control of a special derivative of u at some order implies control of all other derivatives of u at that order - is known as *elliptic regularity*, and relies crucially on Δ being an *elliptic* differential operator. We will discus ellipticity a little bit more later in Exercise 1.14.36. The theory of Schauder estimates is by now extremely well developed, and applies to large classes of elliptic operators on quite general domains, but we will not discuss these estimates and their applications to various linear and nonlinear elliptic PDE here.

Exercise 1.14.15 (Rellich-Kondrakov type embedding theorem for Hölder spaces)**.** Let $0 \leq \alpha < \beta \leq 1$. Show that any bounded sequence of functions $f_n \in C^{0,\beta}(\mathbf{R}^d)$ that are all supported in the same compact subset of $\mathbf{R}^n$ will have a subsequence that converges in $C^{0,\alpha}(\mathbf{R}^d)$. (*Hint*: use the *Arzelá-Ascoli theorem* (Theorem 1.8.23) to first obtain uniform convergence, then upgrade this convergence.) This is part of a more general phenomenon: sequences bounded in a high regularity space, and constrained to lie in a compact domain, will tend to have convergent subsequences in low regularity spaces.

1.14.2. Classical Sobolev spaces. We now turn to the "classical" Sobolev spaces $W^{k,p}(\mathbf{R}^d)$, which involve only an integral amount k of regularity.

Definition 1.14.3. Let $1 \leq p \leq \infty$, and let $k \geq 0$ be a natural number. A function f is said to lie in $W^{k,p}(\mathbf{R}^d)$ if its *weak derivatives* $\nabla^j f$ exist and lie in $L^p(\mathbf{R}^d)$ for all $j = 0, \ldots, k$. If f lies in $W^{k,p}(\mathbf{R}^d)$,

we define the $W^{k,p}$ norm of f by the formula

$$\|f\|_{W^{k,p}(\mathbf{R}^d)} := \sum_{j=0}^{k} \|\nabla^j f\|_{L^p(\mathbf{R}^d)}.$$

(As before, the exact choice of convention in which one measures the L^p norm of ∇^j is not particularly relevant for most applications, as all such conventions are equivalent up to multiplicative constants.)

The space $W^{k,p}(\mathbf{R}^d)$ is also denoted $L^p_k(\mathbf{R}^d)$ in some texts.

Example 1.14.4. $W^{0,p}(\mathbf{R}^d)$ is of course the same space as $L^p(\mathbf{R}^d)$, thus the Sobolev spaces generalise the Lebesgue spaces. From Exercise 1.14.8 we see that $W^{1,\infty}(\mathbf{R})$ is the same space as $C^{0,1}(\mathbf{R})$, with an equivalent norm. More generally, one can see from induction that $W^{k+1,\infty}(\mathbf{R})$ is the same space as $C^{k,1}(\mathbf{R})$ for $k \geq 0$, with an equivalent norm. It is also clear that $W^{k,p}(\mathbf{R}^d)$ contains $W^{k+1,p}(\mathbf{R}^d)$ for any k, p.

Example 1.14.5. The function $|\sin x|$ lies in $W^{1,\infty}(\mathbf{R})$, but is not everywhere differentiable in the classical sense; nevertheless, it has a bounded weak derivative of $\cos x \operatorname{sgn}(\sin(x))$. On the other hand, the *Cantor function* (aka the "Devil's staircase") is not in $W^{1,\infty}(\mathbf{R})$, despite having a classical derivative of zero at almost every point; the weak derivative is a Cantor measure, which does not lie in any L^p space. Thus one really does need to work with weak derivatives rather than classical derivatives to define Sobolev spaces properly (in contrast to the $C^{k,\alpha}$ spaces).

Exercise 1.14.16. Let $\phi \in C_c^\infty(\mathbf{R}^d)$ be a bump function, $k \geq 0$, and $1 \leq p \leq \infty$. Show that if $\xi \in \mathbf{R}^d$ with $|\xi| \geq 1$, $R \geq 1/|\xi|$, and $A > 0$, then the function $\phi(x/R)\sin(\xi x)$ has a $W^{k,p}(\mathbf{R})$ norm of at most $CA|\xi|^k R^{d/p}$, where C is a constant depending only on ϕ, p and k. (Compare this with Exercise 1.14.3 and Exercise 1.14.11.) What happens when the condition $R \geq 1/|\xi|$ is dropped?

Exercise 1.14.17. Show that $W^{k,p}(\mathbf{R}^d)$ is a Banach space for any $1 \leq p \leq \infty$ and $k \geq 0$.

The fact that Sobolev spaces are defined using weak derivatives is a technical nuisance, but in practice one can often end up working with classical derivatives anyway by means of the following lemma:

Lemma 1.14.6. *Let $1 \leq p < \infty$ and $k \geq 0$. Then the space $C_c^\infty(\mathbf{R}^d)$ of test functions is a dense subspace of $W^{k,p}(\mathbf{R}^d)$.*

Proof. It is clear that $C_c^\infty(\mathbf{R}^d)$ is a subspace of $W^{k,p}(\mathbf{R}^d)$. We first show that the smooth functions $C^\infty_{\mathrm{loc}}(\mathbf{R}^d) \cap W^{k,p}(\mathbf{R}^d)$ is a dense subspace of $W^{k,p}(\mathbf{R}^d)$, and then show that $C_c^\infty(\mathbf{R}^d)$ is dense in $C^\infty_{\mathrm{loc}}(\mathbf{R}^d) \cap W^{k,p}(\mathbf{R}^d)$.

We begin with the former claim. Let $f \in W^{k,p}(\mathbf{R}^d)$, and let ϕ_n be a sequence of smooth, compactly supported approximations to the identity. Since $f \in L^p(\mathbf{R}^d)$, we see that $f * \phi_n$ converges to f in $L^p(\mathbf{R}^d)$. More generally, since $\nabla^j f$ is in $L^p(\mathbf{R}^d)$ for $0 \leq j \leq k$, we see that $(\nabla^j f) * \phi_n = \nabla^j (f * \phi_n)$ converges to $\nabla^j f$ in $L^p(\mathbf{R}^d)$. Thus we see that $f * \phi_n$ converges to f in $W^{k,p}(\mathbf{R}^d)$. On the other hand, as ϕ_n is smooth, $f * \phi_n$ is smooth; and the claim follows.

Now we prove the latter claim. Let f be a smooth function in $W^{k,p}(\mathbf{R}^d)$, thus $\nabla^j f \in L^p(\mathbf{R}^d)$ for all $0 \leq j \leq k$. We let $\eta \in C_c^\infty(\mathbf{R}^d)$ be a compactly supported function which equals 1 near the origin, and consider the functions $f_R(x) := f(x)\eta(x/R)$ for $R > 0$. Clearly, each f_R lies in $C_c^\infty(\mathbf{R}^d)$. As $R \to \infty$, dominated convergence shows that f_R converges to f in $L^p(\mathbf{R}^d)$. An application of the product rule then lets us write $\nabla f_R(x) = (\nabla f)(x)\eta(x/R) + \frac{1}{R} f(x)(\nabla \eta)(x/R)$. The first term converges to ∇f in $L^p(\mathbf{R}^d)$ by dominated convergence, while the second term goes to zero in the same topology; thus ∇f_R converges to ∇f in $L^p(\mathbf{R}^d)$. A similar argument shows that $\nabla^j f_R$ converges to $\nabla^j f$ in $L^p(\mathbf{R}^d)$ for all $0 \leq j \leq k$, and so f_R converges to f in $W^{k,p}(\mathbf{R}^d)$, and the claim follows. □

As a corollary of this lemma we also see that the space $\mathcal{S}(\mathbf{R}^d)$ of Schwartz functions is dense in $W^{k,p}(\mathbf{R}^d)$.

Exercise 1.14.18. Let $k \geq 0$. Show that the closure of $C_c^\infty(\mathbf{R}^d)$ in $W^{k,\infty}(\mathbf{R}^d)$ is $C^{k+1}(\mathbf{R}^d)$, thus Lemma 1.14.6 fails at the endpoint $p = \infty$.

Now we come to the important *Sobolev embedding theorem*, which allows one to trade regularity for integrability. We illustrate this phenomenon first with some very simple cases. First, we claim that the space $W^{1,1}(\mathbf{R})$ embeds continuously into $W^{0,\infty}(\mathbf{R}) = L^\infty(\mathbf{R})$,

thus trading in one degree of regularity to upgrade L^1 integrability to L^∞ integrability. To prove this claim, it suffices to establish the bound

$$\|f\|_{L^\infty(\mathbf{R})} \leq C\|f\|_{W^{1,1}(\mathbf{R})} \tag{1.123}$$

for all test functions $f \in C_c^\infty(\mathbf{R})$ and some constant C, as the claim then follows by taking limits using Lemma 1.14.6. (Note that any limit in either the L^∞ or $W^{1,1}$ topologies, is also a limit in the sense of distributions, and such limits are necessarily unique. Also, since $L^\infty(\mathbf{R})$ is the dual space of $L^1(\mathbf{R})$, the distributional limit of any sequence bounded in $L^\infty(\mathbf{R})$ remains in $L^\infty(\mathbf{R})$, by Exercise 1.13.28.) To prove (1.123), observe from the fundamental theorem of calculus that

$$|f(x) - f(0)| = |\int_0^x f'(t)\ dt| \leq \|f'\|_{L^1(\mathbf{R})} \leq \|f\|_{W^{1,1}(\mathbf{R})}$$

for all x; in particular, from the triangle inequality

$$\|f\|_{L^\infty(\mathbf{R})} \leq |f(0)| + \|f\|_{W^{1,1}(\mathbf{R})}.$$

Also, taking x to be sufficiently large, we see (from the compact support of f) that

$$|f(0)| \leq \|f\|_{W^{1,1}(\mathbf{R})}$$

and (1.123) follows.

Since the closure of $C_c^\infty(\mathbf{R})$ in $L^\infty(\mathbf{R})$ is $C_0(\mathbf{R})$, we actually obtain the stronger embedding, that $W^{1,1}(\mathbf{R})$ embeds continuously into $C_0(\mathbf{R})$.

Exercise 1.14.19. Show that $W^{d,1}(\mathbf{R}^d)$ embeds continuously into $C_0(\mathbf{R}^d)$, thus there exists a constant C (depending only on d) such that

$$\|f\|_{C_0(\mathbf{R}^d)} \leq C\|f\|_{W^{d,1}(\mathbf{R}^d)}$$

for all $f \in W^{d,1}(\mathbf{R}^d)$.

Now we turn to Sobolev embedding for exponents other than $p = 1$ and $p = \infty$.

Theorem 1.14.7 (Sobolev embedding theorem for one derivative)**.** *Let $1 \leq p \leq q \leq \infty$ be such that $\frac{d}{p} - 1 \leq \frac{d}{q} \leq \frac{d}{p}$, but that one is not in*

the endpoint cases $(p,q) = (d,\infty), (1, \frac{d}{d-1})$. *Then* $W^{1,p}(\mathbf{R}^d)$ *embeds continuously into* $L^q(\mathbf{R}^d)$.

Proof. By Lemma 1.14.6 and the same limiting argument as before, it suffices to establish the *Sobolev embedding inequality*

$$\|f\|_{L^q(\mathbf{R}^d)} \leq C_{p,q,d}\|f\|_{W^{1,p}(\mathbf{R}^d)}$$

for all test functions $f \in C_c^\infty(\mathbf{R}^d)$, and some constant $C_{p,q,d}$ depending only on p,q,d, as the inequality will then extend to all $f \in W^{1,p}(\mathbf{R}^d)$. To simplify the notation we shall use $X \lesssim Y$ to denote an estimate of the form $X \leq C_{p,q,d}Y$, where $C_{p,q,d}$ is a constant depending on p,q,d (the exact value of this constant may vary from instance to instance).

The case $p = q$ is trivial. Now let us look at another extreme case, namely when $\frac{d}{p} - 1 = \frac{d}{q}$; by our hypotheses, this forces $1 < p < d$. Here, we use the fundamental theorem of calculus (and the compact support of f) to write

$$f(x) = -\int_0^\infty \omega \cdot \nabla f(x + r\omega)\ dr$$

for any $x \in \mathbf{R}^d$ and any direction $\omega \in S^{d-1}$. Taking absolute values, we conclude in particular that

$$|f(x)| \lesssim \int_0^\infty |\nabla f(x + r\omega)|\ dr.$$

We can average this over all directions ω:

$$|f(x)| \lesssim \int_{S^{d-1}} \int_0^\infty |\nabla f(x + r\omega)|\ dr d\omega.$$

Switching from polar coordinates back to Cartesian (multiplying and dividing by r^{d-1}) we conclude that

$$|f(x)| \lesssim \int_{\mathbf{R}^d} \frac{1}{|y|^{d-1}} |\nabla f(x-y)|\ dy,$$

thus f is pointwise controlled by the convolution of $|\nabla f|$ with the fractional integration $\frac{1}{|x|^{d-1}}$. By the Hardy-Littlewood-Sobolev theorem on fractional integration (Corollary 1.11.18) we conclude that

$$\|f\|_{L^q(\mathbf{R}^d)} \lesssim \|\nabla f\|_{L^p(\mathbf{R}^d)}$$

and the claim follows. (Note that the hypotheses $1 < p < d$ are needed here in order to be able to invoke this theorem.)

Now we handle intermediate cases, when $\frac{d}{p} - 1 < \frac{d}{q} < \frac{d}{p}$. (Many of these cases can be obtained from the endpoints already established by interpolation, but unfortunately not all such cases can be, so we will treat this case separately.) Here, the trick is not to integrate out to infinity, but instead to integrate out to a bounded distance. For instance, the fundamental theorem of calculus gives

$$f(x) = f(x + R\omega) - \int_0^R \omega \cdot \nabla f(x + r\omega)\ dr$$

for any $R > 0$, hence

$$|f(x)| \lesssim |f(x + R\omega)| + \int_0^R |\nabla f(x + r\omega)|\ dr$$

What value of R should one pick? If one picks any specific value of R, one would end up with an average of f over spheres, which looks somewhat unpleasant. But what one can do here is average over a *range* of R's, for instance between 1 and 2. This leads to

$$|f(x)| \lesssim \int_1^2 |f(x + R\omega)|\ dR + \int_0^2 |\nabla f(x + r\omega)|\ dr;$$

averaging over all directions ω and converting back to Cartesian coordinates, we see that

$$|f(x)| \lesssim \int_{1 \leq |y| \leq 2} |f(x - y)|\ dy + \int_{|y| \leq 2} \frac{1}{|y|^{d-1}} |\nabla f(x - y)|\ dy.$$

Thus one is bounding $|f|$ pointwise (up to constants) by the convolution of $|f|$ with the kernel $K_1(y) := 1_{1 \leq |y| \leq 2}$, plus the convolution of $|\nabla f|$ with the kernel $K_2(y) := 1_{|y| \leq 2} \frac{1}{|y|^{d-1}}$. A short computation shows that both kernels lie in $L^r(\mathbf{R}^d)$, where r is the exponent in Young's inequality, and more specifically that $\frac{1}{q} + 1 = \frac{1}{p} + \frac{1}{r}$ (and in particular $1 < r < \frac{d}{d-1}$). Applying Young's inequality (Exercise 1.11.25), we conclude that

$$\|f\|_{L^q(\mathbf{R}^d)} \lesssim \|f\|_{L^p(\mathbf{R}^d)} + \|\nabla f\|_{L^p(\mathbf{R}^d)}$$

and the claim follows. □

Remark 1.14.8. It is instructive to insert the example in Exercise 1.14.16 into the Sobolev embedding theorem. By replacing the $W^{1,p}(\mathbf{R}^d)$ norm with the $L^q(\mathbf{R}^d)$ norm, one trades one factor of the frequency scale $|\xi|$ for $\frac{1}{q}-\frac{1}{p}$ powers of the width R^d. This is consistent with the Sobolev embedding theorem so long as $R^d \gtrsim 1/|\xi|^d$, which is essentially one of the hypotheses in that exercise. Thus, one can view Sobolev embedding as an assertion that the width of a function must always be greater than or comparable to the wavelength scale (the reciprocal of the frequency scale), raised to the power of the dimension; this is a manifestation of the *uncertainty principle* (see Section 2.6 for further discussion).

Exercise 1.14.20. Let $d \geq 2$. Show that the Sobolev endpoint estimate fails in the case $(p,q) = (d,\infty)$. (*Hint*: experiment with functions f of the form $f(x) := \sum_{n=1}^N \phi(2^n x)$, where ϕ is a test function supported on the annulus $\{1 \leq |x| \leq 2\}$.) Conclude in particular that $W^{1,d}(\mathbf{R}^d)$ is not a subset of $L^\infty(\mathbf{R}^d)$. (*Hint*: Either use the closed graph theorem, or use some variant of the function f used in the first part of this exercise.) Note that when $d = 1$, the Sobolev endpoint theorem for $(p,q) = (1,\infty)$ follows from the fundamental theorem of calculus, as mentioned earlier. There are substitutes known for the endpoint Sobolev embedding theorem, but they involve more sophisticated function spaces, such as the space BMO of spaces of *bounded mean oscillation*, which we will not discuss here.

The $p = 1$ case of the Sobolev inequality cannot be proven via the Hardy-Littlewood-Sobolev inequality; however, there are other proofs available. One of these (due to Gagliardo and Nirenberg) is based on

Exercise 1.14.21 (Loomis-Whitney inequality)**.** Let $d \geq 1$, let $f_1,\ldots,f_d \in L^p(\mathbf{R}^{d-1})$ for some $0 < p \leq \infty$, and let $F : \mathbf{R}^d \to \mathbf{C}$ be the function

$$F(x_1,\ldots,x_d) := \prod_{i=1}^d f_i(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_d).$$

Show that

$$\|F\|_{L^{p/(d-1)}(\mathbf{R}^d)} \leq \prod_{i=1}^d \|f_i\|_{L^p(\mathbf{R}^d)}.$$

(*Hint*: induct on d, using Hölder's inequality and Fubini's theorem.)

Lemma 1.14.9 (Endpoint Sobolev inequality)**.** $W^{1,1}(\mathbf{R}^d)$ *embeds continuously into* $L^{d/(d-1)}(\mathbf{R}^d)$.

Proof. It will suffice to show that

$$\|f\|_{L^{d/(d-1)}(\mathbf{R}^d)} \leq \|\nabla f\|_{L^1(\mathbf{R}^d)}$$

for all test functions $f \in C_c^\infty(\mathbf{R}^d)$. From the fundamental theorem of calculus we see that

$$|f(x_1,\ldots,x_d)| \leq \int_{\mathbf{R}} |\frac{\partial f}{\partial x_i}(x_1,\ldots,x_{i-1},t,x_{i+1},\ldots,x_d)|\ dt$$

and thus

$$|f(x_1,\ldots,x_d)| \leq f_i(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_d)$$

where

$$f_i(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_d) := \int_{\mathbf{R}} |\nabla f(x_1,\ldots,x_{i-1},t,x_{i+1},\ldots,x_d)|\,dt.$$

From Fubini's theorem we have

$$\|f_i\|_{L^1(\mathbf{R}^d)} = \|\nabla f\|_{L^1(\mathbf{R}^d)}$$

and hence by the Loomis-Whitney inequality

$$\|f_1 \ldots f_d\|_{L^{1/(d-1)}(\mathbf{R}^d)} \leq \|\nabla f\|^d_{L^1(\mathbf{R}^d)},$$

and the claim follows. □

Exercise 1.14.22 (Connection between Sobolev embedding and isoperimetric inequality)**.** Let $d \geq 2$, and let Ω be an open subset of $\mathbf{R}^d$ whose boundary $\partial\Omega$ is a smooth $d-1$-dimensional manifold. Show that the surface area $|\partial\Omega|$ of Ω is related to the volume $|\Omega|$ of Ω by the *isoperimetric inequality*

$$|\Omega| \leq C_d |\partial\Omega|^{d/(d-1)}$$

for some constant C_d depending only on d. (*Hint*: Apply the endpoint Sobolev theorem to a suitably smoothed out version of 1_Ω.) It is also possible to reverse this implication and deduce the endpoint Sobolev embedding theorem from the isoperimetric inequality and the coarea formula, which we will do in later notes.

Exercise 1.14.23. Use dimensional analysis to argue why the Sobolev embedding theorem should fail when $\frac{d}{q} < \frac{d}{p} - 1$. Then create a rigorous counterexample to that theorem in this case.

Exercise 1.14.24. Show that $W^{k,p}(\mathbf{R}^d)$ embeds into $W^{l,q}(\mathbf{R}^d)$ whenever $k \geq l \geq 0$ and $1 < p < q \leq \infty$ are such that $\frac{d}{p} - k \leq \frac{d}{q} - l$, and such that at least one of the two inequalities $q \leq \infty$, $\frac{d}{p} - k \leq \frac{d}{q} - l$ is strict.

Exercise 1.14.25. Show that the Sobolev embedding theorem fails whenever $q < p$. (*Hint*: experiment with functions of the form $f(x) = \sum_{j=1}^n \phi(x - x_j)$, where ϕ is a test function and the x_j are widely separated points in space.)

Exercise 1.14.26 (Hölder-Sobolev embedding)**.** Let $d < p < \infty$. Show that $W^{1,p}(\mathbf{R}^d)$ embeds continuously into $C^{0,\alpha}(\mathbf{R}^d)$, where $0 < \alpha < 1$ is defined by the scaling relationship $\frac{d}{p} - 1 = -\alpha$. Use dimensional analysis to justify why one would expect this scaling relationship to arise naturally, and give an example to show that α cannot be improved to any higher exponent.

More generally, with the same assumptions on p, α, show that $W^{k+1,p}(\mathbf{R}^d)$ embeds continuously into $C^{k,\alpha}(\mathbf{R}^d)$ for all natural numbers $k \geq 0$.

Exercise 1.14.27 (Sobolev product theorem, special case)**.** Let $k \geq 1$, $1 < p, q < d/k$, and $1 < r < \infty$ be such that $\frac{1}{p} + \frac{1}{q} - \frac{k}{d} = \frac{1}{r}$. Show that whenever $f \in W^{k,p}(\mathbf{R}^d)$ and $g \in W^{k,q}(\mathbf{R}^d)$, then $fg \in W^{k,r}(\mathbf{R}^d)$, and that

$$\|fg\|_{W^{k,r}(\mathbf{R}^d)} \leq C_{p,q,k,d,r} \|f\|_{W^{k,p}(\mathbf{R}^d)} \|g\|_{W^{k,q}(\mathbf{R}^d)}$$

for some constant $C_{p,q,k,d,r}$ depending only on the subscripted parameters. (This is not the most general range of parameters for which this sort of product theorem holds, but it is an instructive special case.)

Exercise 1.14.28. Let L be a differential operator of order m whose coefficients lie in $C^\infty(\mathbf{R}^d)$. Show that L maps $W^{k+m,p}(\mathbf{R}^d)$ continuously to $W^{k,p}(\mathbf{R}^d)$ for all $1 \leq p \leq \infty$ and all integers $k \geq 0$.

1.14.3. L^2-based Sobolev spaces. It is possible to develop more general Sobolev spaces $W^{s,p}(\mathbf{R}^d)$ than the integer-regularity spaces $W^{k,p}(\mathbf{R}^d)$ defined above, in which s is allowed to take any real number (including negative numbers) as a value, although the theory becomes

somewhat pathological unless one restricts attention to the range $1 < p < \infty$, for reasons having to do with the theory of *singular integrals*.

As the theory of singular integrals is beyond the scope of this course, we will illustrate this theory only in the model case $p = 2$, in which Plancherel's theorem is available, which allows one to avoid dealing with singular integrals by working purely on the frequency space side.

To explain this, we begin with the Plancherel identity

$$\int_{\mathbf{R}^d} |f(x)|^2 \, dx = \int_{\mathbf{R}^d} |\hat{f}(\xi)|^2 \, d\xi,$$

which is valid for all $L^2(\mathbf{R}^d)$ functions and in particular for Schwartz functions $f \in \mathcal{S}(\mathbf{R}^d)$. Also, we know that the Fourier transform of any derivative $\frac{\partial f}{\partial x_j} f$ of f is $-2\pi i \xi_j \hat{f}(\xi)$. From this we see that

$$\int_{\mathbf{R}^d} |\frac{\partial f}{\partial x_j}(x)|^2 \, dx = \int_{\mathbf{R}^d} (2\pi|\xi_j|)^2 |\hat{f}(\xi)|^2 \, d\xi,$$

for all $f \in \mathcal{S}(\mathbf{R}^d)$ and so on summing in j we have

$$\int_{\mathbf{R}^d} |\nabla f(x)|^2 \, dx = \int_{\mathbf{R}^d} (2\pi|\xi|)^2 |\hat{f}(\xi)|^2 \, d\xi.$$

A similar argument then gives

$$\int_{\mathbf{R}^d} |\nabla^j f(x)|^2 \, dx = \int_{\mathbf{R}^d} (2\pi|\xi|)^{2j} |\hat{f}(\xi)|^2 \, d\xi$$

and so on summing in j we have

$$\|f\|^2_{W^{k,2}(\mathbf{R}^d)} = \int_{\mathbf{R}^d} \sum_{j=0}^{k} (2\pi|\xi|)^{2j} |\hat{f}(\xi)|^2 \, d\xi$$

for all $k \geq 0$ and all Schwartz functions $f \in \mathcal{S}(\mathbf{R}^d)$. Since the Schwartz functions are dense in $W^{k,2}(\mathbf{R}^d)$, a limiting argument (using the fact that L^2 is complete) then shows that the above formula also holds for all $f \in W^{k,2}(\mathbf{R}^d)$.

Now observe that the quantity $\sum_{j=0}^{k} (2\pi|\xi|)^{2j}$ is comparable (up to constants depending on k, d) to the expression $\langle \xi \rangle^{2k}$, where $\langle x \rangle := (1 + |x|^2)^{1/2}$ (this quantity is sometimes known as the "Japanese bracket" of x). We thus conclude that

$$\|f\|_{W^{k,2}(\mathbf{R}^d)} \sim \|\langle \xi \rangle^k \hat{f}(\xi)\|_{L^2(\mathbf{R}^d)},$$

where we use $x \sim y$ here to denote the fact that x and y are comparable up to constants depending on d, k, and ξ denotes the variable of independent variable on the right-hand side. If we then define, for any real number s, the space $H^s(\mathbf{R}^d)$ to be the space of all tempered distributions f such that the distribution $\langle\xi\rangle^s \hat{f}(\xi)$ lies in L^2, and give this space the norm

$$\|f\|_{H^s(\mathbf{R}^d)} := \|\langle\xi\rangle^s \hat{f}(\xi)\|_{L^2(\mathbf{R}^d)},$$

then we see that $W^{k,2}(\mathbf{R}^d)$ embeds into $H^k(\mathbf{R}^d)$, and that the norms are equivalent.

Actually, the two spaces are equal:

Exercise 1.14.29. For any $s \in \mathbf{R}$, show that $\mathcal{S}(\mathbf{R}^d)$ is a dense subspace of $H^s(\mathbf{R}^d)$. Use this to conclude that $W^{k,2}(\mathbf{R}^d) = H^k(\mathbf{R}^d)$ for all non-negative integers k.

It is clear that $H^0(\mathbf{R}^d) \equiv L^2(\mathbf{R}^d)$, and that $H^s(\mathbf{R}^d) \subset H^{s'}(\mathbf{R}^d)$ whenever $s > s'$. The spaces $H^s(\mathbf{R}^d)$ are also (complex) *Hilbert spaces*, with the Hilbert space inner product

$$\langle f, g\rangle_{H^s(\mathbf{R}^d)} := \int_{\mathbf{R}^d} \langle\xi\rangle^{2s} f(\xi)\overline{g(\xi)}\, d\xi.$$

It is not hard to verify that this inner product does indeed give $H^s(\mathbf{R}^d)$ the structure of a Hilbert space (indeed, it is isomorphic under the Fourier transform to the Hilbert space $L^2(\langle\xi\rangle^{2s} d\xi)$ which is isomorphic in turn under the map $F(\xi) \mapsto \langle\xi\rangle^s F(\xi)$ to the standard Hilbert space $L^2(\mathbf{R}^d)$).

Being a Hilbert space, $H^s(\mathbf{R}^d)$ is isomorphic to its dual $H^s(\mathbf{R}^d)^*$ (or more precisely, to the complex conjugate of this dual). There is another duality relationship which is also useful:

Exercise 1.14.30 (Duality between H^s and H^{-s})**.** Let $s \in \mathbf{R}$, and $f \in H^s(\mathbf{R}^d)$. Show also for any continuous linear functional $\lambda : H^s(\mathbf{R}^d) \to \mathbf{C}$ there exists a unique $g \in H^{-s}(\mathbf{R}^d)$ such that

$$\lambda(f) = \langle f, g\rangle_{L^2(\mathbf{R}^d)}$$

for all $f \in H^s(\mathbf{R}^d)$, where the inner product $\langle f, g\rangle_{L^2(\mathbf{R}^d)}$ is defined via the Fourier transform as

$$\langle f, g\rangle_{L^2(\mathbf{R}^d)} := \int_{\mathbf{R}^d} \hat{f}(\xi)\overline{\hat{g}(\xi)}\, d\xi.$$

Also show that

$$\|f\|_{H^s(\mathbf{R}^d)} := \sup\{|\langle f, g\rangle_{L^2(\mathbf{R}^d)} : g \in \mathcal{S}(\mathbf{R}^d); \|g\|_{H^{-s}(\mathbf{R}^d)} \leq 1\}$$

for all $f \in H^s(\mathbf{R}^d)$.

The H^s Sobolev spaces also enjoy the same type of embedding estimates as their classical counterparts:

Exercise 1.14.31 (Sobolev embedding for H^s, I)**.** If $s > d/2$, show that $H^s(\mathbf{R}^d)$ embeds continuously into $C^{0,\alpha}(\mathbf{R}^d)$ whenever $0 < \alpha \leq \min(s - \frac{d}{2}, 1)$. (*Hint*: use the Fourier inversion formula and the Cauchy-Schwarz inequality.)

Exercise 1.14.32 (Sobolev embedding for H^s, II)**.** If $0 < s < d/2$, show that $H^s(\mathbf{R}^d)$ embeds continuously into $L^q(\mathbf{R}^d)$ whenever $\frac{d}{2} - s \leq \frac{d}{q} \leq \frac{d}{2}$. (*Hint*: it suffices to handle the extreme case $\frac{d}{q} = \frac{d}{2} - s$. For this, first reduce to establishing the bound $\|f\|_{L^q(\mathbf{R}^d)} \leq C\|f\|_{H^s(\mathbf{R}^d)}$ to the case when $f \in H^s(\mathbf{R}^d)$ is a Schwartz function whose Fourier transform vanishes near the origin (and C depends on s, d, q), and write $\hat{f}(\xi) = \hat{g}(\xi)/|\xi|^s$ for some g which is bounded in $L^2(\mathbf{R}^d)$. Then use Exercise 1.13.35 and Corollary 1.11.18.

Exercise 1.14.33. In this exercise we develop a more elementary variant of Sobolev spaces, the *L^p Hölder spaces*. For any $1 \leq p \leq \infty$ and $0 < \alpha < 1$, let $\Lambda^p_\alpha(\mathbf{R}^d)$ be the space of functions f whose norm

$$\|f\|_{\Lambda^p_\alpha(\mathbf{R}^d)} := \|f\|_{L^p(\mathbf{R}^d)} + \sup_{x \in \mathbf{R}^d \backslash \{0\}} \frac{\|\tau_x f - f\|_{L^p(\mathbf{R}^d)}}{|x|^\alpha}$$

is finite, where $\tau_x(y) := f(y - x)$ is the translation of f by x. Note that $\Lambda^\infty_\alpha(\mathbf{R}^d) = C^{0,\alpha}(\mathbf{R}^d)$ (with equivalent norms).

(i) For any $0 < \alpha < 1$, establish the inclusions $\Lambda^2_{\alpha+\varepsilon}(\mathbf{R}^d) \subset H^\alpha(\mathbf{R}^d) \subset \Lambda^2_\alpha(\mathbf{R}^d)$ for any $0 < \varepsilon < 1 - \alpha$. (*Hint*: take Fourier transforms and work in frequency space.)

(ii) Let $\phi \in C^\infty_c(\mathbf{R}^d)$ be a bump function, and let ϕ_n be the approximations to the identity $\phi_n(x) := 2^{dn}\phi(2^n x)$. If $f \in \Lambda^p_\alpha(\mathbf{R}^d)$, show that one has the equivalence

$$\|f\|_{\Lambda^p_\alpha(\mathbf{R}^d)} \sim \|f\|_{L^p(\mathbf{R}^d)} + \sup_{n \geq 0} 2^{\alpha n}\|f * \phi_{n+1} - f * \phi_n\|_{L^p(\mathbf{R}^d)}$$

where we use $x \sim y$ to denote the assertion that x and y are comparable up to constants depending on p, d, α. (*Hint*: To upper bound $\|\tau_x f - f\|_{L^p(\mathbf{R}^d)}$ for $|x| \leq 1$, express f as a telescoping sum of $f * \phi_{n+1} - f * \phi_n$ for $2^{-n} \leq x$, plus a final term $f * \phi_{n_0}$ where 2^{-n_0} is comparable to x.)

(iii) If $1 \leq p \leq q \leq \infty$ and $0 < \alpha < 1$ are such that $\frac{d}{p} - \alpha < \frac{d}{q}$, show that $\Lambda^p_\alpha(\mathbf{R}^d)$ embeds continuously into $L^q(\mathbf{R}^d)$. (*Hint*: express $f(x)$ as $f * \phi_1 * \phi_0$ plus a telescoping series of $f * \phi_{n+1} * \phi_n - f * \phi_n * \phi_{n-1}$, where ϕ_n is as in the previous exercise. The additional convolution is in place in order to apply Young's inequality.)

The functions $f * \phi_{n+1} - f * \phi_n$ are crude versions of *Littlewood-Paley projections*, which play an important role in harmonic analysis and nonlinear wave and dispersive equations.

Exercise 1.14.34 (Sobolev trace theorem, special case)**.** Let $s > 1/2$. For any $f \in C^\infty_c(\mathbf{R}^d)$, establish the *Sobolev trace inequality*

$$\|f \restriction_{\mathbf{R}^{d-1}}\|_{H^{s-1/2}(\mathbf{R}^d)} \leq C\|f\|_{H^s(\mathbf{R}^d)}$$

where C depends only on d and s, and $f \restriction_{\mathbf{R}^{d-1}}$ is the restriction of f to the standard hyperplane $\mathbf{R}^{d-1} \equiv \mathbf{R}^{d-1} \times \{0\} \subset \mathbf{R}^d$. (*Hint*: Convert everything to L^2-based statements involving the Fourier transform of f, and use Schur's test, see Lemma 1.11.14.)

Exercise 1.14.35. (i) Show that if $f \in H^s(\mathbf{R}^d)$ for some $s \in \mathbf{R}$, and $g \in C^\infty(\mathbf{R}^d)$, then $fg \in H^s(\mathbf{R}^d)$ (note that this product has to be defined in the sense of tempered distributions if s is negative), and the map $f \mapsto fg$ is continuous from $H^s(\mathbf{R}^d)$ to $H^s(\mathbf{R}^d)$. (*Hint*: As with the previous exercise, convert everything to L^2-based statements involving the Fourier transform of f, and use Schur's test.)

(ii) Let L be a partial differential operator of order m with coefficients in $C^\infty(\mathbf{R}^d)$ for some $m \geq 0$. Show that L maps $H^s(\mathbf{R}^d)$ continuously to $H^{s-m}(\mathbf{R}^d)$ for all $s \in \mathbf{R}$.

Now we consider a partial converse to Exercise 1.14.35.

Exercise 1.14.36 (Elliptic regularity)**.** Let $m \geq 0$, and let

$$L = \sum_{j_1,\ldots,j_d \geq 0; j_1+\ldots+j_d=m} c_{j_1,\ldots,j_d} \frac{\partial^d}{\partial x_{j_1} \ldots \partial x_{j_d}}$$

be a constant-coefficient homogeneous differential operator of order m. Define the *symbol* $l : \mathbf{R}^d \to \mathbf{C}$ of L to be the homogeneous polynomial of degree m, defined by the formula

$$L(\xi_1, \ldots, \xi_d) := \sum_{j_1,\ldots,j_d \geq 0; j_1+\ldots+j_d=m} c_{j_1,\ldots,j_d} \xi_{j_1} \ldots \xi_{j_d}.$$

We say that L is *elliptic* if one has the lower bound

$$l(\xi) \geq c|\xi|^m$$

for all $\xi \in \mathbf{R}^d$ and some constant $c > 0$. Thus, for instance, the Laplacian is elliptic. Another example of an elliptic operator is the Cauchy-Riemann operator $\frac{\partial}{\partial x_1} - i\frac{\partial}{\partial x_2}$ in $\mathbf{R}^2$. On the other hand, the heat operator $\frac{\partial}{\partial t} - \Delta$, the Schrödinger operator $i\frac{\partial}{\partial t} + \Delta$, and the wave operator $-\frac{\partial^2}{\partial t^2} + \Delta$ are not elliptic on $\mathbf{R}^{1+d}$.

(i) Show that if L is elliptic of order m, and f is a tempered distribution such that $f, Lf \in H^s(\mathbf{R}^d)$, then $f \in H^{s+m}(\mathbf{R}^d)$, and that one has the bound

$$\|f\|_{H^{s+m}(\mathbf{R}^d)} \leq C(\|f\|_{H^s(\mathbf{R}^d)} + \|Lf\|_{H^s(\mathbf{R}^d)}) \tag{1.124}$$

for some C depending on s, m, d, L. (*Hint*: Once again, rewrite everything in terms of the Fourier transform $\hat{f}$ of f.)

(ii) Show that if L is a constant-coefficient differential operator of m which is *not* elliptic, then the estimate (1.124) fails.

(iii) Let $f \in L^2_{\mathrm{loc}}(\mathbf{R}^d)$ be a function which is locally in L^2, and let L be an elliptic operator of order m. Show that if $Lf = 0$, then f is smooth. (*Hint*: First show inductively that $f\phi \in H^k(\mathbf{R}^d)$ for every test function ϕ and every natural number $k \geq 0$.)

Remark 1.14.10. The symbol l of an elliptic operator (with real coefficients) tends to have level sets that resemble ellipsoids, hence the name. In contrast, the symbol of *parabolic* operators such as the heat operator $\frac{\partial}{\partial t} - \Delta$ has level sets resembling paraboloids, and the

symbol of *hyperbolic* operators such as the wave operator $-\frac{\partial^2}{\partial t^2} + \Delta$ has level sets resembling hyperboloids. The symbol in fact encodes many important features of linear differential operators, in particular controlling whether singularities can form, and how they must propagate in space and/or time; but this topic is beyond the scope of this course.

Notes. This lecture first appeared at `terrytao.wordpress.com/2009/04/30`. Thanks to Antonio, bk, lutfu, PDEbeginner, Polam, timur, and anonymous commenters for corrections.

1.15. Hausdorff dimension

A fundamental characteristic of many mathematical spaces (e.g. vector spaces, metric spaces, topological spaces, etc.) is their *dimension*, which measures the "complexity" or "degrees of freedom" inherent in the space. There is no single notion of dimension; instead, there are a variety of different versions of this concept, with different versions being suitable for different classes of mathematical spaces. Typically, a single mathematical object may have several subtly different notions of dimension that one can place on it, which will be related to each other, and which will often agree with each other in "non-pathological" cases, but can also deviate from each other in many other situations. For instance:

- One can define the dimension of a space X by seeing how it compares to some standard reference spaces, such as $\mathbf{R}^n$ or $\mathbf{C}^n$; one may view a space as having dimension n if it can be (locally or globally) identified with a standard n-dimensional space. The dimension of a vector space or a manifold can be defined in this fashion.
- Another way to define dimension of a space X is as the largest number of "independent" objects one can place inside that space; this can be used to give an alternate notion of dimension for a vector space, or of an algebraic variety, as well as the closely related notion of the *transcendence degree* of a field. The concept of *VC dimension* in machine learning also broadly falls into this category.

- One can also try to define dimension inductively, for instance declaring a space X to be n-dimensional if it can be "separated" somehow by an $n-1$-dimensional object; thus an n-dimensional object will tend to have "maximal chains" of sub-objects of length n (or $n+1$, depending on how one initialises the chain and how one defines length). This can give a notion of dimension for a topological space or of a commutative ring (*Krull dimension*).

The notions of dimension as defined above tend to necessarily take values in the natural numbers (or the cardinal numbers); there is no such space as $\mathbf{R}^{\sqrt{2}}$, for instance, nor can one talk about a basis consisting of π linearly independent elements, or a chain of maximal ideals of length e. There is however a somewhat different approach to the concept of dimension which makes no distinction between integer and non-integer dimensions, and is suitable for studying "rough" sets such as *fractals*. The starting point is to observe that in the d-dimensional space $\mathbf{R}^d$, the volume V of a ball of radius R grows like R^d, thus giving the following heuristic relationship

$$\frac{\log V}{\log R} \approx d \tag{1.125}$$

between volume, scale, and dimension. Formalising this heuristic leads to a number of useful notions of dimension for subsets of $\mathbf{R}^n$ (or more generally, for metric spaces), including (upper and lower) *Minkowski dimension* (also known as box-packing dimension or Minkowski-Bougliand dimension), and *Hausdorff dimension*.

Remark 1.15.1. In *K-theory*, it is also convenient to work with "virtual" vector spaces or vector bundles, such as formal differences of such spaces, and which may therefore have a negative dimension; but as far as I am aware there is no connection between this notion of dimension and the metric ones given here.

Minkowski dimension can either be defined externally (relating the external volume of δ-neighbourhoods of a set E to the scale δ) or internally (relating the internal δ-entropy of E to the scale). Hausdorff dimension is defined internally by first introducing the d-dimensional *Hausdorff measure* of a set E for any parameter $0 \leq d <$

∞, which generalises the familiar notions of length, area, and volume to non-integer dimensions, or to rough sets, and is of interest in its own right. Hausdorff dimension has a lengthier definition than its Minkowski counterpart, but is more robust with respect to operations such as countable unions, and is generally accepted as the "standard" notion of dimension in metric spaces. We will compare these concepts against each other later in these notes.

One use of the notion of dimension is to create finer distinctions between various types of "small" subsets of spaces such as $\mathbf{R}^n$, beyond what can be achieved by the usual Lebesgue measure (or Baire category). For instance, a point, line, and plane in $\mathbf{R}^3$ all have zero measure with respect to three-dimensional Lebesgue measure (and are nowhere dense), but of course have different dimensions (0, 1, and 2 respectively). (Another good example is provided by *Kakeya sets*.) This can be used to clarify the nature of various singularities, such as that arising from non-smooth solutions to PDE; a function which is non-smooth on a set of large Hausdorff dimension can be considered less smooth than one which is non-smooth on a set of small Hausdorff dimension, even if both are smooth almost everywhere. While many properties of the singular set of such a function are worth studying (e.g. their *rectifiability*), understanding their dimension is often an important starting point. The interplay between these types of concepts is the subject of *geometric measure theory*.

1.15.1. Minkowski dimension. Before we study the more standard notion of Hausdorff dimension, we begin with the more elementary concept of the (upper and lower) Minkowski dimension of a subset E of a Euclidean space $\mathbf{R}^n$.

There are several equivalent ways to approach Minkowski dimension. We begin with an "external" approach, based on a study of the δ-neighbourhoods $E_\delta := \{x \in \mathbf{R}^n : \mathrm{dist}(x, E) < \delta\}$ of E, where $\mathrm{dist}(x, E) := \inf\{|x - y| : y \in E\}$ and we use the Euclidean metric on $\mathbf{R}^n$. These are open sets in $\mathbf{R}^n$ and therefore have a d-dimensional volume (or Lebesgue measure) $\mathrm{vol}^d(E_\delta)$. To avoid divergences, let us assume for now that E is bounded, so that the E_δ have finite volume.

Let $0 \leq d \leq n$. Suppose E is a bounded portion of a k-dimensional subspace, e.g. $E = B^d(0,1) \times \{0\}^{n-d}$, where $B^d(0,1) \subset \mathbf{R}^d$ is the unit ball in $\mathbf{R}^d$ and we identify $\mathbf{R}^n$ with $\mathbf{R}^d \times \mathbf{R}^{n-d}$ in the usual manner. Then we see from the triangle inequality that

$$B^d(0,1) \times B^{n-d}(0,\delta) \subset E_\delta \subset B^d(0,2) \times B^{n-d}(0,\delta)$$

for all $0 < \delta < 1$, which implies that

$$c\delta^{n-d} \leq \mathrm{vol}^n(E_\delta) \leq C\delta^{n-d}$$

for some constants $c, C > 0$ depending only on n, d. In particular, we have

$$\lim_{\delta \to 0} n - \frac{\log \mathrm{vol}^n(E_\delta)}{\log \delta} = d$$

(compare with (1.125)). This motivates our first definition of Minkowski dimension:

Definition 1.15.2. Let E be a bounded subset of $\mathbf{R}^n$. The *upper Minkowski dimension* $\overline{\dim}_M(E)$ is defined as

$$\overline{\dim}_M(E) := \limsup_{\delta \to 0} n - \frac{\log \mathrm{vol}^n(E_\delta)}{\log \delta}$$

and the *lower Minkowski dimension* $\underline{\dim}_M(E)$ is defined as

$$\underline{\dim}_M(E) := \liminf_{\delta \to 0} n - \frac{\log \mathrm{vol}^n(E_\delta)}{\log \delta}.$$

If the upper and lower Minkowski dimensions match, we refer to $\dim_M(E) := \overline{\dim}_M(E) = \underline{\dim}_M(E)$ as the Minkowski dimension of E. In particular, the empty set has a Minkowski dimension of $-\infty$.

Unwrapping all the definitions, we have the following equivalent formulation, where E is a bounded subset of $\mathbf{R}^n$ and $\alpha \in \mathbf{R}$:

- We have $\overline{\dim}_M(E) < \alpha$ iff for every $\varepsilon > 0$, one has $\mathrm{vol}^d(E_\delta) \leq C\delta^{n-d-\varepsilon}$ for *all sufficiently small* $\delta > 0$ and some $C > 0$.
- We have $\underline{\dim}_M(E) < \alpha$ iff for every $\varepsilon > 0$, one has $\mathrm{vol}^d(E_\delta) \leq C\delta^{n-d-\varepsilon}$ for *arbitrarily small* $\delta > 0$ and some $C > 0$.
- We have $\overline{\dim}_M(E) > \alpha$ iff for every $\varepsilon > 0$, one has $\mathrm{vol}^d(E_\delta) \geq c\delta^{n-d-\varepsilon}$ for *arbitrarily small* $\delta > 0$ and some $c > 0$.
- We have $\overline{\dim}_M(E) > \alpha$ iff for every $\varepsilon > 0$, one has $\mathrm{vol}^d(E_\delta) \geq c\delta^{n-d-\varepsilon}$ for *all sufficiently small* $\delta > 0$ and some $c > 0$.

Exercise 1.15.1. (i) Let $C \subset \mathbf{R}$ be the *Cantor set* consisting of all base 4 strings $\sum_{i=1}^{\infty} a_i 4^{-i}$, where each a_i takes values in $\{0, 3\}$. Show that C has Minkowski dimension $1/2$. (*Hint*: approximate any small δ by a negative power of 4.)

(ii) Let $C' \subset \mathbf{R}$ be the Cantor set consisting of all base 4 strings $\sum_{i=1}^{\infty} a_i 4^{-i}$, where each a_i takes values in $\{0, 3\}$ when $(2k)! \leq i < (2k+1)!$ for some integer $k \geq 0$, and a_i is arbitrary for the other values of i. Show that C' has a lower Minkowski dimension of $1/2$ and an upper Minkowski dimension of 1.

Exercise 1.15.2. Suppose that $E \subset \mathbf{R}^n$ is a compact set with the property that there exist $0 < r < 1$ and an integer $k > 1$ such that E is equal to the union of k disjoint translates of $r \cdot E := \{rx : x \in E\}$. (This is a special case of a *self-similar fractal*; the *Cantor set* is a typical example.) Show that E has Minkowski dimension $\frac{\log k}{\log 1/r}$.

If the k translates of $r \cdot E$ are allowed to overlap, establish the upper bound $\overline{\dim}_M(E) \leq \frac{\log k}{\log 1/r}$.

It is clear that we have the inequalities

$$0 \leq \underline{\dim}_M(E) \leq \overline{\dim}_M(E) \leq n$$

for non-empty bounded $E \subset \mathbf{R}^n$, and the monotonicity properties

$$\underline{\dim}_M(E) \leq \underline{\dim}_M(F); \quad \overline{\dim}_M(E) \leq \overline{\dim}_M(F)$$

whenever $E \subset F \subset \mathbf{R}^n$ are bounded sets. It is thus natural to extend the definitions of lower and upper Minkowski dimension to unbounded sets E by defining

$$\underline{\dim}_M(E) := \sup_{F \subset E, \text{ bounded}} \underline{\dim}_M(F) \tag{1.126}$$

and

$$\overline{\dim}_M(E) := \sup_{F \subset E, \text{ bounded}} \overline{\dim}_M(F). \tag{1.127}$$

In particular, we easily verify that d-dimensional subspaces of $\mathbf{R}^n$ have Minkowski dimension d.

Exercise 1.15.3. Show that any subset of $\mathbf{R}^n$ with lower Minkowski dimension less than n has Lebesgue measure zero. In particular,

any subset $E \subset \mathbf{R}^n$ of positive Lebesgue measure must have full Minkowski dimension $\dim_M(E) = n$.

Now we turn to other formulations of Minkowski dimension. Given a bounded set E and $\delta > 0$, we make the following definitions:

- $\mathcal{N}_\delta^{\mathrm{ext}}(E)$ (the *external δ-covering number* of E) is the fewest number of open balls of radius δ with centres in $\mathbf{R}^n$ needed to cover E.
- $\mathcal{N}_\delta^{\mathrm{int}}(E)$ (the *internal δ-covering number* of E) is the fewest number of open balls of radius δ with centres in E needed to cover E.
- $\mathcal{N}_\delta^{\mathrm{net}}(E)$ (the *δ-metric entropy*) is the cardinality of the largest δ-net in E, i.e. the largest set $x_1, \ldots, x_k$ in E such that $|x_i - x_j| \geq \delta$ for every $1 \leq i < j \leq k$.
- $\mathcal{N}_\delta^{\mathrm{pack}}(E)$ (the *δ-packing number* of E) is the largest number of disjoint open balls one can find of radius δ with centres in E.

These three quantities are closely related to each other, and to the volumes $\mathrm{vol}^n(E_\delta)$:

Exercise 1.15.4. For any bounded set $E \subset \mathbf{R}^n$ and any $\delta > 0$, show that

$$\mathcal{N}_{2\delta}^{\mathrm{net}}(E) = \mathcal{N}_\delta^{\mathrm{pack}}(E) \leq \frac{\mathrm{vol}^n(E_\delta)}{\mathrm{vol}^n(B^n(0,\delta))} \leq$$
$$\leq \mathcal{N}_\delta^{\mathrm{ext}}(E) \leq \mathcal{N}_\delta^{\mathrm{int}}(E) \leq \mathcal{N}_\delta^{\mathrm{net}}(E).$$

As a consequence of this exercise, we see that

$$\overline{\dim}_M(E) = \limsup_{\delta \to 0} \frac{\mathcal{N}_\delta^*(E)}{\log 1/\delta} \tag{1.128}$$

and

$$\underline{\dim}_M(E) = \liminf_{\delta \to 0} \frac{\mathcal{N}_\delta^*(E)}{\log 1/\delta}. \tag{1.129}$$

where $*$ is any of $\mathrm{ext}, \mathrm{int}, \mathrm{net}, \mathrm{pack}$.

One can now take the formulae (1.128), (1.129) as the *definition* of Minkowski dimension for bounded sets (and then use (1.126), (1.127) to extend to unbounded sets). The formulations (1.128), (1.129) for

$* = \mathrm{int}, \mathrm{net}, \mathrm{pack}$ have the advantage of being *intrinsic* - they only involve E, rather than the ambient space $\mathbf{R}^n$. For metric spaces, one still has a partial analogue of Exercise 1.15.4, namely

$$\mathcal{N}^{\mathrm{net}}_{2\delta}(E) \leq \mathcal{N}^{\mathrm{pack}}_{\delta}(E) \leq \mathcal{N}^{\mathrm{int}}_{\delta}(E) \leq \mathcal{N}^{\mathrm{net}}_{\delta}(E).$$

As such, these formulations of Minkowski dimension extend without any difficulty to arbitrary bounded metric spaces (E, d) (at least when the spaces are locally compact), and then to unbounded metric spaces by (1.126), (1.127).

Exercise 1.15.5. If $\phi : (X, d_X) \to (Y, d_Y)$ is a Lipschitz map between metric spaces, show that $\overline{\dim}_M(\phi(E)) \leq \overline{\dim}_M(E)$ and $\underline{\dim}_M(\phi(E)) \leq \underline{\dim}_M(E)$ for all $E \subset X$. Conclude in particular that the graph $\{(x, \phi(x)) : x \in \mathbf{R}^d\}$ of any Lipschitz function $\phi : \mathbf{R}^d \to \mathbf{R}^{n-d}$ has Minkowski dimension d, and the graph of any measurable function $\phi : \mathbf{R}^d \to \mathbf{R}^{n-d}$ has Minkowski dimension at least d.

Note however that the dimension of graphs can become larger than that of the base in the non-Lipschitz case:

Exercise 1.15.6. Show that the graph $\{(x, \sin \frac{1}{x}) : 0 < x < 1\}$ has Minkowski dimension $3/2$.

Exercise 1.15.7. Let (X, d) be a bounded metric space. For each $n \geq 0$, let E_n be a maximal 2^{-n}-net of X (thus the cardinality of E_n is $\mathcal{N}^{net}_{2^{-n}}(X)$). Show that for any continuous function $f : X \to \mathbf{R}$ and any $x_0 \in X$, one has the inequality

$$\sup_{x \in X} f(x) \leq \sup_{x_0 \in E_0} f(x_0) +$$

$$+ \sum_{n=0}^{\infty} \sup_{x_n \in E_n, x_{n+1} \in E_{n+1} : |x_n - x_{n+1}| \leq \frac{3}{2} 2^{-n}} (f(x_n) - f(x_{n+1})).$$

(*Hint*: For any $x \in X$, define $x_n \in E_n$ to be the nearest point in E_n to x, and use a telescoping series.) This inequality (and variants thereof), which replaces a continuous supremum of a function $f(x)$ by a sum of discrete suprema of differences $f(x_n) - f(x_{n+1})$ of that function, is the basis of the *generic chaining* technique in probability, used to estimate the supremum of a continuous family of random processes. It is particularly effective when combined with bounds on

the metric entropy $\mathcal{N}_{2^{-n}}^{net}(X)$, which of course is closely related to the Minkowski dimension of X, and with *large deviation* bounds on the differences $f(x_n) - f(x_{n+1})$. A good reference for generic chaining is [**Ta2005**].

Exercise 1.15.8. If $E \subset \mathbf{R}^n$ and $F \subset \mathbf{R}^m$ are bounded sets, show that

$$\underline{\dim}_M(E) + \underline{\dim}_M(F) \leq \underline{\dim}_M(E \times F)$$

and

$$\overline{\dim}_M(E \times F) \leq \overline{\dim}_M(E) + \overline{\dim}_M(F).$$

Give a counterexample that shows that either of the inequalities here can be strict. (*Hint*: There are many possible constructions; one of them is a modification of Exercise 1.15.1(ii).)

It is easy to see that Minkowski dimension reacts well to finite unions, and more precisely that

$$\overline{\dim}_M(E \cup F) = \max(\overline{\dim}_M(E), \overline{\dim}_M(F))$$

and

$$\underline{\dim}_M(E \cup F) = \max(\underline{\dim}_M(E), \underline{\dim}_M(F))$$

for any $E, F \subset \mathbf{R}^n$. However, it does not respect countable unions. For instance, the rationals $\mathbf{Q}$ have Minkowski dimension 1, despite being the countable union of points, which of course have Minkowski dimension 0. More generally, it is not difficult to see that any set $E \subset \mathbf{R}^n$ has the same upper or lower Minkowski dimension as its topological closure $\overline{E}$, since both sets have the same δ-neighbourhoods. Thus we see that the notion of Minkowski dimension misses some of the fine structure of a set E, in particular the presence of "holes" within the set. We now turn to the notion of Hausdorff dimension, which rectifies some of these defects.

1.15.2. Hausdorff measure. The Hausdorff approach to dimension begins by noting that d-dimensional objects in $\mathbf{R}^n$ tend to have a meaningful d-dimensional measure to assign to them. For instance,

the 1-dimensional boundary of a polygon has a perimeter, the 0-dimensional vertices of that polygon have a cardinality, and the polygon itself has an area. So to define the notion of a d-Hausdorff dimensional set, we will first define the notion of the d-dimensional Hausdorff measure $\mathcal{H}^d(E)$ of a set E.

To do this, let us quickly review one of the (many) constructions of n-dimensional Lebesgue measure, which we are denoting here by vol^n. One way to build this measure is to work with half-open boxes $B = \prod_{i=1}^n [a_i, b_i)$ in $\mathbf{R}^n$, which we assign a volume of $|B| := \prod_{i=1}^n (b_i - a_i)$. Given this notion of volume for boxes, we can then define the *outer Lebesgue measure* $(\mathrm{vol}^n)^*(E)$ of any set $E \subset \mathbf{R}^n$ by the formula

$$(\mathrm{vol}^n)^*(E) := \inf\{\sum_{k=1}^{\infty} |B_k| : B_k \text{ covers } E\}$$

where the infimum ranges over all at most countable collections $B_1, B_2, \ldots$ of boxes that cover E. One easily verifies that $(\mathrm{vol}^n)^*$ is indeed an outer measure (i.e. it is monotone, countably subadditive, and assigns zero to the empty set). We then define a set $A \subset \mathbf{R}^n$ to be $(\mathrm{vol}^n)^*$-*measurable* if one has the additivity property

$$(\mathrm{vol}^n)^*(E) = (\mathrm{vol}^n)^*(E \cap A) + (\mathrm{vol}^n)^*(E \backslash A)$$

for all $E \subset \mathbf{R}^n$. By *Carathéodory's theorem*, the space of $(\mathrm{vol}^n)^*$-measurable sets is a σ-algebra, and outer Lebesgue measure is a countably additive measure on this σ-algebra, which we denote vol^n. Furthermore, one easily verifies that every box B is $(\mathrm{vol}^n)^*$-measurable, which soon implies that every Borel set is also; thus Lebesgue measure is a Borel measure (though it can of course measure some non-Borel sets also).

Finally, one needs to verify that the Lebesgue measure $\mathrm{vol}^n(B)$ of a box is equal to its classical volume $|B|$; the above construction trivially gives $\mathrm{vol}^n(B) \leq |B|$ but the converse is not as obvious. This is in fact a rather delicate matter, relying in particular on the completeness of the reals; if one replaced $\mathbf{R}$ by the rationals $\mathbf{Q}$, for instance, then all the above constructions go through but now boxes have Lebesgue measure zero (why?). See [**Fo2000**, Chapter 1], for instance, for details.

Anyway, we can use this construction of Lebesgue measure as a model for building d-dimensional Hausdorff measure. Instead of using half-open boxes as the building blocks, we will instead work with the open balls $B(x,r)$. For d-dimensional measure, we will assign each ball $B(x,r)$ a measure r^d (cf. (1.125)). We can then define the *unlimited Hausdorff content* $h_{d,\infty}(E)$ of a set $E \subset \mathbf{R}^n$ by the formula

$$h_{d,\infty}(E) := \inf\{\sum_{k=1}^{\infty} r_k^d : B(x_k, r_k) \text{ covers } E\}$$

where the infimum ranges over all at most countable families of balls that cover E. (Note that if E is compact, then it would suffice to use finite coverings, since every open cover of E has a finite subcover. But in general, for non-compact E we must allow the use of infinitely many balls.)

As with Lebesgue measure, $h_{d,\infty}$ is easily seen to be an outer measure, and one could define the notion of a $h_{d,\infty}$-measurable set on which Carathéodory's theorem applies to build a countably additive measre. Unfortunately, a key problem arises: once d is less than n, most sets cease to be $h_{d,\infty}$-measurable! We illustrate this in the one-dimensional case with $n = 1$ and $d = 1/2$, and consider the problem of computing the unlimited Hausdorff content $h_{1/2,\infty}([a,b])$. On the one hand, this content is at most $|\frac{b-a}{2}|^{1/2}$, since one can cover $[a,b]$ by the ball of radius $\frac{b-a}{2} + \varepsilon$ centred at $\frac{a+b}{2}$ for any $\varepsilon > 0$. On the other hand, the content is also at *least* $|\frac{b-a}{2}|^{1/2}$. To see this, suppose we cover $[a,b]$ by a finite or countable family of balls $B(x_k, r_k)$ (one can reduce to the finite case by compactness, though it isn't necessary to do so here). The total one-dimensional Lebesgue measure $\sum_k 2r_k$ of these balls must equal or exceed the Lebesgue measure of the entire interval $|b-a|$, thus

$$\sum_k r_k \geq \frac{|b-a|}{2}.$$

From the inequality $\sum_k r_k \leq (\sum_k r_k^{1/2})^2$ (which is obvious after expanding the RHS and discarding cross-terms) we see that

$$\sum_k r_k^{1/2} \geq (\frac{|b-a|}{2})^{1/2}$$

and the claim follows.

We now see some serious breakdown of additivity: for instance, the unlimited $1/2$-dimensional content of $[0,2]$ is 1, despite being the disjoint union of $[0,1]$ and $(1,2]$, which each have an unlimited content of $1/\sqrt{2}$. In particular, this shows that $[0,1]$ (for instance) is not measurable with respect to the unlimited content. The basic problem here is that the most efficient cover of a union such as $[0,1] \cup (1,2]$ for the purposes of unlimited $1/2$-dimensional content is not coming from covers of the separate components $[0,1]$ and $(1,2]$ of that union, but is instead coming from one giant ball that covers $[0,2]$ directly.

To fix this, we will *limit* the Hausdorff content by working only with small balls. More precisely, for any $r > 0$, we define the Hausdorff content $h_{d,r}(E)$ of a set $E \subset \mathbf{R}^n$ by the formula

$$h_{d,r}(E) := \inf\{\sum_{k=1}^{\infty} r_k^d : B(x_k, r_k) \text{ covers } E; r_k \leq r\}$$

where the balls $B(x_k, r_k)$ are now restricted to be less than or equal to r in radius. This quantity is increasing in r, and we then define the *Hausdorff outer measure* $(\mathcal{H}^d)^*(E)$ by the formula

$$(\mathcal{H}^d)^*(E) := \lim_{r \to 0} h_{d,r}(E).$$

(This is analogous to the Riemann integral approach to volume of sets, covering them by balls, boxes, or rectangles of increasingly small size; this latter approach is also closely connected to the Minkowski dimension concept studied earlier. The key difference between the Lebesgue/Hausdorff approach and the Riemann/Minkowski approach is that in the former approach one allows the balls or boxes to be countable in number, and to be variable in size, whereas in the latter approach the cover is finite and uniform in size.)

Exercise 1.15.9. Show that if $d > n$, then $(\mathcal{H}^d)^*(E) = 0$ for all $E \subset \mathbf{R}^n$. Thus d-dimensional Hausdorff measure is only a non-trivial concept for subsets of $\mathbf{R}^n$ in the regime $0 \leq d \leq n$.

Since each of the $h_{d,r}$ are outer measures, $(\mathcal{H}^d)^*$ is also. But the key advantage of moving to the Hausdorff measure rather than Hausdorff content is that we obtain a lot more additivity. For instance:

Exercise 1.15.10. Let E, F be subsets of $\mathbf{R}^n$ which have a non-zero separation, i.e. the quantity $\text{dist}(E,F) = \inf\{|x-y| : x \in E, y \in F\}$

is strictly positive. Show that $(\mathcal{H}^d)^*(E \cup F) = (\mathcal{H}^d)^*(E) + (\mathcal{H}^d)^*(F)$. (*Hint*: one inequality is easy. For the other, observe that any small ball can intersect E or intersect F, but not both.)

One consequence of this is that there is a large class of measurable sets:

Proposition 1.15.3. *Let $d \geq 0$. Then every Borel subset of $\mathbf{R}^n$ is $(\mathcal{H}^d)^*$-measurable.*

Proof. Since the collection of $(\mathcal{H}^d)^*$-measurable sets is a σ-algebra, it suffices to show the claim for closed sets A. (It will be slightly more convenient technically to work with closed sets rather than open ones here.) Thus, we take an arbitrary set $E \subset \mathbf{R}^n$ and seek to show that

$$(\mathcal{H}^d)^*(E) = (\mathcal{H}^d)^*(E \cap A) + (\mathcal{H}^d)^*(E \backslash A).$$

We may assume that $(\mathcal{H}^d)^*(E \cap A)$ and $(\mathcal{H}^d)^*(E \backslash A)$ are both finite, since the claim is obvious otherwise from monotonicity.

From Exercise 1.15.10 and the fact that $(\mathcal{H}^d)^*$ is an outer measure, we already have

$$(\mathcal{H}^d)^*(E \cap A) + (\mathcal{H}^d)^*(E \backslash A_{1/m}) \leq (\mathcal{H}^d)^*(E) \leq (\mathcal{H}^d)^*(E \cap A) + (\mathcal{H}^d)^*(E \backslash A),$$

where $A_{1/m}$ is the $1/m$-neighbourhood of A. So it suffices to show that

$$\lim_{m \to \infty} (\mathcal{H}^d)^*(E \backslash A_{1/m}) = (\mathcal{H}^d)^*(E \backslash A).$$

For any m, we have the telescoping sum $E \backslash A = (E \backslash A_{1/m}) \cup \bigcup_{l > m} F_l$, where $F_l := (E \backslash A_{1/(l+1)}) \cap A_l$, and thus by countable subadditivity and monotonicity

$$(\mathcal{H}^d)^*(E \backslash A_{1/m}) \leq (\mathcal{H}^d)^*(E \backslash A) \leq (\mathcal{H}^d)^*(E \backslash A_{1/m}) + \sum_{l > m} (\mathcal{H}^d)^*(F_l)$$

so it suffices to show that the sum $\sum_{l=1}^{\infty} (\mathcal{H}^d)^*(F_l)$ is absolutely convergent.

Consider the even-indexed sets $F_2, F_4, F_6, \ldots$. These sets are separated from each other, so by many applications of Exercise 1.15.10 followed by monotonicity we have

$$\sum_{l=1}^{L} (\mathcal{H}^d)^*(F_{2l}) = (\mathcal{H}^d)^*(\bigcup_{l=1}^{L} F_{2l}) \leq (\mathcal{H}^d)^*(E \backslash A) < \infty$$

for all L, and thus $\sum_{l=1}^{\infty} (\mathcal{H}^d)^*(F_{2l})$ is absolutely convergent. Similarly for $\sum_{l=1}^{\infty} (\mathcal{H}^d)^*(F_{2l-1})$, and the claim follows. □

On the $(\mathcal{H}^d)^*$-measurable sets E, we write $\mathcal{H}^d(E)$ for $(\mathcal{H}^d)^*(E)$, thus $\mathcal{H}^d$ is a Borel measure on $\mathbf{R}^n$. We now study what this measure looks like for various values of d. The case $d = 0$ is easy:

Exercise 1.15.11. Show that every subset of $\mathbf{R}^n$ is $(\mathcal{H}^0)^*$-measurable, and that $\mathcal{H}^0$ is counting measure.

Now we look at the opposite case $d = n$. It is easy to see that any Lebesgue-null set of $\mathbf{R}^n$ has n-dimensional Hausdorff measure zero (since it may be covered by balls of arbitrarily small total content). Thus n-dimensional Hausdorff measure is absolutely continuous with respect to Lebesgue measure, and we thus have $\frac{d\mathcal{H}^n}{d\,\mathrm{vol}^n} = c$ for some locally integrable function c. As Hausdorff measure and Lebesgue measure are clearly translation-invariant, c must also be translation-invariant and thus constant. We therefore have

$$\mathcal{H}^n = c\,\mathrm{vol}^n$$

for some constant $c \geq 0$.

We now compute what this constant is. If ω_n denotes the volume of the unit ball $B(0,1)$, then we have

$$\sum_k r_k^n = \frac{1}{\omega_n} \sum_k \mathrm{vol}^n(B(x_k, r_k)) \geq \frac{1}{\omega_n} \mathrm{vol}^n(\bigcup_k B(x_k, r_k))$$

for any at most countable collection of balls $B(x_k, r_k)$. Taking infima, we conclude that

$$\mathcal{H}^n \geq \frac{1}{\omega_n} \mathrm{vol}^n$$

and so $c \geq \frac{1}{\omega_n}$.

In the opposite direction, observe from Exercise 1.15.4 that given any $0 < r < 1$, one can cover the unit cube $[0,1]^n$ by at most $C_n r^{-n}$ balls of radius r, where C_n depends only on n; thus

$$\mathcal{H}^n([0,1]^n) \leq C_n$$

and so $c \leq C_n$; in particular, c is finite.

We can in fact compute c explicitly (although knowing that c is finite and non-zero already suffices for many applications):

Lemma 1.15.4. *We have $c = \frac{1}{\omega_n}$, or in other words $\mathcal{H}^n = \frac{1}{\omega_n} \mathrm{vol}^n$. (In particular, a ball $B^n(x,r)$ has n-dimensional Hausdorff measure r^n.)*

Proof. Let us consider the Hausdorff measure $\mathcal{H}^n([0,1]^n)$ of the unit cube. By definition, for any $\varepsilon > 0$ one can find an $0 < r < 1/2$ such that

$$h_{n,r}([0,1]^n) \geq \mathcal{H}^n([0,1]^n) - \varepsilon.$$

Observe (using Exercise 1.15.4) that we can find at least $c_n r^{-n}$ disjoint balls $B(x_1,r), \ldots, B(x_k,r)$ of radius r inside the unit cube. We then observe that

$$h_{n,r}([0,1]^n) \leq kr^n + \mathcal{H}^n([0,1]^n \backslash \bigcup_{i=1}^k B(x_k,r)).$$

On the other hand,

$$\mathcal{H}^n([0,1]^n \backslash \bigcup_{i=1}^k B(x_k,r)) = c\,\mathrm{vol}^n([0,1]^n \backslash \bigcup_{i=1}^k B(x_k,r)) = c(1-k\omega_n r^n);$$

putting all this together, we obtain

$$c = \mathcal{H}^n([0,1]^n) \leq kr^n + c(1 - k\omega_n r^n) + \varepsilon$$

which rearranges as

$$1 - c\omega_n \geq \frac{\varepsilon}{kr^n}.$$

Since kr^n is bounded below by c_n, we can then send $\varepsilon \to 0$ and conclude that $c \geq \frac{1}{\omega_n}$; since we already showed $c \leq \frac{1}{\omega_n}$, the claim follows. □

Thus n-dimensional Hausdorff measure is an explicit constant multiple of n-dimensional Lebesgue measure. The same argument shows that for integers $0 < d < n$, the restriction of d-dimensional Hausdorff measure to any d-dimensional linear subspace (or affine subspace) V is equal to the constant $\frac{1}{\omega_d}$ times d-dimensional Lebesgue measure on V. (This shows, by the way, that $\mathcal{H}^d$ is not a σ-finite measure on $\mathbf{R}^n$ in general, since one can partition $\mathbf{R}^n$ into uncountably many d-dimensional affine subspaces. In particular, it is *not* a Radon measure in general).

One can then compute d-dimensional Hausdorff measure for other sets than subsets of d-dimensional affine subspaces by changes of variable. For instance:

Exercise 1.15.12. Let $0 \leq d \leq n$ be an integer, let Ω be an open subset of $\mathbf{R}^d$, and let $\phi : \Omega \to \mathbf{R}^n$ be a smooth injective map which is *non-degenerate* in the sense that the Hessian $D\phi$ (which is a $d \times n$ matrix) has full rank at every point of Ω. For any compact subset E of Ω, establish the formula

$$\mathcal{H}^d(\phi(E)) = \int_E J \; d\mathcal{H}^d = \frac{1}{\omega_d} \int_E J \; d\,\mathrm{vol}^d$$

where the *Jacobian* J is the square root of the sum of squares of all the determinants of the $d \times d$ minors of the $d \times n$ matrix $D\phi$. (*Hint*: By working locally, one can assume that ϕ is the graph of some map from Ω to $\mathbf{R}^{n-d}$, and so can be inverted by the projection function; by working even more locally, one can assume that the Jacobian is within an epsilon of being constant. The image of a small ball in Ω then resembles a small ellipsoid in $\phi(\Omega)$, and conversely the projection of a small ball in $\phi(\Omega)$ is a small ellipsoid in Ω. Use some linear algebra and several variable calculus to relate the content of these ellipsoids to the radius of the ball.) It is possible to extend this formula to Lipschitz maps $\phi : \Omega \to \mathbf{R}^n$ that are not necessarily injective, leading to the *area formula*

$$\int_{\phi(E)} \#(\phi^{-1}(y)) \; d\mathcal{H}^d(y) = \frac{1}{\omega_d} \int_E J \; d\,\mathrm{vol}^d$$

for such maps, but we will not prove this formula here.

From this exercise we see that d-dimensional Hausdorff measure does coincide to a large extent with the d-dimensional notion of surface area; for instance, for a simple smooth curve $\gamma : [a,b] \to \mathbf{R}^n$ with everywhere non-vanishing derivative, the $\mathcal{H}^1$ measure of $\gamma([a,b])$ is equal to its classical length $|\gamma| = \int_a^b |\gamma'(t)| \; dt$. One can also handle a certain amount of singularity (e.g. piecewise smooth non-degenerate curves rather than everywhere smooth non-degenerate curves) by exploiting the countable additivity of $\mathcal{H}^1$ measure, or by using the area formula alluded to earlier.

Now we see how the Hausdorff dimension varies in d.

Exercise 1.15.13. Let $0 \leq d < d'$, and let $E \subset \mathbf{R}^n$ be a Borel set. Show that if $\mathcal{H}^{d'}(E)$ is finite, then $\mathcal{H}^d(E)$ is zero; equivalently, if $\mathcal{H}^d(E)$ is positive, then $\mathcal{H}^{d'}$ is infinite.

Example 1.15.5. Let $0 \leq d \leq n$ be integers. The unit ball $B^d(0,1) \subset \mathbf{R}^d \subset \mathbf{R}^n$ has a d-dimensional Hausdorff measure of 1 (by Lemma 1.15.4), and so it has zero d'-dimensional Hausdorff dimensional measure for $d' > d$ and infinite d'-dimensional measure for $d' < d$.

On the other hand, we know from Exercise 1.15.11 that $\mathcal{H}^0(E)$ is positive for any non-empty set E, and that $\mathcal{H}^d(E) = 0$ for every $d > n$. We conclude (from the least upper bound property of the reals) that for any Borel set $E \subset \mathbf{R}^n$, there exists a unique number in $[0,n]$, called the *Hausdorff dimension* $\dim_H(E)$ of E, such that $\mathcal{H}^d(E) = 0$ for all $d > \dim_H(E)$ and $\mathcal{H}^d(E) = \infty$ for all $d < \dim_H(E)$. Note that at the critical dimension $d = \dim_H$ itself, we allow $\mathcal{H}^d(E)$ to be zero, finite, or infinite, and we shall shortly see in fact that all three possibilities can occur. By convention, we give the empty set a Hausdorff dimension of $-\infty$. One can also assign Hausdorff dimension to non-Borel sets, but we shall not do so to avoid some (very minor) technicalities.

Example 1.15.6. The unit ball $B^d(0,1) \subset \mathbf{R}^d \subset \mathbf{R}^n$ has Hausdorff dimension d, as does $\mathbf{R}^d$ itself. Note that the former set has finite d-dimensional Hausdorff measure, while the latter has an infinite measure. More generally, any d-dimensional smooth manifold in $\mathbf{R}^n$ has Hausdorff dimension d.

Exercise 1.15.14. Show that the graph $\{(x, \sin\frac{1}{x}) : 0 < x < 1\}$ has Hausdorff dimension 1; compare this with Exercise 1.15.6.

It is clear that Hausdorff dimension is monotone: if $E \subset F$ are Borel sets, then $\dim_H(E) \leq \dim_H(F)$. Since Hausdorff measure is countably additive, it is also not hard to see that Hausdorff dimension interacts well with countable unions:

$$\dim_H(\bigcup_{i=1}^{\infty} E_i) = \sup_{1 \leq i \leq \infty} \dim_H(E_i).$$

Thus for instance the rationals, being a countable union of 0-dimensional points, have Hausdorff dimension 0, in contrast to their Minkowski

dimension of 1. On the other hand, we at least have an inequality between Hausdorff and Minkowski dimension:

Exercise 1.15.15. For any Borel set $E \subset \mathbf{R}^n$, show that $\dim_H(E) \leq \underline{\dim}_M(E) \leq \overline{\dim}_M(E)$. (*Hint*: use (1.129). Which of the choices of $*$ is most convenient to use here?)

It is instructive to compare Hausdorff dimension and Minkowski dimension as follows.

Exercise 1.15.16. Let E be a bounded Borel subset of $\mathbf{R}^n$, and let $d \geq 0$.

- Show that $\overline{\dim}_M(E) \leq d$ if and only if, for every $\varepsilon > 0$ and arbitrarily small $r > 0$, one can cover E by finitely many balls $B(x_1, r_1), \ldots, B(x_k, r_k)$ of radii $r_i = r$ equal to r such that $\sum_{i=1}^k r_i^{d+\varepsilon} \leq \varepsilon$.
- Show that $\underline{\dim}_M(E) \leq d$ if and only if, for every $\varepsilon > 0$ and all sufficiently small $r > 0$, one can cover E by finitely many balls $B(x_1, r_1), \ldots, B(x_k, r_k)$ of radii $r_i = r$ equal to r such that $\sum_{i=1}^k r_i^{d+\varepsilon} \leq \varepsilon$.
- Show that $\dim_H(E) \leq d$ if and only if, for every $\varepsilon > 0$ and $r > 0$, one can cover E by countably many balls $B(x_1, r_1), \ldots$ of radii $r_i \leq r$ at most r such that $\sum_{i=1}^k r_i^{d+\varepsilon} \leq \varepsilon$.

The previous two exercises give ways to upper-bound the Hausdorff dimension; for instance, we see from Exercise 1.15.2 that self-similar fractals E of the type in that exercise (i.e. E is k translates of $r \cdot E$) have Hausdorff dimension at most $\frac{\log k}{\log 1/r}$. To lower bound the Hausdorff dimension of a set E, one convenient way to do so is to find a measure with a certain "dimension" property (analogous to (1.125)) that assigns a positive mass to E:

Exercise 1.15.17. Let $d \geq 0$. A Borel measure μ on $\mathbf{R}^n$ is said to be a *Frostman measure of dimension at most d* if it is compactly supported there exists a constant C such that $\mu(B(x,r)) \leq Cr^d$ for all balls $B(x,r)$ of radius $0 < r < 1$. Show that if μ has dimension at most d, then any Borel set E with $\mu(E) > 0$ has positive d-dimensional Hausdorff content; in particular, $\dim_H(E) \geq d$.

Note that this gives an alternate way to justify the fact that smooth d-dimensional manifolds have Hausdorff dimension d, since on the one hand they have Minkowski dimension d, and on the other hand they support a non-trivial d-dimensional measure, namely Lebesgue measure.

Exercise 1.15.18. Show that the Cantor set in Exercise 1.15.1(i) has Hausdorff dimension $1/2$. More generally, establish the analogue of the first part of Exercise 1.15.2 for Hausdorff measure.

Exercise 1.15.19. Construct a subset of $\mathbf{R}$ of Hausdorff dimension 1 that has zero Lebesgue measure. (*Hint*: A modified Cantor set, vaguely reminiscent of Exercise 1.15.1(ii), can work here.)

A useful fact is that Exercise 1.15.17 can be reversed:

Lemma 1.15.7 (Frostman's lemma)**.** *Let $d \geq 0$, and let $E \subset \mathbf{R}^n$ be a compact set with $\mathcal{H}^d(E) > 0$. Then there exists a non-trivial Frostman measure of dimension at least d supported on E (thus $\mu(E) > 0$ and $\mu(\mathbf{R}^d \backslash E) = 0$).*

Proof. Without loss of generality we may place the compact set E in the half-open unit cube $[0,1)^n$. It is convenient to work dyadically. For each integer $k \geq 0$, we subdivide $[0,1)^n$ into 2^{kn} half-open cubes $Q_{k,1}, \ldots, Q_{k,2^{nk}}$ of sidelength $\ell(Q_{k,i}) = 2^{-k}$ in the usual manner, and refer to such cubes as *dyadic cubes*. For each k and any $F \subset [0,1)^n$, we can define the *dyadic Hausdorff content* $h^\Delta_{d,k}(F)$ to be the quantity

$$h^\Delta_{d,2^{-k}}(F) := \inf\{\sum_j \ell(Q_{k_j,i_j})^d : Q_{k_j,i_j} \text{ cover } F; k_j \geq k\}$$

where the Q_{k_j,i_j} range over all at most countable families of dyadic cubes of sidelength at most 2^{-k} that cover F. By covering cubes by balls and vice versa, it is not hard to see that

$$c h_{d,C2^{-k}}(F) \leq h^\Delta_{d,2^{-k}}(F) \leq C h_{d,c2^{-k}}(F)$$

for some absolute constants c, C depending only on d, n. Thus, if we define the dyadic Hausdorff measure

$$(\mathcal{H}^d)^\Delta(F) := \lim_{k\to\infty} h^\Delta_{d,2^{-k}}(F)$$

then we see that the dyadic and non-dyadic Huausdorff measures are comparable:

$$c\mathcal{H}^d(F) \leq (\mathcal{H}^d)^\Delta(F) \leq C(\mathcal{H}^d)^\Delta(F).$$

In particular, the quantity $\sigma := (\mathcal{H}^d)^\Delta(E)$ is strictly positive.

Given any dyadic cube Q of length $\ell(Q) = 2^{-k}$, define the upper Frostman content $\mu^+(Q)$ to be the quantity

$$\mu^+(Q) := h^\Delta_{d,k}(E \cap Q).$$

Then $\mu^+([0,1)^n) \geq \sigma$. By covering $E \cap Q$ by Q, we also have the bound

$$\mu^+(Q) \leq \ell(Q)^d.$$

Finally, by the subadditivity property of Hausdorff content, if we decompose Q into 2^n cubes Q' of sidelength $\ell(Q') = 2^{-k-1}$, we have

$$\mu^+(Q) \leq \sum_{Q'} \mu^+(Q').$$

The quantity μ^+ behaves like a measure, but is subadditive rather than additive. Nevertheless, one can easily find another quantity $\mu(Q)$ to assign to each dyadic cube such that

$$\mu([0,1)^n) = \mu^+([0,1)^n)$$

and

$$\mu(Q) \leq \mu^+(Q)$$

for all dyadic cubes, and such that

$$\mu(Q) = \sum_{Q'} \mu(Q')$$

whenever a dyadic cube is decomposed into 2^n sub-cubes of half the sidelength. Indeed, such a μ can be constructed by a greedy algorithms starting at the largest cube $[0,1)^n$ and working downward; we omit the details. One can then use this "measure" μ to integrate any continuous compactly supported function on $\mathbf{R}^n$ (by approximating such a function by one which is constant on dyadic cubes of a certain scale), and so by the Riesz representation theorem, it extends to a Radon measure μ supported on $[0,1]^n$. (One could also have used the Caratheódory extension theorem at this point.) Since $\mu([0,1)^n) \geq \sigma$, μ is non-trivial; since $\mu(Q) \leq \mu^+(Q) \leq \ell(Q)^d$ for all dyadic cubes Q,

it is not hard to see that μ is a Frostman measure of dimension at most d, as desired. □

The study of Hausdorff dimension is then intimately tied to the study of the dimensional properties of various measures. We give some examples in the next few exercises.

Exercise 1.15.20. Let $0 < d \leq n$, and let $E \subset \mathbf{R}^n$ be a compact set. Show that $\dim_H(E) \geq d$ if and only if, for every $0 < \varepsilon < d$, there exists a compactly supported probability Borel measure μ with

$$\int_{\mathbf{R}^d} \int_{\mathbf{R}^d} \frac{1}{|x-y|^{d-\varepsilon}}\, d\mu(x) d\mu(y) < \infty.$$

Show that this condition is also equivalent to μ lying in the Sobolev space $H^{-(n-d+\varepsilon)/2}(\mathbf{R}^n)$. Thus we see a link here between Hausdorff dimension and Sobolev norms: the lower the dimension of a set, the rougher the measures that it can support, where the Sobolev scale is used to measure roughness.

Exercise 1.15.21. Let E be a compact subset of $\mathbf{R}^n$, and let μ be a Borel probability measure supported on E. Let $0 \leq d \leq n$.

- Suppose that for every $\varepsilon > 0$, every $0 < \delta < 1/10$, and every subset E' of E with $\mu(E') \geq \frac{1}{\log^2(1/\delta)}$, one could establish the bound $\mathcal{N}^*_\delta(E') \geq c_\varepsilon (\frac{1}{\delta})^{d-\varepsilon}$ for $*$ equal to any of $\mathrm{ext}, \mathrm{int}, \mathrm{net}, \mathrm{pack}$ (the exact choice of $*$ is irrelevant thanks to Exercise 1.15.4). Show that E has Hausdorff dimension at least d. (*Hint*: cover E by small balls, then round the radius of each ball to the nearest power of 2. Now use countable additivity and the observation that sum $\sum_\delta \frac{1}{\log^2(1/\delta)}$ is small when δ ranges over sufficiently small powers of 2.)
- Show that one can replace $\mu(E') \geq \frac{1}{\log^2(1/\delta)}$ with $\mu(E') \geq \frac{1}{\log\log^2(1/\delta)}$ in the previous statement. (*Hint*: instead of rounding the radius to the nearest power of 2, round instead to radii of the form $1/2^{2^{\varepsilon n}}$ for integers n.) This trick of using a hyper-dyadic range of scales rather than a dyadic range of scales is due to Bourgain[**Bo1999**]. The exponent 2 in the double logarithm can be replaced by any other exponent strictly greater than 1.

This should be compared with the task of lower bounding the lower Minkowski dimension, which only requires control on the entropy of E itself, rather than of large subsets E' of E. The results of this exercise are exploited to establish lower bounds on the Hausdorff dimension of Kakeya sets (and in particular, to conclude such bounds from the Kakeya maximal function conjecture).

Exercise 1.15.22. Let $E \subset \mathbf{R}^n$ be a Borel set, and let $\phi : E \to \mathbf{R}^m$ be a locally Lipschitz map. Show that $\dim_H(\phi(E)) \leq \dim_H(E)$, and that if E has zero d-dimensional Hausdorff measure then so does $\phi(E)$.

Exercise 1.15.23. Let $\phi : \mathbf{R}^n \to \mathbf{R}$ be a smooth function, and let $g : \mathbf{R}^n \to \mathbf{R}$ be a test function such that $|\nabla\phi| > 0$ on the support of g. Establish the *co-area formula*

$$\int_{\mathbf{R}^n} g(x)|\nabla\phi(x)|\ dx = \int_{\mathbf{R}} (\int_{\phi^{-1}(t)} g(x)\ d\mathcal{H}^{n-1}(x))\ dt. \tag{1.130}$$

(*Hint*: Subdivide the support of g to be small, and then apply a change of variables to make ϕ linear, e.g. $\phi(x) = x_1$.) This formula is in fact valid for all absolutely integrable g and Lipschitz ϕ, but is difficult to prove for this level of generality, requiring a version of *Sard's theorem*.

The coarea formula (1.130) can be used to link geometric inequalities to analytic ones. For instance, the sharp isoperimetric inequality

$$\mathrm{vol}^n(\Omega)^{\frac{n-1}{n}} \leq \frac{1}{n\omega_n^{1/n}} \mathcal{H}^{n-1}(\partial\Omega),$$

valid for bounded open sets Ω in $\mathbf{R}^n$, can be combined with the coarea formula (with $g := 1$) to give the sharp Sobolev inequality

$$\|\phi\|_{L^{\frac{n}{n-1}}(\mathbf{R}^n)} \leq \frac{1}{n\omega_n^{1/n}} \int_{\mathbf{R}^n} |\nabla\phi(x)|\ dx$$

for any test function ϕ, the main point being that $\phi^{-1}(t) \cup \phi^{-1}(-t)$ is the boundary of $\{|\phi| \geq t\}$ (one also needs to do some manipulations relating the volume of those level sets to $\|\phi\|_{L^{\frac{n}{n-1}}(\mathbf{R}^n)}$). We omit the details.

Notes. This lecture first appeared at `terrytao.wordpress.com/2009/05/19`. Thanks to Vicky for corrections.

Further discussion of Hausdorff dimension can be found in [**Fa2003**], [**Ma1995**], [**Wo2003**], as well as in many other places.

There was some interesting discussion online as to whether there could be an analogue of K-theory for Hausdorff dimension, although the results of the discussion were inconclusive.

Chapter 2

Related articles

2.1. An alternate approach to the Carathéodory extension theorem

In this section, I would like to give an alternate proof of a (weak form of the) Carathéodory extension theorem (Theorem 1.1.17). This argument is restricted to the σ-finite case, and does not extend the measure to quite as large a σ-algebra as is provided by the standard proof of this theorem, but I find it conceptually clearer (in particular, hewing quite closely to *Littlewood's principles*, and the general Lebesgue philosophy of treating sets of small measure as negligible), and suffices for many standard applications of this theorem, in particular the construction of Lebesgue measure.

Let us first state the precise statement of the theorem:

Theorem 2.1.1 (Weak Carathéodory extension theorem)**.** *Let $\mathcal{A}$ be a Boolean algebra of subsets of a set X, and let $\mu : \mathcal{A} \to [0, +\infty]$ be a function obeying the following three properties:*

(i) $\mu(\emptyset) = 0$.

(ii) *(Pre-countable additivity) If $A_1, A_2, \ldots \in \mathcal{A}$ are disjoint and such that $\bigcup_{n=1}^\infty A_n$ also lies in $\mathcal{A}$, then $\mu(\bigcup_{n=1}^\infty A_n) = \sum_{n=1}^\infty \mu(A_n)$.*

(iii) *(σ-finiteness) X can be covered by at most countably many sets in $\mathcal{A}$, each of which has finite μ-measure.*

Let $\mathcal{X}$ be the σ-algebra generated by $\mathcal{A}$. Then μ can be uniquely extended to a countably additive measure on $\mathcal{X}$.

We will refer to sets in $\mathcal{A}$ as elementary sets and sets in $\mathcal{X}$ as *measurable sets*. A typical example is when $X = [0, 1]$ and $\mathcal{A}$ is the collection of all sets that are unions of finitely many intervals; in this case, $\mathcal{X}$ are the Borel-measurable sets.

2.1.1. Some basics. Let us first observe that the hypotheses on the premeasure μ imply some other basic and useful properties:

From properties (i) and (ii) we see that μ is finitely additive (thus $\mu(A_1 \cup \ldots \cup A_n) = \mu(A_1) + \ldots + \mu(A_n)$ whenever $A_1, \ldots, A_n$ are disjoint elementary sets).

As particular consequences of finite additivity, we have monotonicity ($\mu(A) \leq \mu(B)$ whenever $A \subset B$ are elementary sets) and finite subadditivity ($\mu(A_1 \cup \ldots \cup A_n) \leq \mu(A_1) + \ldots + \mu(A_n)$ for all elementary $A_1, \ldots, A_n$, not necessarily disjoint).

We also have pre-countable subadditivity: $\mu(A) \leq \sum_{n=1}^{\infty} \mu(A_n)$ whenever the elementary sets $A_1, A_2, \ldots$ cover the elementary set A. To see this, first observe by replacing A_n with $A_n \backslash \bigcup_{i=1}^{n-1} A_i$ and using monotonicity that we may take the A_i to be disjoint; next, by restricting all the A_i to A and using monotonicity we may assume that A is the union of the A_i, and now the claim is immediate from pre-countable additivity.

2.1.2. Existence. Let us first verify existence. As is standard in measure-theoretic proofs for σ-finite spaces, we first handle the finite case (when $\mu(X) < \infty$), and then rely on countable additivity or sub-additivity to recover the σ-finite case.

The basic idea, following Littlewood's principles, is to view the measurable sets as lying in the "completion" of the elementary sets, or in other words to exploit the fact that measurable sets can be approximated to arbitrarily high accuracy by elementary sets.

Define the outer measure $\mu_*(A)$ of a set $A \subset X$ to be the infimum of $\sum_{n=1}^{\infty} \mu(A_n)$, where $A_1, A_2, \ldots$ range over all at most countable collections of elementary sets that cover A. It is clear that outer measure is monotone and countably subadditive. Also, since μ is pre-countably subadditive, we see that $\mu_*(A) \geq \mu(A)$ for all elementary A. Since we also have the trivial inequality $\mu_*(A) \leq \mu(A)$, we conclude that μ_* and μ agree on elementary sets.

The outer measure naturally defines a pseudometric[1] (and thus a topology) on the space of subsets of X, with the distance between A and B being defined as $\mu_*(A \Delta B)$, where Δ denotes symmetric difference. (The subadditivity of μ_* ensures the triangle inequality; furthermore, we see that the Boolean operations (union, intersection, complement, etc.) are all continuous with respect to this pseudometric.) With this pseudometric, we claim that the measurable sets lie in

[1] A pseudometric is a metric in which distinct objects are allowed to be separated by a zero distance.

the closure of the elementary sets. Indeed, it is not difficult to see (using subadditivity and monotonicity properties of μ_*) that the closure of the elementary sets are closed under finite unions, under complements, and under countable disjoint unions (here we need finiteness of $\mu(X)$ to keep the measure of all the pieces absolutely summable), and thus form a σ-algebra. Since this σ-algebra clearly contains the elementary sets, it must contain the measurable sets also.

By subadditivity of μ_*, the function $A \mapsto \mu_*(A)$ is Lipschitz continuous. Since this function is finitely additive on elementary sets, we see on taking limits (using subadditivity to control error terms) that it must be finitely additive on measurable sets also. Since μ_* is finitely additive, monotone, and countably sub-additive, it must be countably additive, and so μ_* is the desired extension of μ to the measurable sets. This completes the proof of the theorem in the finite measure case.

To handle the σ-finite case, we partition X into countably many elementary sets of finite measure, and use the above argument to extend μ to measurable subsets of each such elementary set. It is then a routine matter to sum together these localised measures to recover a measure on all measurable sets; the pre-countable additivity property ensures that this sum still agrees with μ on elementary sets.

2.1.3. Uniqueness. Now we verify uniqueness. Again, we begin with the finite measure case.

Suppose first that $\mu(X) < \infty$, and that we have two different extensions $\mu_1, \mu_2 : \mathcal{X} \to [0, +\infty]$ of μ to $\mathcal{X}$ that are countably additive. Observe that μ_1, μ_2 must both be continuous with respect to the μ_* pseudometric used in the existence argument, from countable subadditivity; since every measurable set is a limit of elementary sets in this pseudometric, we obtain uniqueness in the finite measure case.

When instead X is σ-finite, we cover X by elementary sets of finite measure. The previous argument shows that any two extensions μ_1, μ_2 of μ agree when restricted to each of these sets, and the claim then follows by countable additivity. This proves Theorem 2.1.1.

Remark 2.1.2. The uniqueness claim fails when the σ-finiteness condition is dropped. Consider for instance the rational numbers $X = \mathbf{Q}$,

and let the elementary sets be the finite unions of intervals $[a,b) \cap \mathbf{Q}$. Define the measure $\mu(A)$ of an elementary set to be zero if A is empty, and $+\infty$ otherwise. As the rationals are countable, we easily see that every set of rationals is measurable. One easily verifies the pre-countable additivity condition (though the σ-finiteness condition fails horribly). However, μ has multiple extensions to the measurable sets; for instance, any positive scalar multiple of counting measure is such an extension.

Remark 2.1.3. It is not difficult to show that the measure completion $\overline{\mathcal{X}}$ of $\mathcal{X}$ with respect to μ is the same as the topological closure of $\mathcal{X}$ (or of $\mathcal{A}$) with respect to the above pseudometric. Thus, for instance, a subset of $[0,1]$ is Lebesgue measurable if and only if it can be approximated to arbitrary accuracy (with respect to outer measure) by a finite union of intervals.

A particularly simple case of Theorem 2.1.1 occurs when X is a compact Hausdorff totally disconnected space (i.e. a *Stone space*), such as the infinite discrete cube $\{0,1\}^{\mathbf{N}}$ or any other Cantor space. Then (see forthcoming lecture notes) the Borel σ-algebra $\mathcal{X}$ is generated by the Boolean algebra $\mathcal{A}$ of clopen sets. Also, as clopen sets here are simultaneously compact and open, we see that any infinite cover of one clopen set by others automatically has a finite subcover. From this, we conclude

Corollary 2.1.4. *Let X be a compact Hausdorff totally disconnected space. Then any finitely additive σ-finite measure on the clopen sets uniquely extends to a countably additive measure on the Borel sets.*

By identifying $\{0,1\}^{\mathbf{N}}$ with $[0,1]$ up to a countable set, this provides one means to construct Lebesgue measure on $[0,1]$; similar constructions are available for $\mathbf{R}$ or $\mathbf{R}^n$.

Notes. This lecture first appeared at `terrytao.wordpress.com/2009/01/03`. Thanks to Américo Tavares, JB, Max Menzies, and mmailliw/william for corrections.

2.2. Amenability, the ping-pong lemma, and the Banach-Tarski paradox

Notational convention: In this section (and in Section 2.4) only, I will colour a statement red if it assumes the axiom of choice. (For the rest of this text, the axiom of choice will be implicitly assumed throughout.)

The famous *Banach-Tarski paradox* asserts that one can take the unit ball in three dimensions, divide it up into finitely many pieces, and then translate and rotate each piece so that their union is now two disjoint unit balls. As a consequence of this paradox, it is not possible to create a finitely additive measure on $\mathbf{R}^3$ that is both translation and rotation invariant, which can measure every subset of $\mathbf{R}^3$, and which gives the unit ball a non-zero measure. This paradox helps explain why Lebesgue measure (which is countably additive and both translation and rotation invariant, and gives the unit ball a non-zero measure) cannot measure every set, instead being restricted to measuring sets that are Lebesgue measurable.

On the other hand, it is not possible to replicate the Banach-Tarski paradox in one or two dimensions; the unit interval in $\mathbf{R}$ or unit disk in $\mathbf{R}^2$ cannot be rearranged into two unit intervals or two unit disks using only finitely many pieces, translations, and rotations, and indeed there do exist non-trivial finitely additive measures on these spaces. However, it is possible to obtain a Banach-Tarski type paradox in one or two dimensions using countably many such pieces; this rules out the possibility of extending Lebesgue measure to a countably additive translation invariant measure on all subsets of $\mathbf{R}$ (or any higher-dimensional space).

In this section we will establish all of the above results, and tie them in with some important concepts and tools in modern group theory, most notably amenability and the ping-pong lemma.

2.2.1. One-dimensional equidecomposability. Before we study the three-dimensional situation, let us first review the simpler one-dimensional situation. To avoid having to say "X can be cut up into finitely many pieces, which can then be moved around to create Y" all the time, let us make a convenient definition:

Definition 2.2.1 (Equidecomposability)**.** Let $G = (G, \cdot)$ be a group acting on a space X, and let A, B be subsets of X.

- We say that A, B are *finitely G-equidecomposable* if there exist finite partitions $A = \bigcup_{i=1}^{n} A_i$ and $B = \bigcup_{i=1}^{n} B_i$ and group elements $g_1, \ldots, g_n \in G$ such that $B_i = g_i A_i$ for all $1 \leq i \leq n$.
- We say that A, B are *countably G-equidecomposable* if there exist countable partitions $A = \bigcup_{i=1}^{\infty} A_i$ and $B = \bigcup_{i=1}^{\infty} B_i$ and group elements $g_1, g_2, \ldots \in G$ such that $B_i = g_i A_i$ for all i.
- We say that A is *finitely G-paradoxical* if it can be partitioned into two subsets, each of which is finitely G-equidecomposable with A.
- We say that A is *countably G-paradoxical* if it can be partitioned into two subsets, each of which is countably G-equidecomposable with A.

One can of course make similar definitions when $G = (G, +)$ is an additive group rather than a multiplicative one.

Clearly, finite G-equidecomposability implies countable G-equidecomposability, but the converse is not true. Observe that any finitely (resp. countably) additive and G-invariant measure on X that measures every single subset of X, must give either a zero measure or an infinite measure to a finitely (resp. countably) G-paradoxical set. Thus, paradoxical sets provide significant obstructions to constructing additive measures that can measure all sets.

Example 2.2.2. If $\mathbf{R}$ acts on itself by translation, then $[0, 2]$ is finitely $\mathbf{R}$-equidecomposable with $[10, 11) \cup [21, 22]$, and $\mathbf{R}$ is finitely $\mathbf{R}$-equidecomposable with $(-\infty, -10] \cup (10, +\infty)$.

Example 2.2.3. If G acts transitively on X, then any two finite subsets of X are finitely G-equidecomposable iff they have the same cardinality, and any two countably infinite sets of X are countably G-equidecomposable. In particular, any countably infinite subset of X is countably G-paradoxical.

Exercise 2.2.1. Show that finite G-equidecomposability and countable G-equidecomposability are both equivalence relations.

Exercise 2.2.2 (Banach-Schröder-Bernstein theorem)**.** Let G act on X, and let A, B be subsets of X.

(i) If A is finitely G-equidecomposable with a subset of B, and B is finitely G-equidecomposable with a subset of A, show that A and B are finitely G-equidecomposable with each other. (*Hint*: adapt the proof of the *Schröder-Bernstein theorem*, see Section 3.13.)

(ii) If A is finitely G-equidecomposable with a superset of B, and B is finitely G-equidecomposable with a superset of A, show that A and B are finitely G-equidecomposable with each other. (*Hint*: use part (i).)

Show that claims (i) and (ii) also hold when "finitely" is replaced by "countably".

Exercise 2.2.3. Show that if G acts on X, A is a subset of X which is finitely (resp. countably) G-paradoxical, and $x \in X$, then the recurrence set $\{g \in G : gx \in A\}$ is also finitely (resp. countably) G-paradoxical (where G acts on itself by translation).

Let us first establish countable equidecomposability paradoxes in the reals.

Proposition 2.2.4. *Let* $\mathbf{R}$ *act on itself by translations. Then* $[0,1]$ *and* $\mathbf{R}$ *are countably* $\mathbf{R}$*-equidecomposable.*

Proof. By Exercise 2.2.2, it will suffice to show that some set contained in $[0,1]$ is countably $\mathbf{R}$-equidecomposable with $\mathbf{R}$. Consider the space $\mathbf{R}/\mathbf{Q}$ of all cosets $x+\mathbf{Q}$ of the rationals. By the axiom of choice, we can express each such coset as $x+\mathbf{Q}$ for some $x \in [0,1/2]$, thus we can partition $\mathbf{R} = \bigcup_{x\in E} x+\mathbf{Q}$ for some $E \subset [0,1/2]$. By Example 2.2.3, $\mathbf{Q}\cap[0,1/2]$ is countably $\mathbf{Q}$-equidecomposable with $\mathbf{Q}$, which implies that $\bigcup_{x\in E} x+(\mathbf{Q}\cap[0,1/2])$ is countably $\mathbf{R}$-equidecomposable with $\bigcup_{x\in E} x+\mathbf{Q}$. Since latter set is $\mathbf{R}$ and the former set is contained in $[0,1]$, the claim follows. □

Of course, the same proposition holds if $[0,1]$ is replaced by any other interval. As a quick consequence of this proposition and Exercise 2.2.2, we see that any subset of $\mathbf{R}$ containing an interval is $\mathbf{R}$-equidecomposable with $\mathbf{R}$. In particular, we have

Corollary 2.2.5. *Any subset of $\mathbf{R}$ containing an interval is countably $\mathbf{R}$-paradoxical.*

In particular, we see that any countably additive translation-invariant measure that measures every subset of $\mathbf{R}$, must assign a zero or infinite measure to any set containing an interval. In particular, it is not possible to extend Lebesgue measure to measure all subsets of $\mathbf{R}$.

We now turn from countably paradoxical sets to finitely paradoxical sets. Here, the situation is quite different: we can rule out many sets from being finitely paradoxical. The simplest example is that of a finite set:

Proposition 2.2.6. *If G acts on X, and A is a non-empty finite subset of X, then A is not finitely (or countably) G-paradoxical.*

Proof. One easily sees that any two sets that are finitely or countably G-equidecomposable must have the same cardinality. The claim follows. □

Now we consider the integers.

Proposition 2.2.7. *Let the integers $\mathbf{Z}$ act on themselves by translation. Then $\mathbf{Z}$ is not finitely $\mathbf{Z}$-paradoxical.*

Proof. The integers are of course infinite, and so Proposition 2.2.6 does not apply directly. However, the key point is that the integers can be efficiently truncated to be finite, and so we will be able to adapt the argument used to prove Proposition 2.2.6 to this setting.

Let's see how. Suppose for contradiction that we could partition $\mathbf{Z}$ into two sets A and B, which are in turn partitioned into finitely many pieces $A = \bigcup_{i=1}^n A_i$ and $B = \bigcup_{j=1}^m B_j$, such that $\mathbf{Z}$ can be partitioned as $\mathbf{Z} = \bigcup_{i=1}^n A_i + a_i$ and $\mathbf{Z} = \bigcup_{j=1}^m B_j + b_j$ for some integers $a_1, \ldots, a_n, b_1, \ldots, b_m$.

Now let N be a large integer (much larger than $n, m, a_1, \ldots, a_n, b_1, \ldots, b_m$). We truncate $\mathbf{Z}$ to the interval $[-N, N] := \{-N, \ldots, N\}$. Clearly

$$A \cap [-N, N] = \bigcup_{i=1}^{n} A_i \cap [-N, N] \tag{2.1}$$

and

$$[-N, N] = \bigcup_{i=1}^{n} (A_i + a_i) \cap [-N, N]. \tag{2.2}$$

From (2.2) we see that the set $\bigcup_{i=1}^{n}(A_i \cap [-N, N]) + a_i$ differs from $[-N, N]$ by only $O(1)$ elements, where the bound in the $O(1)$ expression can depend on $n, a_1, \ldots, a_n$ but does not depend on N. (The point here is that $[-N, N]$ is "almost" translation-invariant in some sense.) Comparing this with (2.1) we see that

$$|[-N, N]| \leq |A \cap [-N, N]| + O(1). \tag{2.3}$$

Similarly with A replaced by B. Summing, we obtain

$$2|[-N, N]| \leq |[-N, N]| + O(1), \tag{2.4}$$

but this is absurd for N sufficiently large, and the claim follows. □

Exercise 2.2.4. Use the above argument to show that in fact no infinite subset of $\mathbf{Z}$ is finitely $\mathbf{Z}$-paradoxical; combining this with Example 2.2.3, we see that the only finitely $\mathbf{Z}$-paradoxical set of integers is the empty set.

The above argument can be generalised to an important class of groups:

Definition 2.2.8 (Amenability)**.** Let $G = (G, \cdot)$ be a discrete, at most countable, group. A *Følner sequence* is a sequence $F_1, F_2, F_3, \ldots$ of finite subsets of G with $\bigcup_{N=1}^{\infty} F_N = G$ with the property that $\lim_{N\to\infty} \frac{|gF_N \Delta F_N|}{|F_N|} = 0$ for all $g \in G$, where Δ denotes symmetric difference. A discrete, at most countable, group G is *amenable* if it contains at least one Følner sequence. Of course, one can define the same concept for additive groups $G = (G, +)$.

Remark 2.2.9. One can define amenability for uncountable groups by replacing the notion of a Følner sequence with a Følner net. Similarly, one can define amenability for locally compact Hausdorff groups

equipped with a Haar measure by using that measure in place of cardinality in the above definition. However, we will not need these more general notions of amenability here. The notion of amenability was first introduced (though not by this name, or by this definition) by von Neumann, precisely in order to study these sorts of decomposition paradoxes. We discuss amenability further in Section 2.8.

Example 2.2.10. The sequence $[-N, N]$ for $N = 1, 2, 3, \ldots$ is a Følner sequence for the integers $\mathbf{Z}$, which are hence an amenable group.

Exercise 2.2.5. Show that any abelian discrete group that is at most countable, is amenable.

Exercise 2.2.6. Show that any amenable discrete group G that is at most countable is not finitely G-paradoxical, when acting on itself. Combined with Exercise 2.2.3, we see that if such a group G acts on a non-empty space X, then X is not finitely G-paradoxical.

Remark 2.2.11. Exercise 2.2.6 suggests that an amenable group G hould be able to support a non-trivial finitely additive measure which is invariant under left-translations, and can measure all subsets of G. Indeed, one can even create a finitely additive probability measure, for instance by selecting a non-principal ultrafilter $p \in \beta\mathbf{N}$ and a Følner sequence $(F_n)_{n=1}^\infty$ and defining $\mu(A) := \lim_{n \to p} |A \cap F_n|/|F_n|$ for all $A \in G$.

The reals $\mathbf{R} = (\mathbf{R}, +)$ (which we will give the discrete topology!) are uncountable, and thus not amenable by the narrow definition of Definition 2.2.8. However, observe from Exercise 2.2.5 that any finitely generated subgroup of the reals is amenable (or equivalently, that the reals themselves with the discrete topology are amenable, using the Følner net generalisation of Definition 2.2.8. Also, we have the following easy observation:

Exercise 2.2.7. Let G act on X, and let A be a subset of X which is finitely G-paradoxical. Show that there exists a finitely generated subgroup H of G such that A is finitely H-paradoxical.

From this, we see that $\mathbf{R}$ is not finitely $\mathbf{R}$-paradoxical. But we can in fact say much more:

Proposition 2.2.12. *Let A be a non-empty subset of $\mathbf{R}$. Then A is not finitely $\mathbf{R}$-paradoxical.*

Proof. Suppose for contradiction that we can partition A into two sets $A = A_1 \cup A_2$ which are both finitely $\mathbf{R}$-equidecomposable with A. This gives us two maps $f_1 : A \to A_1$, $f_2 : A \to A_2$ which are piecewise given by a finite number of translations; thus there exists a finite set $g_1, \ldots, g_d \in \mathbf{R}$ such that $f_i(x) \in x + \{g_1, \ldots, g_d\}$ for all $x \in A$ and $i = 1, 2$.

For any integer $N \geq 1$, consider the 2^N composition maps $f_{i_1} \circ \ldots \circ f_{i_N} : A \to A$ for $i_1, \ldots, i_N \in \{1, 2\}$. From the disjointness of A_1, A_2 and an easy induction we see that the ranges of all these maps are disjoint, and so for any $x \in A$ the 2^N quantities $f_{i_1} \circ \ldots \circ f_{i_N}(x)$ are distinct. On the other hand, we have

$$f_{i_1} \circ \ldots \circ f_{i_N}(x) \in x + \{g_1, \ldots, g_d\} + \ldots + \{g_1, \ldots, g_d\}. \tag{2.5}$$

Simple combinatorics (relying primarily on the abelian nature of $(\mathbf{R}, +)$ shows that the number of values on the right-hand side of (2.5) is at most N^d. But for sufficiently large N, we have $2^N > N^d$, giving the desired contradiction. □

Let us call a group G *supramenable* if every non-empty subset of G is not finitely G-paradoxical; thus $\mathbf{R}$ is supramenable. From Exercise 2.2.3 we see that if a supramenable group acts on any space X, then the only finitely G-paradoxical subset of X is the empty set.

Exercise 2.2.8. We say that a group $G = (G, \cdot)$ has *subexponential growth* if for any finite subset S of G, we have $\lim_{n\to\infty} |S^n|^{1/n} = 1$, where $S^n = S \cdot \ldots \cdot S$ is the set of n-fold products of elements of S. Show that every group of subexponential growth is supramenable.

Exercise 2.2.9. Show that every abelian group has subexponential growth (and is thus supramenable). More generally, show that every nilpotent group has subexponential growth and is thus also supramenable.

Exercise 2.2.10. Show that if two finite unions of intervals in $\mathbf{R}$ are finitely $\mathbf{R}$-equidecomposable, then they must have the same total length. (*Hint*: reduce to the case when both sets consist of a single

interval. First show that the lengths of these intervals cannot differ by more than a factor of two, and then amplify this fact by iteration to conclude the result.)

Remark 2.2.13. We already saw that amenable groups G admit finitely additive translation-invariant probability measures that measure all subsets of G (Remarkk 2.2.11 can be extended to the uncountable case); in fact, this turns out to be an equivalent definition of amenability. It turns out that supramenable groups G enjoy a stronger property, namely that given any non-empty set A on G, there exists a finitely additive translation-invariant measure on G that assigns the measure 1 to A; this is basically a deep result of Tarski.

2.2.2. Two-dimensional equidecomposability. Now we turn to equidecomposability on the plane $\mathbf{R}^2$. The nature of equidecomposability depends on what group G of symmetries we wish to act on the plane.

Suppose first that we only allow ourselves to translate various sets in the planes, but not to rotate them; thus $G = \mathbf{R}^2$. As this group is abelian, it is supramenable by Exercise 2.2.9, and so any non-empty subset A of the plane will not be finitely $\mathbf{R}^2$-paradoxical; indeed, by Remark 2.2.13, there exists a finitely additive translation-invariant measure that gives A the measure 1. On the other hand, it is easy to adapt Corollary 2.2.5 to see that any subset of the plane containing a ball will be countably $\mathbf{R}^2$-paradoxical.

Now suppose we allow both translations and rotations, thus G is now the group $SO(2) \ltimes \mathbf{R}^2$ of (orientation-preserving) isometries $x \mapsto e^{i\theta}x + v$ for $v \in \mathbf{R}^2$ and $\theta \in \mathbf{R}/2\pi\mathbf{Z}$, where $e^{i\theta}$ denotes the anticlockwise rotation by θ around the origin. This group is no longer abelian, or even nilpotent, so Exercise 2.2.9 no longer applies. Indeed, it turns out that G is no longer supramenable. This is a consequence of the following three lemmas:

Lemma 2.2.14. *Let G be a group which contains a free semigroup on two generators (in other words, there exist group elements $g, h \in G$ such that all the words involving g and h (but not g^{-1} or h^{-1}) are distinct). Then G contains a non-empty finitely G-paradoxical set. In other words, G is not supramenable.*

Proof. Let S be the semigroup generated by g and h (i.e. the set of all words formed by g and h, including the empty word (i.e. group identity). Observe that gS and hS are disjoint subsets of S that are clearly G-equidecomposable with S. The claim then follows from Exercise 2.2.2. □

Lemma 2.2.15 (Semigroup ping-pong lemma)**.** *Let G act on a space X, let g, h be elements of G, and suppose that there exists a non-empty subset A of X such that gA and hA are disjoint subsets of A. Then g, h generate a free semigroup.*

Proof. As in the proof of Proposition 2.2.12, we see from induction that for two different words w, w' generated by g, h, the sets wA and $w'A$ are disjoint, and the claim follows. □

Lemma 2.2.16. . *The group $G = SO(2) \ltimes \mathbf{R}^2$ contains a free semigroup on two generators.*

Proof. It is convenient to identify $\mathbf{R}^2$ with the complex plane $\mathbf{C}$. We set g to be the rotation $gx := \omega x$ for some transcendental phase $\omega = e^{2\pi i\theta}$ be such that $\omega := e^{2\pi i\theta}$ is transcendental (such a phase must exist, since the set of algebraic complex numbers is countable), and let h be the translation $hx := x + 1$. Observe that g and h act on the set A of polynomials in ω with non-negative integer coefficients, and that gA and hA are disjoint. The claim now follows from Lemma 2.2.15. □

Combining Lemma 2.2.14 and Lemma 2.2.16 to create a countable, finitely paradoxical subset of $SO(2) \ltimes \mathbf{R}^2$, and then letting that set act on a generic point in the plane (noting that each group element in $SO(2) \ltimes \mathbf{R}^2$ has at most one fixed point), we obtain

Corollary 2.2.17 (Sierpinski-Mazurkiewicz paradox)**.** *There exist non-empty finitely $SO(2) \ltimes \mathbf{R}^2$-paradoxical subsets of the plane.*

We have seen that the group of rigid motions is not supramenable. Nevertheless, it is still amenable, thanks to the following lemma:

Lemma 2.2.18. *Suppose one has a short exact sequence $0 \to H \to G \to K \to 0$ of discrete, at most countable, groups, and suppose one*

has a choice function $\phi : K \to G$ that inverts the projection of G to K (the existence of which is automatic, from the axiom of choice, and also follows if G is finitely generated). If H and K are amenable, then so is G.

Proof. Let $(A_n)_{n=1}^\infty$ and $(B_n)_{n=1}^\infty$ be Følner sequences for H and K respectively. Let $f : \mathbf{N} \to \mathbf{N}$ be a rapidly growing function, and let $(F_n)_{n=1}^\infty$ be the set $F_n := \bigcup_{x \in B_n} \phi(x) \cdot A_{f(n)}$. One easily verifies that this is a Følner sequence for G if f is sufficiently rapidly growing. □

Exercise 2.2.11. Show that any finitely generated solvable group is amenable. More generally, show that any discrete, at most countable, solvable group is amenable.

Exercise 2.2.12. Show that any finitely generated subgroup of $SO(2) \ltimes \mathbf{R}^2$ is amenable. (*Hint*: use the short exact sequence $0 \to \mathbf{R}^2 \to SO(2) \ltimes \mathbf{R}^2 \to SO(2) \to 0$, which shows that $SO(2) \ltimes \mathbf{R}^2$ is solvable (in fact it is metabelian)). Conclude that $\mathbf{R}^2$ is not finitely $SO(2) \ltimes \mathbf{R}^2$-paradoxical.

Finally, we show a result of Banach.

Proposition 2.2.19. *The unit disk D in $\mathbf{R}^2$ is not finitely $SO(2) \ltimes \mathbf{R}^2$-paradoxical.*

Proof. If the claim failed, then D would be finitely $SO(2) \ltimes \mathbf{R}^2$-equidecomposable with a disjoint union of two copies of D, say D and $D + v$ for some vector v of length greater than 2. By Exercise 2.2.7, we can then find a subgroup G of $SO(2) \times \mathbf{R}^2$ generated by a finite number of rotations $x \mapsto e^{i\theta_j} x$ for $j = 1, \ldots, J$ and translations $x \mapsto x + v_k$ for $k = 1, \ldots, K$ such that D and $D \cup (D + v)$ are finitely G-equidecomposable. Indeed, we may assume that the rigid motions that move pieces of D to pieces of $D \cup (D + v)$ are of the form $x \mapsto e^{i\theta_j} x + v_k$ for some $1 \leq j \leq J, 1 \leq k \leq K$, thus

$$D \cup (D + v) = \bigcup_{j=1}^{J} \sum_{k=1}^{K} e^{i\theta_j} D_{j,k} + v_k \tag{2.6}$$

for some partition $D = \bigcup_{j=1}^{J} \sum_{k=1}^{K} D_{j,k}$ of the disk.

By amenability of the rotation group $SO(2)$, one can find a finite set $\Phi \subset SO(2)$ of rotations such that $e^{i\theta_j}\Phi$ differs from Φ by at most $0.01|\Phi|$ elements for all $1 \leq j \leq J$. Let N be a large integer, and let $\Gamma_N \subset \mathbf{R}^2$ be the set of all linear combinations of $e^{i\theta}v_k$ for $\theta \in \Phi$ and $1 \leq k \leq K$ with coefficients in $\{-N, \ldots, N\}$. Observe that Γ_N is a finite set whose cardinality grows at most polynomially in N. Thus, by the pigeonhole principle, one can find arbitrarily large N such that

$$|D \cap \Gamma_{N+10}| \leq 1.01|D \cap \Gamma_N|. \tag{2.7}$$

On the other hand, from (2.6) and the rotation-invariance of the disk we have

$$\begin{aligned} 2|D \cap \Gamma_N| &= 2|e^{i\theta}(D) \cap \Gamma_N| \\ &\leq |e^{i\theta}(D \cup (D+v)) \cap \Gamma_{N+5}| \\ &\leq \sum_{j=1}^{J}\sum_{k=1}^{K} |e^{i(\theta+\theta_j)}D_{j,k} \cap \Gamma_{N+10}| \end{aligned} \tag{2.8}$$

for all $\theta \in \Phi$. Averaging this over all $\theta \in \Phi$ we conclude

$$2|D \cap \Gamma_N| \leq 1.01|D \cap \Gamma_{N+10}|, \tag{2.9}$$

contradicting (2.7). □

Remark 2.2.20. Banach in fact showed the slightly stronger statement that any two finite unions of polygons of differing area were not finitely $SO(2) \times \mathbf{R}^2$-equidecomposable. (The converse is also true, and is known as the *Bolyai-Gerwien theorem*.)

Exercise 2.2.13. Show that all the claims in this section continue to hold if we replace $SO(2) \ltimes \mathbf{R}^2$ by the slightly larger group $\mathrm{Isom}(\mathbf{R})^2 = O(2) \ltimes \mathbf{R}^2$ of isometries (not necessarily orientation-preserving.

Remark 2.2.21. As a consequence of Remark 2.2.20, the unit square is not $SO(2) \times \mathbf{R}^2$-paradoxical. However, it is $SL(2) \times \mathbf{R}^2$-paradoxical; this is known as the *von Neumann paradox*.

2.2.3. Three-dimensional equidecomposability. We now turn to the three-dimensional setting. The new feature here is that the group $SO(3) \times \mathbf{R}^3$ of rigid motions is no longer abelian (as in one dimension) or solvable (as in two dimensions), but now contains a free

group on two generators (not just a free semigroup, as per Lemma 2.2.16. The significance of this fact comes from

Lemma 2.2.22. *The free group F_2 on two generators is finitely F_2-paradoxical.*

Proof. Let a, b be the two generators of F_2. We can partition $F_2 = \{1\} \cup W_a \cup W_b \cup W_{a^{-1}} \cup W_{b^{-1}}$, where W_c is the collection of reduced words of F_2 that begin with c. From the identities

$$W_{a^{-1}} = a^{-1} \cdot (F_2 \backslash W_a); \quad W_{b^{-1}} = b^{-1} \cdot (F_2 \backslash W_b) \tag{2.10}$$

we see that F_2 is finitely F_2-equidecomposable with both $W_a \cup W_{a^{-1}}$ and $W_c \cup W_{c^{-1}}$, and the claim now follows from Exercise 2.2.2. □

Corollary 2.2.23. *Suppose that F_2 acts freely on a space X (i.e. $gx \neq x$ whenever $x \in X$ and $g \in F_2$ is not the identity). Then X is finitely F_2-paradoxical.*

Proof. Using the axiom of choice, we can partition X as $X = \bigcup_{x \in \Gamma} F_2 x$ for some subset Γ of X. The claim now follows from Lemma 2.2.22. □

Next, we embed the free group inside the rotation group $SO(3)$ using the following useful lemma (cf. Lemma 2.2.15).

Exercise 2.2.14 (Ping-pong lemma)**.** Let G be a group acting on a set X. Suppose that there exist disjoint subsets A_+, A_-, B_+, B_- of X, whose union is not all of X, and elements $a, b \in G$, such that[2]

$$a(X \backslash A_-) \subset A_+; \quad a^{-1}(X \backslash A_+) \subset A_-; \quad b(X \backslash B_-) \subset B_+; \quad b^{-1}(X \backslash B_+) \subset B_-. \tag{2.11}$$

Show that a, b generate a free group.

Proposition 2.2.24. *$SO(3)$ contains a copy of the free group on two generators.*

Proof. It suffices to find a space X that two elements of $SO(3)$ act on in a way that Exercise 2.2.14 applies. There are many such constructions. One such construction[3], based on passing from the reals to the

[2] If drawn correctly, a diagram of the inclusions in (2.11) resembles a game of doubles ping-pong of A_+, A_- versus B_+, B_-, hence the name.

[3] See `http://sbseminar.wordpress.com/2007/09/17/` for more details and motivation for this construction.

5-adics, where -1 is a square root and so $SO(3)$ becomes isomorphic to $PSL(2)$. At the end of the day, one takes

$$a = \begin{pmatrix} 3/5 & 4/5 & 0 \\ -4/5 & 3/5 & 0 \\ 0 & 0 & 1 \end{pmatrix}; \quad b = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 3/5 & -4/5 \\ 0 & 4/5 & 3/5 \end{pmatrix} \tag{2.12}$$

and

(2.13)

$$A_\pm := 5^{\mathbf{Z}} \cdot \{ \begin{pmatrix} x \\ y \\ z \end{pmatrix} : x, y, z \in \mathbf{Z}, x = \pm 3y \bmod 5, z = 0 \bmod 5 \}$$

$$B_\pm := 5^{\mathbf{Z}} \cdot \{ \begin{pmatrix} x \\ y \\ z \end{pmatrix} : x, y, z \in \mathbf{Z}, z = \pm 3y \bmod 5, x = 0 \bmod 5 \}$$

$$X := A_- \cup A_+ \cup B_- \cup B_+ \cup \{ \begin{pmatrix} 0 & 1 & 0 \end{pmatrix} \},$$

where $5^{\mathbf{Z}}$ denotes the integer powers of 5 (which act on column vectors in the obvious manner). The verification of the ping-pong inclusions (2.11) is a routine application of modular arithmetic. □

Remark 2.2.25. This is a special case of the *Tits alternative*.

Corollary 2.2.26 (Hausdorff paradox)**.** *There exists a countable subset E of the sphere S^2 such that $S^2 \backslash E$ is finitely $SO(3)$-paradoxical, where $SO(3)$ of course acts on S^2 by rotations.*

Proof. Let $F_2 \subset SO(3)$ be a copy of the free group on two generators, as given by Proposition 2.2.24. Each rotation in F_2 fixes exactly two points on the sphere. Let E be the union of all these points; this is countable since F_2 is countable. The action of F_2 on $SO(3) \backslash E$ is free, and the claim now follows from Corollary 2.2.23. □

Corollary 2.2.27 (Banach-Tarski paradox on the sphere)**.** *S^2 is finitely $SO(3)$-paradoxical.*

Proof. (Sketch) Iterating the Hausdorff paradox, we see that $S^2 \backslash E$ is finitely $SO(3)$-equidecomposable to four copies of $S^2 \backslash E$, which can easily be used to cover two copies of S^2 (with some room to spare), by randomly rotating each of the copies. The claim now follows from Exercise 2.2.2. □

Exercise 2.2.15 (Banach-Tarski paradox on $\mathbf{R}^3$)**.** Show that the unit ball in $\mathbf{R}^3$ is finitely $SO(3) \ltimes \mathbf{R}^3$-paradoxical.

Exercise 2.2.16. Extend these three-dimensional paradoxes to higher dimensions.

Notes. This lecture first appeared at `terrytao.wordpress.com/2009/01/08`. Thanks to Harald Helfgott for corrections.

2.3. The Stone and Loomis-Sikorski representation theorems

A (concrete) *Boolean algebra* is a pair $(X, \mathcal{B})$, where X is a set, and $\mathcal{B}$ is a collection of subsets of X which contain the empty set $\emptyset$, and which is closed under unions $A, B \mapsto A \cup B$, intersections $A, B \mapsto A \cap B$, and complements $A \mapsto A^c := X \backslash A$. The subset relation $\subset$ also gives a relation on $\mathcal{B}$. Because the $\mathcal{B}$ is concretely represented as subsets of a space X, these relations automatically obey various axioms, in particular, for any $A, B, C \in \mathcal{B}$,

(i) $\subset$ is a partial ordering on $\mathcal{B}$, and A and B have join $A \cup B$ and meet $A \cap B$.

(ii) We have the distributive laws $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ and $A \cap (B \cup C) = A \cup (B \cap C)$.

(iii) $\emptyset$ is the minimal element of the partial ordering $\subset$, and $\emptyset^c$ is the maximal element.

(iv) $A \cap A^c = \emptyset$ and $A \cup A^c = \emptyset^c$.

(More succinctly: $\mathcal{B}$ is a lattice which is distributive, bounded, and complemented.)

We can then define an *abstract Boolean algebra* $\mathcal{B} = (\mathcal{B}, \emptyset, \cdot^c, \cup, \cap, \subset)$ to be an abstract set $\mathcal{B}$ with the specified objects, operations, and relations that obey the axioms (i)-(iv). Of course, some of these operations are redundant; for instance, intersection can be defined in terms of complement and union by de Morgan's laws. In the literature, different authors select different initial operations and axioms when defining an abstract Boolean algebra, but they are all easily seen to be equivalent to each other. To emphasise the abstract nature of

these algebras, the symbols $\emptyset, \cdot^c, \cup, \cap, \subset$ are often replaced with other symbols such as $0, \neg, \vee, \wedge, <$.

Clearly, every concrete Boolean algebra is an abstract Boolean algebra. In the converse direction, we have *Stone's representation theorem* (see below), which asserts (among other things) that every abstract Boolean algebra is isomorphic to a concrete one (and even constructs this concrete representation of the abstract Boolean algebra canonically). So, up to (abstract) isomorphism, there is really no difference between a concrete Boolean algebra and an abstract one.

Now let us turn from Boolean algebras to σ-algebras.

A concrete σ-algebra (also known as a *measurable space*) is a pair $(X, \mathcal{B})$, where X is a set, and $\mathcal{B}$ is a collection of subsets of X which contains $\emptyset$ and are closed under countable unions, countable intersections, and complements; thus every concrete σ-algebra is a concrete Boolean algebra, but not conversely. As before, concrete σ-algebras come equipped with the structures $\emptyset, \cdot^c, \cup, \cap, \subset$ which obey axioms (i)-(iv), but they also come with the operations of countable union $(A_n)_{n=1}^\infty \mapsto \bigcup_{n=1}^\infty A_n$ and countable intersection $(A_n)_{n=1}^\infty \mapsto \bigcap_{n=1}^\infty A_n$, which obey an additional axiom:

(v) Any countable family $A_1, A_2, \ldots$ of elements of $\mathcal{B}$ has supremum $\bigcup_{n=1}^\infty A_n$ and infimum $\bigcap_{n=1}^\infty A_n$.

As with Boolean algebras, one can now define an *abstract σ-algebra* to be a set $\mathcal{B} = (\mathcal{B}, \emptyset, \cdot^c, \cup, \cap, \subset, \bigcup_{n=1}^\infty, \bigcap_{n=1}^\infty)$ with the indicated objects, operations, and relations, which obeys axioms (i)-(v). Again, every concrete σ-algebra is an abstract one; but is it still true that every abstract σ-algebra is representable as a concrete one?

The answer turns out to be no, but the obstruction can be described precisely (namely, one needs to quotient out an ideal of "null sets" from the concrete σ-algebra), and there is a satisfactory representation theorem, namely the *Loomis-Sikorski representation theorem* (see below). As a corollary of this representation theorem, one can also represent abstract measure spaces $(\mathcal{B}, \mu)$ (also known as measure algebras) by concrete measure spaces, $(X, \mathcal{B}, \mu)$, after quotienting out by null sets.

In the rest of this section, I will state and prove these representation theorems. These theorems help explain why it is "safe" to focus attention primarily on concrete σ-algebras and measure spaces when doing measure theory, since the abstract analogues of these mathematical concepts are largely equivalent to their concrete counterparts. (The situation is quite different for non-commutative measure theories, such as *quantum probability*, in which there is basically no good representation theorem available to equate the abstract with the classically concrete, but I will not discuss these theories here.)

2.3.1. Stone's representation theorem. We first give the class of Boolean algebras the structure of a *category*:

Definition 2.3.1 (Boolean algebra morphism)**.** A *morphism* $\phi : \mathcal{A} \to \mathcal{B}$ from one abstract Boolean algebra to another is a map which preserves the empty set, complements, unions, intersections, and the subset relation (e.g. $\phi(A \cup B) = \phi(A) \cup \phi(B)$ for all $A, B \in \mathcal{A}$. An isomorphism is an morphism $\phi : \mathcal{A} \to \mathcal{B}$ which has an inverse morphism $\phi^{-1} : \mathcal{B} \to \mathcal{A}$. Two Boolean algebras are *isomorphic* if there is an isomorphism between them.

Note that if $(X, \mathcal{A}), (Y, \mathcal{B})$ are concrete Boolean algebras, and if $f : X \to Y$ is a map which is measurable in the sense that $f^{-1}(B) \in \mathcal{A}$ for all $B \in \mathcal{B}$, then the inverse of f is a Boolean algebra morphism $f^{-1} : \mathcal{B} \to \mathcal{A}$ which goes in the reverse (i.e. contravariant) direction to that of f. To state Stone's representation theorem we need another definition.

Definition 2.3.2 (Stone space)**.** A *Stone space* is a topological space $X = (X, \mathcal{F})$ which is compact, Hausdorff, and totally disconnected. Given a Stone space, define the *clopen algebra* $Cl(X)$ of X to be the concrete Boolean algebra on X consisting of the clopen sets (i.e. sets that are both closed and open).

It is easy to see that $Cl(X)$ is indeed a concrete Boolean algebra for any topological space X. The additional properties of being compact, Hausdorff, and totally disconnected are needed in order to recover the topology $\mathcal{F}$ of X uniquely from the clopen algebra. Indeed, we have

Lemma 2.3.3. *If X is a Stone space, then the topology $\mathcal{F}$ of X is generated by the clopen algebra $Cl(X)$. Equivalently, the clopen algebra forms an open base for the topology.*

Proof. Let $x \in X$ be a point, and let K be the intersection of all the clopen sets containing x. Clearly, K is closed. We claim that $K = \{x\}$. If this is not the case, then (since X is totally disconnected) K must be disconnected, thus K can be separated non-trivially into two closed sets $K = K_1 \cup K_2$. Since compact Hausdorff spaces are normal, we can write $K_1 = K \cap U_1$ and $K_2 = K \cap U_2$ for some disjoint open U_1, U_2. Since the intersection of all the clopen sets containing x with the closed set $(U_1 \cup U_2)^c$ is empty, we see from the finite intersection property that there must be a finite intersection K' of clopen sets containing x that is contained inside $U_1 \cup U_2$. In particular, $K' \cap U_1$ and $K' \cap U_2$ are clopen and do not contain K. But this contradicts the definition of K (since x is contained in one of $K' \cap U_1$ and $K' \cap U_2$). Thus $K = \{x\}$.

Another application of the finite intersection property then reveals that every open neighbourhood of x contains at least one clopen set containing x, and so the clopen sets form a base as required. □

Exercise 2.3.1. Show that two Stone spaces have isomorphic clopen algebras if and only if they are homeomorphic.

Now we turn to the representation theorem.

Theorem 2.3.4 (Stone representation theorem)**.** *Every abstract Boolean algebra $\mathcal{B}$ is equivalent to the clopen algebra $Cl(X)$ of a Stone space X.*

Proof. We will need the binary abstract Boolean algebra $\{0, 1\}$, with the usual Boolean logic operations. We define $X := \mathrm{Hom}(\mathcal{B}, \{0, 1\})$ be the space of all morphisms from $\mathcal{B}$ to {0,1}. Observe that each point $x \in X$ can be viewed as a finitely additive measure $\mu_x : \mathcal{B} \to \{0, 1\}$ that takes values in $\{0, 1\}$. In particular, this makes X a closed subset of $\{0, 1\}^{\mathcal{B}}$ (endowed with the product topology). The space $\{0, 1\}^{\mathcal{B}}$ is Hausdorff, totally disconnected, and (by Tychonoff's theorem, Theorem 1.8.14) compact, and so X is also; in other words, X is a Stone space. Every $B \in \mathcal{B}$ induces a cylinder set $C_B \subset \{0, 1\}^{\mathcal{B}}$, consisting of

all maps $\mu : \mathcal{B} \to \{0,1\}$ that map B to 1. If we define $\phi(B) := C_B \cap X$, it is not hard to see that ϕ is a morphism from $\mathcal{B}$ to $Cl(X)$. Since the cylinder sets are clopen and generate the topology of $\{0,1\}^{\mathcal{B}}$, we see that $\phi(\mathcal{B})$ of clopen sets generates the topology of X. Using compactness, we then conclude that every clopen set is the finite union of finite intersections of elements of $\phi(\mathcal{B})$; since $\phi(\mathcal{B})$ is an algebra, we thus see that ϕ is surjective.

The only remaining task is to check that ϕ is injective. It is sufficient to show that $\phi(A)$ is non-empty whenever $A \in \mathcal{B}$ is not equal to $\emptyset$. But by Zorn's lemma (Section 2.4), we can place A inside a maximal proper filter (i.e. an *ultrafilter*) p. The indictator $1_p : \mathcal{B} \to \{0,1\}$ of p can then be verified to be an element of $\phi(A)$, and the claim follows. □

Remark 2.3.5. If $\mathcal{B} = 2^Y$ is the power set of some set Y, then the Stone space given by Theorem 2.3.4 is the *Stone-Čech compactification* of Y (which we give the discrete topology); see Section 2.5.

Remark 2.3.6. Lemma 2.3.3 and Theorem 2.3.4 can be interpreted as giving a duality between the category of Boolean algebras and the category of Stone spaces, with the duality maps being $\mathcal{B} \mapsto \mathrm{Hom}(\mathcal{B}, \{0,1\})$ and $X \mapsto Cl(X)$. (The duality maps are (contravariant) functors which are inverses up to natural transformations.) It is the model example of the more general *Stone duality* between certain partially ordered sets and certain topological spaces. The idea of dualising a space X by considering the space of its morphisms to a fundamental space (in this case, $\{0,1\}$) is a common one in mathematics; for instance, *Pontryagin duality* in the context of Fourier analysis on locally compact abelian groups provides another example (with the fundamental space in this case being the unit circle $\mathbf{R}/\mathbf{Z}$); see Section 1.12. Other examples include the *Gelfand representation* of C^* algebras (here the fundamental space is the complex numbers $\mathbf{C}$; see Section 1.10.4) and the *ideal-variety correspondence* that provides the duality between algebraic geometry and commutative algebra (here the fundamental space is the base field k). In fact there are various connections between all of the dualities mentioned above.

Exercise 2.3.2. Show that any finite Boolean algebra is isomorphic to the power set of a finite set. (This is a special case of *Birkhoff's representation theorem.*)

2.3.2. The Loomis-Sikorski representation theorem. Now we turn to abstract σ-algebras. We can of course adapt Definition 2.3.1 to define the notion of a morphism or isomorphism between abstract σ-algebras, and to define when two abstract σ-algebras are isomorphic. Another important notion for us will be that of a quotient σ-algebra.

Definition 2.3.7 (Quotient σ-algebras)**.** Let $\mathcal{B}$ be an abstract σ-algebra. A *σ-ideal* in $\mathcal{B}$ is a subset $\mathcal{N}$ of $\mathcal{B}$ which contains $\emptyset$, is closed under countable unions, and is downwardly closed (thus if $N \in \mathcal{N}$ and $A \in \mathcal{B}$ is such that $A \subset N$, then $A \in \mathcal{N}$). If $\mathcal{N}$ is a σ-ideal, then we say that two elements of $\mathcal{B}$ are equivalent modulo $\mathcal{N}$ if their symmetric difference lies in $\mathcal{N}$. The quotient of $\mathcal{B}$ by this equivalence relation is denoted $\mathcal{B}/\mathcal{N}$, and can be given the structure of an abstract σ-algebra in a straightforward manner.

Example 2.3.8. If $(X, \mathcal{B}, \mu)$ is a measure space, then the collection $\mathcal{N}$ of sets of measure zero is a σ-ideal, so that we can form the abstract σ-algebra $\mathcal{B}/\mathcal{N}$. This freedom to quotient out the null sets is only available in the abstract setting, not the concrete one, and is perhaps the primary motivation for introducing abstract σ-algebras into measure theory in the first place.

One might hope that there is an analogue of Stone's representation theorem holds for σ-algebras. Unfortunately, this is not the case:

Proposition 2.3.9. *Let $\mathcal{B}$ be the Borel σ-algebra on $[0,1]$, and let $\mathcal{N}$ be the σ-ideal consisting of those sets with Lebesgue measure zero. Then the abstract σ-algebra $\mathcal{B}/\mathcal{N}$ is not isomorphic to a concrete σ-algebra.*

Proof. Suppose for contradiction that we had an isomorphism $\phi : \mathcal{B}/\mathcal{N} \to \mathcal{A}$ to some concrete σ-algebra $(X, \mathcal{A})$; this induces a map $\phi : \mathcal{B} \to \mathcal{A}$ which sends null sets to the empty set. Let x be a point in X. (It is clear that X must be non-empty.) Observe that any Borel set E in $[0,1]$ can be partitioned into two Borel subsets whose

Lebesgue measure is exactly half that of E. As a consequence, we see that if there exists a Borel set B such that $\phi(B)$ contains x, then there exists another Borel set B' of half the measure with $\phi(B')$ contains x. Iterating this (starting with $[0,1]$) we see that there exist Borel sets B of arbitrarily small measure with $\phi(B)$ containing x. Taking countable intersections, we conclude that there exists a null set N whose image $\phi(N)$ contains x; but $\phi(N)$ is empty, a contradiction. □

However, it turns out that quotienting out by ideals is the only obstruction to having a Stone-type representation theorem. Namely, we have

Theorem 2.3.10 (Loomis-Sikorski representation theorem)**.** *Let $\mathcal{B}$ be an abstract σ-algebra. Then there exists a concrete σ-algebra $(X, \mathcal{A})$ and a σ-ideal $\mathcal{N}$ of $\mathcal{A}$ such that $\mathcal{B}$ is isomorphic to $\mathcal{A}/\mathcal{N}$.*

Proof. We use the argument of Loomis[**Lo1946**]. Applying Stone's representation theorem, we can find a Stone space X such that there is a Boolean algebra isomorphism $\phi : \mathcal{B} \to Cl(X)$ from $\mathcal{B}$ (viewed now only as a Boolean algebra rather than a σ-algebra to the clopen algebra of X. Let $\mathcal{A}$ be the Baire σ-algebra of X, i.e. the σ-algebra generated by $Cl(X)$. The map ϕ need not be a σ-algebra isomorphism, being merely a Boolean algebra isomorphism one instead; it preserves finite unions and intersections, but need not preserve countable ones. In particular, if $B_1, B_2, \ldots \in \mathcal{B}$ are such that $\bigcap_{n=1}^\infty B_n = \emptyset$, then $\bigcap_{n=1}^\infty \phi(B_n) \in \mathcal{A}$ need not be empty.

Let us call sets $\bigcap_{n=1}^\infty \phi(B_n)$ of this form basic null sets, and let $\mathcal{N}$ be the collection of sets in $\mathcal{A}$ which can be covered by at most countably many basic null sets.

It is not hard to see that $\mathcal{N}$ is a σ-ideal in $\mathcal{A}$. The map ϕ then descends to a map $\phi : \mathcal{B} \to \mathcal{A}/\mathcal{N}$. It is not hard to see that ϕ is a Boolean algebra morphism. Also, if $B_1, B_2, \ldots \in \mathcal{B}$ are such that $\bigcap_{n=1}^\infty B_n = \emptyset$, then from construction we have $\bigcap_{n=1}^\infty \phi(B_n) = \emptyset$. From these two facts one can easily show that ϕ is in fact a σ-algebra morphism. Since $\phi(\mathcal{B}) = Cl(X)$ generates $\mathcal{A}$, $\phi(\mathcal{B})$ must generate $\mathcal{A}/\mathcal{N}$, and so ϕ is surjective.

The only remaining task is to show that ϕ is injective. As before, it suffices to show that $\phi(A) \neq \emptyset$ when $A \neq \emptyset$. Suppose for contradiction that $A \neq \emptyset$ and $\phi(A) = \emptyset$; then $\phi(A)$ can be covered by a countable family $\bigcap_{n=1}^{\infty} \phi(A_n^{(i)})$ of basic null sets, where $\bigcap_{n=1}^{\infty} A_n^{(i)} = \emptyset$ for each i. Since $A \neq \emptyset$ and $\bigcap_{n=1}^{\infty} A_n^{(1)} = \emptyset$, we can find n_1 such that $A \backslash A_{n_1}^{(1)} \neq \emptyset$ (where of course $A \backslash B := A \cap B^c$). Iterating this, we can find $n_2, n_3, n_4, \ldots$ such that $A \backslash (A_{n_1}^{(1)} \cup \ldots \cup A_{n_k}^{(k)}) \neq \emptyset$ for allk. Since ϕ is a Boolean space isomorphism, we conclude that $\phi(A)$ is not covered by any finite subcollection of the $\phi(A_{n_1}^{(1)}), \phi(A_{n_2}^{(2)}), \ldots$. But all of these sets are clopen, so by compactness, $\phi(A)$ is not covered by the entire collection $\phi(A_{n_1}^{(1)}), \phi(A_{n_2}^{(2)}), \ldots$. But this contradicts the fact that $\phi(A)$ is covered by the $\bigcap_{n=1}^{\infty} \phi(A_n^{(i)})$. □

,

Remark 2.3.11. The proof above actually gives a little bit more structure on $X, \mathcal{A}$, namely it gives X the structure of a Stone space, with $\mathcal{A}$ being its Baire σ-algebra. Furthermore, the ideal $\mathcal{N}$ constructed in the proof is in fact the ideal of meager Baire sets. The only difficult step is to show that every closed Baire set S with empty interior is in N, i.e. is a countable intersection of clopen sets. To see this, note that S is generated by a countable subalgebra of B which corresponds to a continuous map f from X to the Cantor set K (since K is dual to the free Boolean algebra on countably many generators). Then $f(S)$ is closed in K and is hence a countable intersection of clopen sets in K, which pull back to countably many clopen sets on X whose intersection is $f^{-1}(f(S))$. But the fact that S is generated by the subalgebra defining f can easily be seen to imply that $f^{-1}(f(S)) = S$.

Remark 2.3.12. The Stone representation theorem relies in an essential way on the axiom of choice (or at least the *boolean prime ideal theorem*, which is slightly weaker than this axiom). However, it is possible to prove the Loomis-Sikorski representation theorem without choice; see for instance [**BudePvaR2008**].

Remark 2.3.13. The construction of $X, \mathcal{A}, \mathcal{N}$ in the above proof was canonical, but it is not unique (in contrast to the situation with the

Stone representation theorem, where Lemma 2.3.3 provides uniqueness up to homeomorphisms). Nevertheless, using Remark 2.3.11, one can make the Loomis-Sikorski representation functorial. Let A and B be σ-algebras with Stone spaces X and Y. A map $Y \to X$ induces a σ-homomorphism $\mathrm{Bor}(X) \to \mathrm{Bor}(Y)$, and if the inverse image of a Borel meager set is meager then it induces a σ-homomorphism $A \to B$. Conversely a σ-homomorphism $A \to B$ induces a map $Y \to X$ under which the inverse image of a Borel meager set is meager (using the fact above that Borel meager sets are generated by countable intersections of clopen sets). The correspondence is bijective since it is just a restriction of the correspondence for ordinary Boolean algebras. This gives a duality between the category of σ-algebras and σ-homomorphisms and the category of "σ-Stone spaces" and continuous maps such that the inverse image of a Borel meager set is meager. In fact, "σ-Stone spaces" can be abstractly characterized as Stone spaces such that the closure of a countable union of clopen sets is clopen.

A (concrete) measure space $(X, \mathcal{B}, \mu)$ is a concrete σ-algebra $(X, \mathcal{B})$ together with a countably additive measure $\mu : \mathcal{B} \to [0, +\infty]$. One can similarly define an abstract measure space $(\mathcal{B}, \mu)$ (or measure algebra) to be an abstract σ-algebra $\mathcal{B}$ with a countably additive measure $mu : \mathcal{B} \to [0, +\infty]$. (Note that one does not need the concrete space X in order to define the notion of a countably additive measure.)

One can obtain an abstract measure space from a concrete one by deleting X and then quotienting out by some σ-ideal of null sets - sets of measure zero with respect to mu. (For instance, one could quotient out the space of all null sets, which is automatically a σ-ideal.) Thanks to the Loomis-Sikorski representation theorem, we have a converse:

Exercise 2.3.3. Show that every abstract measure space is isomorphic to a concrete measure space after quotieting out by a σ-ideal of null sets (where the notion of morphism, isomorphism, etc. on abstract measure spaces is defined in the obvious manner.)

Notes. This lecture first appeared at `terrytao.wordpress.com/2009/01/12`. Thanks to Eric for Remark 2.3.11, and for the functoriality remark in Remark 2.3.13.

Eric and Tom Leinster pointed out a subtlety that two concrete Boolean algebras which are abstractly isomorphic need not be concretely isomorphic. In particular, the modifier "abstract" is essential in the statement that "up to (abstract) isomorphism, there is no difference between a concrete Boolean algebra and an abstract one".

2.4. Well-ordered sets, ordinals, and Zorn's lemma

Notational convention: As in Section 2.2, I will colour a statement red in this post if it assumes the axiom of choice. We will, of course, rely on every other axiom of Zermelo-Frankel set theory here (and in the rest of the course).

In analysis, one often needs to iterate some sort of operation "infinitely many times" (e.g. to create a infinite basis by choosing one basis element at a time). In order to do this rigorously, we will rely on *Zorn's lemma*:

Lemma 2.4.1 (Zorn's Lemma)**.** *Let $(X, \leq)$ be a non-empty partially ordered set, with the property that every chain (i.e. a totally ordered set) in X has an upper bound. Then X contains a maximal element (i.e. an element with no larger element).*

Indeed, we have used this lemma several times already in previous sections. Given the other standard axioms of set theory, this lemma is logically equivalent to

Axiom 2.4.2 (Axiom of choice)**.** *Let X be a set, and let $\mathcal{F}$ be a collection of non-empty subsets of X. Then there exists a choice function $f : \mathcal{F} \to X$, i.e. a function such that $f(A) \in A$ for all $A \in \mathcal{F}$.*

One implication is easy:

Proof of axiom of choice using Zorn's lemma. Define a partial choice function to be a pair $(\mathcal{F}', f')$, where $\mathcal{F}'$ is a subset of $\mathcal{F}$ and $f' : \mathcal{F}' \to X$ is a choice function for $\mathcal{F}'$. We can partially order the

collection of partial choice functions by writing $(\mathcal{F}', f') \leq (\mathcal{F}'', f'')$ if $\mathcal{F}' \subset \mathcal{F}''$ and f'' extends f'. The collection of partial choice functions is non-empty (since it contains the pair $(\emptyset, ())$ consisting of the empty set and the empty function), and it is easy to see that any chain of partial choice functions has an upper bound (formed by gluing all the partial choices together). Hence, by Zorn's lemma, there is a maximal partial choice function $(\mathcal{F}_*, f_*)$. But the domain $\mathcal{F}_*$ of this function must be all of $\mathcal{F}$, since otherwise one could enlarge $\mathcal{F}_*$ by a single set A and extend f_* to A by choosing a single element of A. (One does not need the axiom of choice to make a single choice, or finitely many choices; it is only when making infinitely many choices that the axiom becomes necessary.) The claim follows. □

In the rest of this section I would like to supply the reverse implication, using the machinery of well-ordered sets. Instead of giving the shortest or slickest proof of Zorn's lemma here, I would like to take the opportunity to place the lemma in the context of several related topics, such as *ordinals* and *transfinite induction*, noting that much of this material is in fact independent of the axiom of choice. The material here is standard, but for the purposes of real analysis, one may simply take Zorn's lemma as a "black box" and not worry about the proof.

2.4.1. Well-ordered sets. To prove Zorn's lemma, we first need to strengthen the notion of a totally ordered set.

Definition 2.4.3. A *well-ordered set* is a totally ordered set $X = (X, \leq)$ such that every non-empty subset A of X has a minimal element $\min(A) \in A$. Two well-ordered sets X, Y are *isomorphic* if there is an order isomorphism $\phi : X \to Y$ between them, i.e. a bijection ϕ which is monotone ($\phi(x) < \phi(x')$ whenever $x < x'$).

Example 2.4.4. The natural numbers are well-ordered (this is the *well-ordering principle*), as is any finite totally ordered set (including the empty set), but the integers, rationals, or reals are not well-ordered.

Example 2.4.5. Any subset of a well-ordered set is again well-ordered. In particular, if a, b are two elements of a well-ordered

set, then *intervals* such as $[a,b] := \{c \in X : a \leq c \leq b\}$, $[a,b) := \{c \in X : a \leq c < b\}$, etc. are also well-ordered.

Example 2.4.6. If X is a well-ordered set, then the ordered set $X \oplus \{+\infty\}$, defined by adjoining a new element $+\infty$ to X and declaring it to be larger than all the elements of X, is also well-ordered. More generally, if X and Y are well-ordered sets, then the ordered set $X \oplus Y$, defined as the *disjoint union* of X and Y, with any element of Y declared to be larger than any element of X, is also well-ordered. Observe that the operation $\oplus$ is associative (up to isomorphism), but not commutative in general: for instance, $\mathbf{N} \oplus \{\infty\}$ is not isomorphic to $\{\infty\} \oplus \mathbf{N}$.

Example 2.4.7. If X, Y are well-ordered sets, then the ordered set $X \otimes Y$, defined as the Cartesian product $X \times Y$ with the lexicographical ordering (thus $(x,y) \leq (x',y')$ if $x < x'$, or if $x = x'$ and $y \leq y'$), is again a well-ordered set. Again, this operation is associative (up to isomorphism) but not commutative. Note that we have one-sided distributivity: $(X \oplus Y) \otimes Z$ is isomorphic to $(X \otimes Z) \oplus (Y \otimes Z)$, but $Z \otimes (X \oplus Y)$ is not isomorphic to $(Z \otimes X) \oplus (Z \otimes Y)$ in general.

Remark 2.4.8. The axiom of choice is trivially true in the case when X is well-ordered, since one can take min to be the choice function. Thus, the axiom of choice follows from the well-ordering theorem (every set has at least one well-ordering). Conversely, we will be able to deduce the well-ordering theorem from Zorn's lemma (and hence from the axiom of choice): see Exercise 2.4.11 below.

One of the reasons that well-ordered sets are useful is that one can perform induction on them. This is easiest to describe for the principle of strong induction:

Exercise 2.4.1 (Strong induction on well-ordered sets)**.** Let X be a well-ordered set, and let $P : X \mapsto \{\text{true}, \text{false}\}$ be a property of elements of X. Suppose that whenever $x \in X$ is such that $P(y)$ is true for all $y < x$, then $P(x)$ is true. Then $P(x)$ is true for every $x \in X$. This is called the *principle of strong induction*. Conversely, show that a totally ordered set X enjoys the principle of strong induction if and only if it is well-ordered. (For partially ordered sets, the corresponding notion is that of being *well-founded*.)

To describe the analogue of the ordinary principle of induction for well-ordered sets, we need some more notation. Given a subset A of a non-empty well-ordered set X, we define the *supremum* $\sup(A) \in X \oplus \{+\infty\}$ of A to be the least upper bound

$$\sup(A) := \min(\{y \in X \oplus \{+\infty\} : x \leq y \text{ for all } x \in X\} \tag{2.14}$$

of A (thus for instance the supremum of the empty set is $\min(X)$). If $x \in X$, we define the *successor* $\mathrm{succ}(x) \in X \oplus \{+\infty\}$ of x by the formula

$$\mathrm{succ}(x) := \min((x, +\infty]). \tag{2.15}$$

We have the following Peano-type axioms:

Exercise 2.4.2. If x is an element of a non-empty well-ordered set X, show that exactly one of the following statements hold:

- (Limit case) $x = \sup([\min(X), x))$.
- (Successor case) $x = \mathrm{succ}(y)$ for some Y.

In particular, $\min(X)$ is not the successor of any element in X.

Exercise 2.4.3. Show that if x, y are elements of a well-ordered set such that $\mathrm{succ}(x) = \mathrm{succ}(y)$, then $x = y$.

Exercise 2.4.4 (Transfinite induction for well-ordered sets)**.** Let X be a non-empty well-ordered set, and let $P : X \mapsto \{\mathrm{true}, \mathrm{false}\}$ be a property of elements of X. Suppose that

- (Base case) $P(\min(X))$ is true.
- (Successor case) If $x \in X$ and $P(x)$ is true, then $P(\mathrm{succ}(x))$ is true.
- (Limit case) If $x = \sup([\min(X), x))$ and $P(y)$ is true for all $y < x$, then $P(x)$ is true. [Note that this subsumes the base case.]

Then $P(x)$ is true for all $x \in X$.

Remark 2.4.9. The usual Peano axioms for succession are the special case of Exercises 2.4.2-2.4.4 in which the limit case of Exercise 2.4.2 only occurs for $\min(X)$ (which is denoted 0), and the successor function never attains $+\infty$. With these additional axioms, X is necessarily isomorphic to $\mathbf{N}$.

Now we introduce two more key concepts.

Definition 2.4.10. An *initial segment* of a well-ordered set X is a subset Y of X such that $[\min(X), y] \subset Y$ for all $y \in Y$ (i.e. whenever y lies in Y, all elements of X that are less than y also lie in Y).

A *morphism* from one well-ordered set X to another Y is a map $\phi : X \to Y$ which is strictly monotone (thus $\phi(x) < \phi(x')$ whenever $x < x'$) and such that $\phi(X)$ is an initial segment of Y.

Example 2.4.11. The only morphism from $\{1, 2, 3\}$ to $\{1, 2, 3, 4, 5\}$ is the inclusion map. There is no morphism from $\{1, 2, 3, 4, 5\}$ to $\{1, 2, 3\}$.

Remark 2.4.12. With this notion of a morphism, the class of well-ordered sets becomes a *category*.

We can identify the initial segments of X with elements of $X \cup \{+\infty\}$:

Exercise 2.4.5. Let X be a non-empty well-ordered set. Show that every initial segment I of X is of the form $I = [\min(X), a)$ for exactly one $a \in X \cup \{+\infty\}$.

Exercise 2.4.6. Show that an arbitrary union or arbitrary intersection of initial segments is again an initial segment.

Exercise 2.4.7. Let $\phi : X \to Y$ be a morphism. Show that ϕ maps initial segments of X to initial segments of Y. If $x, x' \in X$ is such that x' is the successor of x, show that $\phi(x')$ is the successor of $\phi(x)$.

As Example 2.4.11 suggests, there are very few morphisms between well-ordered sets. Indeed, we have

Proposition 2.4.13 (Uniqueness of morphisms)**.** *Given two well-ordered sets X and Y, there is at most one morphism from X and Y.*

Proof. Suppose we have two morphisms $\phi : X \to Y$, $\psi : X \to Y$. By using transfinite induction (Exercise 2.4.4 and Exercise 2.4.7), we see that ϕ, ψ agree on $[\min(X), a)$ for every $a \in X \oplus \{+\infty\}$; setting $a = +\infty$ gives the claim. □

Exercise 2.4.8 (Schroder-Bernstein theorem for well-ordered sets)**.** Show that two well-ordered sets X, Y are isomorphic if and only if there is a morphism from X to Y, and a morphism from Y to X.

We can complement the uniqueness in Proposition 2.4.13 with existence:

Proposition 2.4.14 (Existence of morphisms)**.** *Given two well-ordered sets X and Y, there is either a morphism from X to Y or a morphism from Y to X.*

Proof. Call an element $a \in X \oplus \{+\infty\}$ good if there is a morphism ϕ_a from $[\min(X), a)$ to Y, thus $\min(X)$ is good. If $+\infty$ is good, then we are done. From uniqueness we see that if every element in a set A is good, then the supremum $\sup(A)$ is also good. Applying transfinite induction (Exercise 2.4.5), we thus see that we are done unless there exists a good $a \in X$ such that $\mathrm{succ}(a)$ is not good. By Exercise 2.4.5, $\phi_a([\min(X), a)) = [\min(Y), b)$ for some $b \in Y \oplus \{+\infty\}$. If $b \in Y$ then we could extend the morphism ϕ_a to $[\min(X), a] = [\min(X), \mathrm{succ}(a))$ by mapping a to b, contradicting the fact that $\mathrm{succ}(a)$ is not good; thus $b = +\infty$ and so ϕ_a is surjective. It is then easy to check that ϕ_a^{-1} exists and is a morphism from Y to X, and the claim follows. □

Remark 2.4.15. Formally, Proposition 2.4.13, Exercise 2.4.8, and Proposition 2.4.14 tell us that the collection of all well-ordered sets, modulo isomorphism, is totally ordered by declaring one well-ordered set X to be at least as large as another Y when there is a morphism from Y to X. However, this is not quite the case, because the collection of well-ordered sets is only a class rather than a set. Indeed, as we shall soon see, this is not a technicality, but is in fact a fundamental fact about well-ordered sets that lies at the heart of Zorn's lemma. (From *Russell's paradox* we know that the notions of class and set are necessarily distinct; see Section 3.15.)

2.4.2. Ordinals. As we learn very early on in our mathematics education, a finite set of a certain cardinality (e.g. a set $\{a, b, c, d, e\}$) can be put in one-to-one correspondence with a "standard" set of the same cardinality (e.g. the set $\{1, 2, 3, 4, 5\}$); two finite sets have the same cardinality if and only if they correspond to the same "standard"

set $\{1, \ldots, N\}$). (The same fact is true for infinite sets; see Exercise 2.4.12 below.) Similarly, we would like to place every well-ordered set in a "standard" form. This motivates

Definition 2.4.16. A *representation* ρ of the well-ordered sets is an assignment of a well-ordered set $\rho(X)$ to every well-ordered set X such that

- $\rho(X)$ is isomorphic to X for every well-ordered set X. (In particular, if $\rho(X)$ and $\rho(Y)$ are equal, then X and Y are isomorphic.)
- If there exists a morphism from X to Y, then $\rho(X)$ is a subset of $\rho(Y)$ (and the order structure on $\rho(X)$ is induced from that on $\rho(Y)$. (In particular, if X and Y are isomorphic, then $\rho(X)$ and $\rho(Y)$ are equal.)

Remark 2.4.17. In the language of category theory, a representation is a covariant functor from the category of well-ordered sets to itself which turns all morphisms into inclusions, and which is naturally isomorphic to the identity functor.

Remark 2.4.18. Because the collection of all well-ordered sets is a class rather than a set, ρ is not actually a function (it is sometimes referred to as a *class function*).

It turns out that several representations of the well-ordered sets exist. The most commonly used one is that of the *ordinals*, defined by von Neumann as follows.

Definition 2.4.19 (Ordinals)**.** An *ordinal* is a well-ordered set α with the property that $x = \{y \in \alpha : y < x\}$ for all $x \in \alpha$. (In particular, each element of α is also a subset of α, and the strict order relation $<$ on α is identical to the set membership relation $\in$.)

Example 2.4.20. For each natural number $n = 0, 1, 2, \ldots$, define the ordinal number n^{th} recursively by setting $0^{\text{th}} := \emptyset$ and $n^{\text{th}} :=$

$\{0^{\text{th}}, 1^{\text{th}}, \ldots, (n-1)^{\text{th}}\}$ for all $n \geq 1$, thus for instance

$$\begin{aligned} 0^{\text{th}} &:= \emptyset \\ 1^{\text{th}} &:= \{0^{\text{th}}\} = \{\emptyset\} \\ 2^{\text{th}} &:= \{0^{\text{th}}, 1^{\text{th}}\} = \{\emptyset, \{\emptyset\}\} \\ 3^{\text{th}} &:= \{0^{\text{th}}, 1^{\text{th}}, 2^{\text{th}}\} = \{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}\}, \end{aligned} \tag{2.16}$$

and so forth. (Of course, to be compatible with the English language conventions for ordinals, we should write 1^{st} instead of 1^{th}, etc., but let us ignore this discrepancy.) One can easily check by induction that n^{th} is an ordinal for every n. Furthermore, if we define $\omega := \{n^{\text{th}} : n \in \mathbf{N}\}$, then ω is also an ordinal. (In the foundations of set theory, this construction, together with the axiom of infinity, is sometimes used to define the natural numbers (so that $n = n^{\text{th}}$ for all natural numbers n), although this construction can lead to some conceptually strange-looking consequences that blur the distinction between numbers and sets, such as $3 \in 5$ and $4 = \{0, 1, 2, 3\}$.)

The fundamental theorem about ordinals is

Theorem 2.4.21. (i) *Given any two ordinals α, β, one is a subset of the other (and the order structure on α is induced from that on β).*

(ii) *Every well-ordered set X is isomorphic to exactly one ordinal* $\text{ord}(X)$.

In particular, ord *is a representation of the well-ordered sets.*

Proof. We first prove (i). From Proposition 2.4.14 and symmetry, we may assume that there is a morphism ϕ from α to β. By strong induction (Exercise 2.4.1) and Definition 2.4.19, we see that $\phi(x) = x$ for all $x \in \alpha$, and so ϕ is the inclusion map from α into β. The claim follows.

Now we prove (ii). If uniqueness failed, then we would have two distinct ordinals that are isomorphic to each other, but as one ordinal is a subset of the other, this would contradict Proposition 2.4.13 (the inclusion morphism is not an isomorphism); so it suffices to prove existence.

We use transfinite induction. It suffices to show that for every $a \in X \oplus \{+\infty\}$, that $[\min(X), a)$ is isomorphic to an ordinal $\alpha(a)$ (which we know to be unique). This is of course true in the base case $a = \min(X)$. To handle the successor case $a = \mathrm{succ}(b)$, we set $\alpha(a) := \alpha(b) \cup \{\alpha(b)\}$, which is easily verified to be an ordinal isomorphic to $[\min(X).a)$. To handle the limit case $a = \sup([\min(X), a))$, we take all the ordinals associated to elements in $[\min(X), a)$ and take their union (here we rely crucially on the axiom schema of replacement and the axiom of union); by use of (i) one can show that this union is an ordinal isomorphic to a as required. □

Remark 2.4.22. Operations on well-ordered sets, such as the sum $\oplus$ and product $\otimes$ defined in Exercises 2.4.3, 2.4.4, induce corresponding operations on ordinals, leading to ordinal arithmetic, which we will not discuss here. (Note that the convention for which order multiplication proceeds in is swapped in some of the literature, thus $\alpha\beta$ would be the ordinal of $\beta \otimes \alpha$ rather than $\alpha \otimes \beta$.)

Exercise 2.4.9 (Ordinals are themselves well-ordered)**.** Let $\mathcal{F}$ be a non-empty class of ordinals. Show that there is a least ordinal $\min(\mathcal{F})$ in this class, which is a subset of all the other ordinals in this class. In particular, this shows that any set of ordinals is well-ordered by set inclusion.

Remark 2.4.23. Because of Exercise 2.4.9, we can meaningfully talk about "the least ordinal obeying property P", as soon as we can exhibit at least one ordinal with that property P. For instance, once one can demonstrate the existence of an uncountable ordinal (which follows from Exercise 2.4.11 below[4]), one can talk about the least uncountable ordinal.

Exercise 2.4.10 (Transfinite induction for ordinals)**.** Let $P(\alpha)$ be a property pertaining to ordinals α. Suppose that

- (Base case) $P(\emptyset)$ is true.
- (Successor case) If $\alpha = \{\beta, \{\beta\}\}$ for some ordinal β, and $P(\beta)$ is true, then $P(\alpha)$ is true.

[4]One can also create an uncountable ordinal without the axiom of choice by taking starting with all the well-orderings of subsets of the natural numbers, and taking the union of their associated ordinals; this construction is due to Hartogs.

- (Limit case) If $\alpha = \bigcup_{\beta \in \alpha} \beta$, and $P(\beta)$ is true for all $\beta \in \alpha$, then $P(\alpha)$ is true.

Show that $P(\alpha)$ is true for every ordinal α.

Now we show a fundamental fact, that the well-ordered sets are just too "numerous" to all fit inside a single set, even modulo isomorphism.

Theorem 2.4.24. *There does not exist a set A and a representation ρ of the well-ordered sets such that $\rho(X) \in A$ for all well-ordered sets X.*

Proof. By Theorem 2.4.21, any two distinct ordinals are non-isomorphic, and so get mapped under ρ to a different element of A. Thus we can identify the class of ordinals with a subset of A, and so the class of ordinals is in fact a set. In particular, by the axiom of union, we may take the union of all the ordinals, which one can verify to be another ordinal ε_0. But then $\varepsilon_0 \cup \{\varepsilon_0\}$ is another ordinal, which implies that $\varepsilon_0 \in \varepsilon_0$, which contradicts the *axiom of foundation*. □

Remark 2.4.25. It is also possible to prove Theorem 2.4.24 without the theory of ordinals, or the axiom of foundation. One first observes (by transfinite induction) that given two well-ordered sets X, X', one of the sets $\rho(X), \rho(X')$ is a subset of the other. Because of this, one can show that the union S of all the $\rho(X)$ (where X ranges over all well-ordered sets) is well-defined (because the $\rho(X)$ form a subset of A) and well-ordered. Now we look at the well-ordered set $S \cup \{+\infty\}$; by Proposition 2.4.13, it is not isomorphic to any subset of S, but $\rho(S \cup \{+\infty\})$ is necessarily contained in S, a contradiction. See also Section 3.15 for some related results and arguments in this spirit.

Remark 2.4.26. The same argument also shows that there is no representation of the ordinals inside a given set; the ordinals are "too big" to be placed in anything other than a class.

2.4.3. Zorn's lemma. Now we can prove Zorn's lemma. The key proposition is

Proposition 2.4.27. *Let X be a partially ordered set, and let $\mathcal{C}$ be the set of all well-ordered sets in X. Then there does not exist a*

function $g : \mathcal{C} \to X$ such that $g(C)$ is a strict upper bound for C (i.e. $g(C) > x$ for all $x \in C$) for all well-ordered $C \in \mathcal{C}$.

Proof. Suppose for contradiction that there existed X and g with the above properties. Then, given any well-ordered set Y, we claim that there exists exactly one isomorphism $\phi_Y : Y \to \rho(Y)$ from Y to a well-ordered set $\rho(Y)$ in X such that $\phi_Y(y) = g(\phi_Y([\min(Y), y)))$ for all $y \in Y$. Indeed, the uniqueness and existence can both be established by a transfinite induction that we leave as an exercise. (Informally, ϕ_Y is what one gets by "applying g Y times, starting with the empty set".) From uniqueness we see that $\rho(Y) = \rho(Y')$ whenever Y and Y' are isomorphic, and another transfinite induction shows that $\rho(Y) \subset \rho(Y')$ whenever Y is a subset of Y'. Thus ρ is a representation of the ordinals. But this contradicts Theorem 2.4.24. □

Remark 2.4.28. One can use transfinite induction on ordinals rather than well-ordered sets if one wishes here, using Remark 2.4.26 in place of Theorem 2.4.24.

Proof of Zorn's lemma. Suppose for contradiction that one had a non-empty partially ordered set X without maximal elements, such that every chain had an upper bound. As there are no maximal elements, every element in X must be bounded by a strictly larger element in X, and so every chain in fact has a strict upper bound; in particular every well-ordered set has a strict upper bound. Applying the axiom of choice, we may thus find a choice function $g : \mathcal{C} \to X$ from the space of well-ordered sets in X to X, that maps every such set to a strict upper bound. But this contradicts Proposition 2.4.27. □

Remark 2.4.29. It is important for Zorn's lemma that X is a set, rather than a class. Consider for instance the class of all ordinals. Every chain of ordinals has an upper bound (namely, the union of the ordinals in that chain), and the class is certainly non-empty, but there is no maximal ordinal. (Compare also Theorem 2.4.21 and Theorem 2.4.24.)

Remark 2.4.30. It is also important that every chain have an upper bound, and not just countable chains. Indeed, the collection of

countable subsets of an uncountable set (such as $\mathbf{R}$) is non-empty, and every countable chain has an upper bound, but there is no maximal element.

Remark 2.4.31. The above argument shows that the hypothesis of Zorn's lemma can be relaxed slightly; one does not need every chain to have an upper bound, merely every well-ordered set needs to have one. But I do not know of any application in which this apparently stronger version of Zorn's lemma dramatically simplifies an argument. (In practice, either Zorn's lemma can be applied routinely, or it fails utterly to be applicable at all.)

Exercise 2.4.11. Use Zorn's lemma to establish the *well-ordering theorem* (every set has at least one well-ordering).

Remark 2.4.32. By the above exercise, $\mathbf{R}$ can be well-ordered. However, if one drops the axiom of choice from the axioms of set theory, one can no longer prove that $\mathbf{R}$ is well-ordered. Indeed, given a well-ordering of $\mathbf{R}$, it is not difficult (using Remark 2.4.8) to remove the axiom of choice from the Banach-Tarski constructions in Section 2.2, and thus obtain constructions of non-measurable subsets of $\mathbf{R}$. But a deep theorem of Solovay gives a model of set theory (without the axiom of choice) in which every set of reals is measurable.

Exercise 2.4.12. Define a (von Neumann) cardinal to be an ordinal α with the property that all smaller ordinals have strictly lesser cardinality (i.e. cannot be placed in one-to-one correspondence with α). Show that every set can be placed in one-to-one correspondence with exactly one cardinal. (This gives a representation of the category of sets, similar to how ord gives a representation of well-ordered sets.)

It seems appropriate to close these notes with a quote from Jerry Bona:

"The Axiom of Choice is obviously true, the well-ordering principle obviously false, and who can tell about Zorn's Lemma?"

Notes. This lecture first appeared at `terrytao.wordpress.com/2009/01/28`. Thanks to an anonymous commenter for corrections.

Eric remarked that any application of Zorn's lemma can be equivalently rephrased as a transfinite induction, after using a choice function to decide where to go at each limit ordinal.

2.5. Compactification and metrisation

One way to study a general class of mathematical objects is to embed them into a more structured class of mathematical objects; for instance, one could study manifolds by embedding them into Euclidean spaces. In these notes we study two (related) embedding theorems for topological spaces:

- The *Stone-Čech compactification*, which embeds locally compact Hausdorff spaces into compact Hausdorff spaces in a "universal" fashion; and
- The *Urysohn metrization theorem*, that shows that every second-countable normal Hausdorff space is metrizable.

2.5.1. The Stone-Čech compactification. Observe that any dense open subset of a compact Hausdorff space is automatically a locally compact Hausdorff space. We now study the reverse concept:

Definition 2.5.1. A *compactification* of a locally compact Hausdorff space X is an embedding $\iota : X \to \overline{X}$ (i.e. a homeomorphism between X and $\iota(X)$) into a compact Hausdorff space $\overline{X}$ such that the image $\iota(X)$ of X is an open dense subset of $\overline{X}$. We will often abuse notation and refer to $\overline{X}$ as the compactification rather than the embedding $\iota : X \to \overline{X}$, when the embedding is obvious from context.

One compactification $\iota : X \to \overline{X}$ is *finer* than another $\iota' : X \to \overline{X}'$ (or $\iota' : X \to \overline{X}'$ is *coarser* than $\iota : X \to \overline{X}$) if there exists a continuous map $\pi : \overline{X}' \to \overline{X}$ such that $\iota = \pi \circ \iota'$; notice that this map must be surjective and unique, by the open dense nature of $\iota(X)$. Two compactifications are *equivalent* if they are both finer than each other.

Example 2.5.2. Any compact set can be its own compactification. The real line $\mathbf{R}$ can be compactified into $[-\pi/2, \pi/2]$ by using the arctan function as the embedding, or (equivalently) by embedding it into the extended real line $[-\infty, \infty]$. It can also be compactified into

the unit circle $\{(x,y) \in \mathbf{R}^2 : x^2+y^2=1\}$ by using the *stereographic projection* $x \mapsto (\frac{2x}{1+x^2}, \frac{x^2-1}{1+x^2})$. Notice that the former embedding is finer than the latter. The plane $\mathbf{R}^2$ can similarly be compactified into the unit sphere $\{(x,y,z) \in \mathbf{R}^2 : x^2+y^2+z^2=1\}$ by the stereographic projection $(x,y) \mapsto (\frac{2x}{1+x^2+y^2}, \frac{2y}{1+x^2+y^2}, \frac{x^2+y^2-1}{1+x^2+y^2})$.

Exercise 2.5.1. Let X be a locally compact Hausdorff space X that is not compact. Define the *one-point compactification* $X \cup \{\infty\}$ by adjoining one point ∞ to X, with the topology generated by the open sets of X, and the complement (in $X \cup \{\infty\}$) of the compact sets in X. Show that $X \cup \{\infty\}$ (with the obvious embedding map) is a compactification of X. Show that the one-point compactification is coarser than any other compactification of X.

We now consider the opposite extreme to the one-point compactification:

Definition 2.5.3. Let X be a locally compact Hausdorff space. A *Stone-Čech compactification* βX of X is defined as the finest compactification of X, i.e. the compactification of X which is finer than every other compactification of X.

It is clear that the Stone-Čech compactification, if it exists, is unique up to isomorphism, and so one often abuses notation by referring to *the* Stone-Čech compactification. The existence of the compactification can be established by Zorn's lemma (see Section 2.3 of *Poincaré's legacies, Vol. I*). We shall shortly give several other constructions of the compactification. (All constructions, however, rely at some point on the axiom of choice, or a related axiom.)

The Stone-Čech compactification obeys a useful *functorial* property:

Exercise 2.5.2. Let X, Y be locally compact Hausdorff spaces, with Stone-Čech compactifications $\beta X, \beta Y$. Show that every continuous map $f : X \to Y$ has a unique continuous extension $\beta f : \beta X \to \beta Y$. (*Hint*: uniqueness is easy; for existence, look at the closure of the graph $\{(x, f(x)) : x \in X\}$ in $\beta X \times \beta Y$, which compactifies X and thus cannot be strictly finer than βX.) In the converse direction, if $\overline{X}$ is a compactification of X such that every continuous map $f : X \to K$

into a compact space can be extended continuously to $\overline{X}$, show that $\overline{X}$ is the Stone-Čech compactification.

Example 2.5.4. From the above exercise, we can define limits $\lim_{x\to p} f(x) := \beta f(p)$ for any bounded continuous function on X and any $p \in \beta X$. But one for coarser compactifications, one can only take limits for special types of bounded continuous functions; for instance, using the one-point compactification of $\mathbf{R}$, $\lim_{x\to\infty} f(x)$ need not exist for a bounded continuous function $f : \mathbf{R} \to \mathbf{R}$, e.g. $\lim_{x\to\infty} \sin(x)$ or $\lim_{x\to\infty} \arctan(x)$ do not exist. The finer the compactification, the more limits can be defined; for instance the two point compactification $[-\infty, +\infty]$ of $\mathbf{R}$ allows one to define the limits $\lim_{x\to+\infty} f(x)$ and $\lim_{x\to-\infty} f(x)$ for some additional functions f (e.g. $\lim_{x\to\pm\infty} \arctan(x)$ is well-defined); and the Stone-Čech compactification is the only compactification which allows one to take limits for *any* bounded continuous function (e.g. $\lim_{x\to p} \sin(x)$ is well-defined for all $p \in \beta\mathbf{R}$).

Now we turn to the issue of actually constructing the Stone-Čech compactifications.

Exercise 2.5.3. Let X be a locally compact Hausdorff space. Let $C(X \to [0,1])$ be the space of continuous functions from X to the unit interval, let $Q := [0,1]^{C(X\to[0,1])}$ be the space of tuples $(y_f)_{f\in C(X\to[0,1])}$ taking values in the unit interval, with the product topology, and let $\iota : X \to Q$ be the Gelfand transform $\iota(x) := (f(x))_{f\in C(X\to[0,1])}$, and let βX be the closure of ιX in Q.

- Show that βX is a compactification of X. (*Hint*: Use Urysohn's lemma and Tychonoff's theorem.)
- Show that βX is the Stone-Čech compactification of X. (*Hint*: If $\overline{X}$ is any other compactification of X, we can identify $C(\overline{X} \to [0,1])$ as a subset of $C(X \to [0,1])$, and then project Q to $[0,1]^{C(\overline{X}\to[0,1])}$. Meanwhile, we can embed $\overline{X}$ inside $[0,1]^{C(\overline{X}\to[0,1])}$ by the Gelfand transform.)

Exercise 2.5.4. Let X be a discrete topological space, let 2^X be the *Boolean algebra* of all subsets of X. By Stone's representation theorem (Theorem 1.2.2), 2^X is isomorphic to the clopen algebra of a Stone space βX.

- Show that βX is a compactification of X.
- Show that βX is the Stone-Čech compactification of X.
- Identify βX with the space of *ultrafilters* on X. (See Section 1.5 of *Structure and randomness* for further discussion of ultrafilters, and Section 2.3 of *Poincaré's legacies, Vol. I* for further discussion of the relationship of ultrafilters to the Stone-Čech compactification.)

Exercise 2.5.5. Let X be a locally compact Hausdorff space, and let $BC(X \to \mathbf{C})$ be the space of bounded continuous complex-valued functions on X.

- Show that $BC(X \to \mathbf{C})$ is a *unital commutative C^*-algebra* (see Section 1.10.4).
- By the commutative Gelfand-Naimark theorem (Theorem 1.10.24), $BC(X \to \mathbf{C})$ is isomorphic as a unital C^*-algebra to $C(\beta X \to \mathbf{C})$ for some compact Hausdorff space βX (which is in fact the spectrum of $BC(X \to \mathbf{C})$. Show that βX is the Stone-Čech compactification of X.
- More generally, show that given any other compactification $\overline{X}$ of X, that $C(\overline{X} \to \mathbf{C})$ is isomorphic as a unital C^*-algebra to a subalgebra of $BC(X \to \mathbf{C})$ that contains $\mathbf{C} \oplus C_0(X \to \mathbf{C})$ (the space of continuous functions from X to $\mathbf{C}$ that converge to a limit at ∞), with $\overline{X}$ as the spectrum of this algebra; thus we have a canonical identification between compactifications and C^*-algebras between $BC(X \to \mathbf{C})$ and $\mathbf{C} \oplus C_0(X \to \mathbf{C})$, which correspond to the Stone-Čech compactification and one-point compactification respectively.

Exercise 2.5.6. Let X be a locally compact Hausdorff space. Show that the dual $BC(X \to \mathbf{R})^*$ of $BC(X \to \mathbf{R})$ is isomorphic as a Banach space to the space $M(\beta X)$ of real signed Radon measures on the Stone-Čech compactification βX, and similarly in the complex case. In particular, conclude that $\ell^\infty(\mathbf{N})^* \equiv M(\beta\mathbf{N})$.

Remark 2.5.5. The Stone-Čech compactification can be extended from locally compact Hausdorff spaces to the slightly larger class of

Tychonoff spaces, which are those Hausdorff spaces X with the property that any closed set $K \subset X$ and point x not in K can be separated by a continuous function $f \in C(X \to \mathbf{R})$ which equals 1 on K and zero on x. This compactification can be constructed by a modification of the argument used to establish Exercise 2.5.3. However, in this case the space X is merely dense in its compactification βX, rather than open and dense.

Remark 2.5.6. A cautionary note: in general, the Stone-Čech compactification is almost never *sequentially* compact. For instance, it is not hard to show that $\mathbf{N}$ is sequentially closed in $\beta\mathbf{N}$. In particular, these compactifications are usually not metrisable.

2.5.2. Urysohn's metrisation theorem. Recall that a topological space is *metrisable* if there exists a metric on that space which generates the topology. There are various necessary conditions for metrisability. For instance, we have already seen that metric spaces must be normal and Hausdorff. In the converse direction, we have

Theorem 2.5.7 (Urysohn's metrisation theorem)**.** *Let X be a normal Hausdorff space which is* second countable*. Then X is metrisable.*

Proof. (Sketch) This will be a variant of the argument in Exercise 2.5.3, but with a countable family of continuous functions in place of $C(X \to [0,1])$.

Let $U_1, U_2, \ldots$ be a countable base for X. If U_i, U_j are in this base with $\overline{U_i} \subset U_j$, we can apply Urysohn's lemma and find a continuous function $f_{ij} : X \to [0,1]$ which equals 1 on $\overline{U_i}$ and vanishes outside of U_j. Let $\mathcal{F}$ be the collection of all such functions; this is a countable family. We can then embed X in $[0,1]^{\mathcal{F}}$ using the Gelfand transform $x \mapsto (f(x))_{f\in\mathcal{F}}$. By modifying the proof of Exercise 2.5.3 one can show that this is an embedding. On the other hand, $[0,1]^{\mathcal{F}}$ is a countable product of metric spaces and is thus metrisable (e.g. by enumerating $\mathcal{F}$ as $f_1, f_2, \ldots$ and using the metric $d((x_n)_{f_n\in\mathcal{F}}, (y_n)_{f_n\in\mathcal{F}}) := \sum_{n=1}^{\infty} 2^{-n}|x_n - y_n|$). Since a subspace of a metrisable space is clearly also metrisable, the claim follows. □

Recalling that compact metric spaces are second countable (Lemma 1.8.6), thus we have

Corollary 2.5.8. *A compact Hausdorff space is metrisable if and only if it is second countable.*

Of course, non-metrisable compact Hausdorff spaces exist; $\beta\mathbf{N}$ is a standard example. Uncountable products of non-trivial compact metric spaces, such as $\{0,1\}$, are always non-metrisable. Indeed, we already saw in Section 1.8 that $\{0,1\}^X$ is compact but not sequentially compact (and thus not metrisable) when X has the cardinality of the continuum; one can use the first uncountable ordinal to achieve a similar result for any uncountable X, and then by embedding one can obtain non-metrisability for any uncountable product of non-trivial compact metric spaces, thus complementing the metrisability of countable products of such spaces. Conversely, there also exist metrisable spaces which are not second countable (e.g. uncountable discrete spaces). So Urysohn's metrisation theorem does not completely classify the metrisable spaces, however it already covers a large number of interesting cases.

Notes. This lecture first appeared at `terrytao.wordpress.com/2009/03/02`. Thanks to Eric, Javier Lopez, Mark Meckes, Max Baroi, Paul Leopardi, Pete L. Clark, and anonymous commenters for corrections.

2.6. Hardy's uncertainty principle

Many properties of a (sufficiently nice) function $f : \mathbf{R} \to \mathbf{C}$ are reflected in its *Fourier transform* $\hat{f} : \mathbf{R} \to \mathbf{C}$, defined by the formula

$$\hat{f}(\xi) := \int_{-\infty}^{\infty} f(x) e^{-2\pi i x\xi}\, dx. \tag{2.17}$$

For instance, decay properties of f are reflected in smoothness properties of $\hat{f}$, as the following table shows:

If f is...	then $\hat{f}$ is...	and this relates to...
Square-integrable	square-integrable	*Plancherel's theorem*
Absolutely integrable	continuous	*Riemann-Lebesgue lemma*
Rapidly decreasing	smooth	theory of Schwartz functions
Exponentially decreasing	analytic in a strip	
Compactly supported	entire, exponential growth	*Paley-Wiener theorem*

(See Section 1.12 for further discussion of the Fourier transform.)

Another important relationship between a function f and its Fourier transform $\hat{f}$ is the *uncertainty principle*, which roughly asserts that if a function f is highly localised in space, then its Fourier transform $\hat{f}$ must be widely dispersed in space, or to put it another way, f and $\hat{f}$ cannot both decay too strongly at infinity (except of course in the degenerate case $f = 0$). There are many ways to make this intuition precise. One of them is the *Heisenberg uncertainty principle*, which asserts that if we normalise

$$\int_{\mathbf{R}} |f(x)|^2 \, dx = \int_{\mathbf{R}} |\hat{f}(\xi)|^2 \, d\xi = 1$$

then we must have

$$(\int_{\mathbf{R}} |x|^2 |f(x)|^2 \, dx) \cdot (\int_{\mathbf{R}} |\xi|^2 |\hat{f}(\xi)|^2 \, dx) \geq \frac{1}{(4\pi)^2}$$

thus forcing at least one of f or $\hat{f}$ to not be too concentrated near the origin. This principle can be proven (for sufficiently nice f, initially) by observing the integration by parts identity

$$\langle xf, f' \rangle = \int_{\mathbf{R}} xf(x)\overline{f'(x)} \, dx = -\frac{1}{2} \int_{\mathbf{R}} |f(x)|^2 \, dx$$

and then using *Cauchy-Schwarz* and the Plancherel identity.

Another well known manifestation of the uncertainty principle is the fact that it is not possible for f and $\hat{f}$ to both be compactly supported (unless of course they vanish entirely). This can be in fact be seen from the above table: if f is compactly supported, then $\hat{f}$ is an entire function; but the zeroes of a non-zero entire function are isolated, yielding a contradiction unless f vanishes. (Indeed, the table also shows that if one of f and $\hat{f}$ is compactly supported, then the other cannot have exponential decay.)

On the other hand, we have the example of the *Gaussian functions* $f(x) = e^{-\pi a x^2}$, $\hat{f}(\xi) = \frac{1}{\sqrt{a}} e^{-\pi \xi^2/a}$, which both decay faster than exponentially. The classical *Hardy uncertainty principle* asserts, roughly speaking, that this is the fastest that f and $\hat{f}$ can simultaneously decay:

Theorem 2.6.1 (Hardy uncertainty principle)**.** *Suppose that f is a (measurable) function such that $|f(x)| \leq Ce^{-\pi a x^2}$ and $|\hat{f}(\xi)| \leq$*

$Ce^{-\pi a x^2}$ for all x, ξ and some $C, C', a > 0$. Then $f(x)$ is a scalar multiple of the gaussian $e^{-\pi a x^2}$.

This theorem is proven by complex-analytic methods, in particular the *Phragmén-Lindelöf principle*; for sake of completeness we give that proof below. But I was curious to see if there was a real-variable proof of the same theorem, avoiding the use of complex analysis. I was able to find the proof of a slightly weaker theorem:

Theorem 2.6.2 (Weak Hardy uncertainty principle)**.** *Suppose that f is a non-zero (measurable) function such that $|f(x)| \leq Ce^{-\pi a x^2}$ and $|\hat{f}(\xi)| \leq C' e^{-\pi b \xi^2}$ for all x, ξ and some $C, C', a, b > 0$. Then $ab \leq C_0$ for some absolute constant C_0.*

Note that the correct value of C_0 should be 1, as is implied by the true Hardy uncertainty principle. Despite the weaker statement, I thought the proof might still might be of interest as it is a little less "magical" than the complex-variable one, and so I am giving it below.

2.6.1. The complex-variable proof. We first give the complex-variable proof. By dilating f by $\sqrt{a}$ (and contracting $\hat{f}$ by $1/\sqrt{a}$) we may normalise $a = 1$. By multiplying f by a small constant we may also normalise $C = C' = 1$.

The super-exponential decay of f allows us to extend the Fourier transform $\hat{f}$ to the complex plane, thus

$$\hat{f}(\xi + i\eta) = \int_{\mathbf{R}} f(x) e^{-2\pi i x \xi} e^{2\pi \eta x}\, dx$$

for all $\xi, \eta \in \mathbf{R}$. We may differentiate under the integral sign and verify that $\hat{f}$ is entire. Taking absolute values, we obtain the upper bound

$$|\hat{f}(\xi + i\eta)| \leq \int_{\mathbf{R}} e^{-\pi x^2} e^{2\pi \eta x}\, dx;$$

completing the square, we obtain

$$|\hat{f}(\xi + i\eta)| \leq e^{\pi \eta^2} \tag{2.18}$$

for all ξ, η. We conclude that the entire function

$$F(z) := e^{\pi z^2} \hat{f}(z)$$

is bounded in magnitude by 1 on the imaginary axis; also, by hypothesis on $\hat{f}$, we also know that F is bounded in magnitude by 1 on the real axis. *Formally* applying the *Phragmen-Lindelöf principle* (or *maximum modulus principle*), we conclude that F is bounded on the entire complex plane, which by Liouville's theorem implies that F is constant, and the claim follows.

Now let's go back and justify the Phragmén-Lindelöf argument. Strictly speaking, Phragmén-Lindelöf does not apply, since it requires exponential growth on the function F, whereas we have quadratic-exponential growth here. But we can tweak F a bit to solve this problem. Firstly, we pick $0 < \theta < \pi/2$ and work on the sector

$$\Gamma_\theta := \{re^{i\alpha} : r > 0, 0 \leq \alpha \leq \theta\}.$$

Using (2.18) we have

$$|F(\xi + i\eta)| \leq e^{\pi\xi^2}.$$

Thus, if $\delta > 0$, and θ is sufficiently close to $\pi/2$ depending on δ, the function $e^{i\delta z^2}F(z)$ is bounded in magnitude by 1 on the boundary of Γ_θ. Then, for any sufficiently small $\varepsilon > 0$, $e^{-i\varepsilon e^{i\varepsilon}z^{2+\varepsilon}}e^{i\delta z^2}F(z)$ (using the standard branch of $z^{2+\varepsilon}$ on Γ_θ) is also bounded in magnitude by 1 on this boundary, and goes to zero at infinity in the interior of Γ_θ, so is bounded by 1 in that interior by the maximum modulus principle. Sending $\varepsilon \to 0$, and then $\theta \to \pi/2$, and then $\delta \to 0$, we obtain F bounded in magnitude by 1 on the upper right quadrant. Similar arguments work for the other quadrants, and the claim follows.

2.6.2. The real-variable proof. Now we turn to the real-variable proof of Theorem 2.6.2, which is based on the fact that polynomials of controlled degree do not resemble rapidly decreasing functions.

Rather than use complex analyticity $\hat{f}$, we will rely instead on a different relationship between the decay of f and the regularity of $\hat{f}$, as follows:

Lemma 2.6.3 (Derivative bound)**.** *Suppose that $|f(x)| \leq Ce^{-\pi ax^2}$ for all $x \in \mathbf{R}$, and some $C, a > 0$. Then $\hat{f}$ is smooth, and furthermore one has the bound $|\partial_\xi^k \hat{f}(\xi)| \leq \frac{C}{\sqrt{a}} \frac{k!\pi^{k/2}}{(k/2)!a^{(k+1)/2}}$ for all $\xi \in \mathbf{R}$ and every even integer k.*

Proof. The smoothness of $\hat{f}$ follows from the rapid decrease of f. To get the bound, we differentiate under the integral sign (one can easily check that this is justified) to obtain

$$\partial_\xi^k \hat{f}(\xi) = \int_{\mathbf{R}} (-2\pi i x)^k f(x) e^{-2\pi i x\xi}\ dx$$

and thus by the triangle inequality for integrals (and the hypothesis that k is even)

$$|\partial_\xi^k \hat{f}(\xi)| \leq C \int_{\mathbf{R}} e^{-\pi a x^2} (2\pi x)^k\ dx.$$

On the other hand, by differentiating the Fourier analytic identity

$$\frac{1}{\sqrt{a}} e^{-\pi\xi^2/a} = \int_{\mathbf{R}} e^{-\pi a x^2} e^{-2\pi i x\xi}\ dx$$

k times at $\xi = 0$, we obtain

$$\frac{d^k}{d\xi^k}(\frac{1}{\sqrt{a}} e^{-\pi\xi^2/a})|_{\xi=0} = \int_{\mathbf{R}} e^{-\pi a x^2} (2\pi i x)^k\ dx;$$

expanding out $\frac{1}{\sqrt{a}} e^{-\pi\xi^2/a}$ using Taylor series we conclude that

$$\frac{k!}{\sqrt{a}} \frac{(-\pi/a)^{k/2}}{(k/2)!} = \int_{\mathbf{R}} e^{-\pi a x^2} (2\pi i x)^k\ dx$$

□

Using *Stirling's formula* $k! = k^k(e + o(1))^{-k}$, we conclude in particular that

$$|\partial_\xi^k \hat{f}(\xi)| \leq (\frac{\pi e}{a} + o(1))^{k/2} k^{k/2} \tag{2.19}$$

for all large even integers k (where the decay of $o(1)$ can depend on a, C).

We can combine (2.19) with Taylor's theorem with remainder, to conclude that on any interval $I \subset \mathbf{R}$, we have an approximation

$$\hat{f}(\xi) = P_I(\xi) + O(\frac{1}{k!}(\frac{\pi e}{a} + o(1))^{k/2} k^{k/2} |I|^k)$$

where $|I|$ is the length of I and P_I is a polynomial of degree less than k. Using Stirling's formula again, we obtain

$$\hat{f}(\xi) = P_I(\xi) + O((\frac{\pi}{ea} + o(1))^{k/2} k^{-k/2} |I|^k) \tag{2.20}$$

Now we apply a useful bound.

Lemma 2.6.4 (Doubling bound)**.** *Let P be a polynomial of degree at most k for some $k \geq 1$, let $I = [x_0 - r, x_0 + r]$ be an interval, and suppose that $|P(x)| \leq A$ for all $x \in I$ and some $A > 0$. Then for any $N \geq 1$ we have the bound $|P(x)| \leq (CN)^k A$ for all $x \in NI := [x_0 - Nr, x_0 + Nr]$ and for some absolute constant C.*

Proof. By translating we may take $x_0 = 0$; by dilating we may take $r = 1$. By dividing P by A, we may normalise $A = 1$. Thus we have $|P(x)| \leq 1$ for all $-1 \leq x \leq 1$, and the aim is now to show that $|P(x)| \leq (CN)^k$ for all $-N \leq x \leq N$.

Consider the trigonometric polynomial $P(\cos\theta)$. By *de Moivre's formula*, this function is a linear combination of $\cos(j\theta)$ for $0 \leq j \leq k$. By Fourier analysis, we can thus write $P(\cos\theta) = \sum_{j=0}^k c_j \cos(j\theta)$, where

$$c_j = \frac{1}{\pi} \int_{-\pi}^{\pi} P(\cos\theta)\cos(j\theta)\ d\theta.$$

Since $P(\cos\theta)$ is bounded in magnitude by 1, we conclude that c_j is bounded in magnitude by 2. Next, we use de Moivre's formula again to expand $\cos(j\theta)$ as a linear combination of $\cos(\theta)$ and $\sin^2(\theta)$, with coefficients of size $O(1)^k$; expanding $\sin^2(\theta)$ further as $1 - \cos^2(\theta)$, we see that $\cos(j\theta)$ is a polynomial in $\cos(\theta)$ with coefficients $O(1)^k$. Putting all this together, we conclude that the coefficients of P are all of size $O(1)^k$, and the claim follows. □

Remark 2.6.5. One can get slightly sharper results by using the theory of *Chebyshev polynomials*. (Is the best bound for C known? I do not know the recent literature on this subject. I think though that even the sharpest bound for C would not fully recover the sharp Hardy uncertainty principle, at least with the argument given here.)

We return to the proof of Theorem 2.6.2. We pick a large integer k and a parameter $r > 0$ to be chosen later. From (2.20) we have

$$\hat{f}(\xi) = P_r(\xi) + O(\frac{r^2}{ak})^{k/2}$$

for $\xi \in [-r, 2r]$, and some polynomial P_r of degree k. In particular, we have

$$P_r(\xi) = O(e^{-br^2}) + O(\frac{r^2}{ak})^{k/2}$$

for $\xi \in [r, 2r]$. Applying Lemma 2.6.4, we conclude that

$$P_r(\xi) = O(1)^k e^{-br^2} + O(\frac{r^2}{ak})^{k/2}$$

for $\xi \in [-r, r]$. Applying (2.20) again we conclude that

$$\hat{f}(\xi) = O(1)^k e^{-br^2} + O(\frac{r^2}{ak})^{k/2}$$

for $\xi \in [-r, r]$. If we pick $r := \sqrt{\frac{k}{cb}}$ for a sufficiently small absolute constant c, we conclude that

$$|\hat{f}(\xi)| \leq 2^{-k} + O(\frac{1}{ab})^{k/2}$$

(say) for $\xi \in [-r, r]$. If $ab \geq C_0$ for large enough C_0, the right-hand side goes to zero as $k \to \infty$ (which also implies $r \to \infty$), and we conclude that $\hat{f}$ (and hence f) vanishes identically.

Notes. This article first appeared at `terrytao.wordpress.com/2009/02/18`.

Pedro Lauridsen Ribiero noted an old result of Schrödinger, that the only minimisers of the Heisenberg uncertainty principle were the gaussians (up to scaling, translation, and modulation symmetries).

Fabrice Planchon and Phillipe Jaming mentioned several related results and generalisations, including a recent PDE-based proof of the Hardy uncertainty principle (with the sharp constant) in [**EsKePoVe2008**].

2.7. Create an epsilon of room

In this article I would like to discuss a fundamental trick in "soft" analysis, sometimes known as the "limiting argument" or "epsilon regularisation argument".

A quick description of the trick is as follows. Suppose one wants to prove some statement S_0 about some object x_0 (which could be a number, a point, a function, a set, etc.). To do so, pick a small $\varepsilon > 0$, and first prove a weaker statement S_ε (which allows for "losses" which go to zero as $\varepsilon \to 0$) about some perturbed object x_ε. Then,

take limits $\varepsilon \to 0$. Provided that the dependency and continuity of the weaker conclusion S_ε on ε are sufficiently controlled, and x_ε is converging to x_0 in an appropriately strong sense, you will recover the original statement.

One can of course play a similar game when proving a statement S_∞ about some object X_∞, by first proving a weaker statement S_N on some approximation X_N to X_∞ for some large parameter N, and then send $N \to \infty$ at the end.

Here are some typical examples of a target statement S_0, and the approximating statements S_ε that would converge to S:

S_0	S_ε
$f(x_0) = g(x_0)$	$f(x_\varepsilon) = g(x_\varepsilon) + o(1)$
$f(x_0) \leq g(x_0)$	$f(x_\varepsilon) \leq g(x_\varepsilon) + o(1)$
$f(x_0) > 0$	$f(x_\varepsilon) \geq c - o(1)$ for some $c > 0$ independent of ε
$f(x_0)$ is finite	$f(x_\varepsilon)$ is bounded uniformly in ε
$f(x_0) \geq f(x)$ for all $x \in X$	$f(x_\varepsilon) \geq f(x) - o(1)$ for all $x \in X$
(i.e. x_0 maximises f)	(i.e. x_ε nearly maximises f)
$f_n(x_0)$ converges as $n \to \infty$	$f_n(x_\varepsilon)$ fluctuates by at most $o(1)$ for suff. large n
f_0 is a measurable function	f_ε is a measurable function converging pointwise to f_0
f_0 is a continuous function	f_ε is an equicts. family of functions converging pointwise to f_0 OR f_ε is continuous and converges (locally) uniformly to f_0
The event E_0 holds a.s.	The event E_ε holds with probability $1 - o(1)$
The statement $P_0(x)$ holds for a.e. x	The statement $P_\varepsilon(x)$ holds for x outside of a set of measure $o(1)$

Of course, to justify the convergence of S_ε to S_0, it is necessary that x_ε converge to x_0 (or f_ε converge to f_0, etc.) in a suitably strong sense. (But for the purposes of proving just upper bounds, such as $f(x_0) \leq M$, one can often get by with quite weak forms of convergence, thanks to tools such as *Fatou's lemma* or the weak closure of the unit ball.) Similarly, we need some continuity (or at least semi-continuity) hypotheses on the functions f, g appearing above.

It is also necessary in many cases that the control S_ε on the approximating object x_ε is somehow "uniform in ε", although for "σ-closed" conclusions, such as measurability, this is not required[5].

By giving oneself an epsilon of room, one can evade a lot of familiar issues in soft analysis. For instance, by replacing "rough", "infinite-complexity", "continuous", "global", or otherwise "infinitary" objects x_0 with "smooth", "finite-complexity", "discrete", "local", or otherwise "finitary" approximants x_ε, one can finesse most issues regarding the justification of various formal operations (e.g. exchanging limits, sums, derivatives, and integrals)[6]. Similarly, issues such as whether the supremum $M := \sup\{f(x) : x \in X\}$ of a function on a set is actually attained by some maximiser x_0 become moot if one is willing to settle instead for an almost-maximiser x_ε, e.g. one which comes within an epsilon of that supremum M (or which is larger than $1/\varepsilon$, if M turns out to be infinite). Last, but not least, one can use the epsilon room to avoid degenerate solutions, for instance by perturbing a non-negative function to be strictly positive, perturbing a non-strictly monotone function to be strictly monotone, and so forth.

To summarise: one can view the epsilon regularisation argument as a "loan" in which one borrows an epsilon here and there in order to be able to ignore soft analysis difficulties, and can temporarily be able to utilise estimates which are non-uniform in epsilon, but at the end of the day one needs to "pay back" the loan by establishing a final "hard analysis" estimate which is uniform in epsilon (or whose error terms decay to zero as epsilon goes to zero).

A variant: It may seem that the epsilon regularisation trick is useless if one is already in "hard analysis" situations when all objects are already "finitary", and all formal computations easily justified. However, there is an important variant of this trick which applies in this case: namely, instead of sending the epsilon parameter to zero, choose

[5]It is important to note that it is only the final conclusion S_ε on x_ε that needs to have this uniformity in ε; one is permitted to have some intermediate stages in the derivation of S_ε that depend on ε in a non-uniform manner, so long as these non-uniformities cancel out or otherwise disappear at the end of the argument.

[6]It is important to be aware, though, that any quantitative measure on how smooth, discrete, finite, etc. x_ε should be expected to degrade in the limit $\varepsilon \to 0$, and so one should take extreme caution in using such quantitative measures to derive estimates that are uniform in ε.

epsilon to be a sufficiently small (but not infinitesimally small) quantity, depending on other parameters in the problem, so that one can eventually neglect various error terms and to obtain a useful bound at the end of the day. (For instance, any result proven using the *Szemerédi regularity lemma* is likely to be of this type.) Since one is not sending epsilon to zero, not every term in the final bound needs to be uniform in epsilon, though for quantitative applications one still would like the dependencies on such parameters to be as favourable as possible.

2.7.1. Examples. The "soft analysis" components of any real analysis textbook will contain a large number of examples of this trick in action. In particular, any argument which exploits Littlewood's three principles of real analysis is likely to utilise this trick. Of course, this trick also occurs repeatedly in Chapter 1, and thus was chosen as the title of this book.

Example 2.7.1 (Riemann-Lebesgue lemma)**.** Given any absolutely integrable function $f \in L^1(\mathbf{R})$, the Fourier transform $\hat{f} : \mathbf{R} \to \mathbf{C}$ is defined by the formula

$$\hat{f}(\xi) := \int_{\mathbf{R}} f(x) e^{-2\pi i x\xi}\, dx.$$

The *Riemann-Lebesgue lemma* asserts that $\hat{f}(\xi) \to 0$ as $\xi \to \infty$. It is difficult to prove this estimate for f directly, because this function is too "rough": it is absolutely integrable (which is enough to ensure that $\hat{f}$ exists and is bounded), but need not be continuous, differentiable, compactly supported, bounded, or otherwise "nice". But suppose we give ourselves an epsilon of room. Then, as the space C_0^∞ of test functions is dense in $L^1(\mathbf{R})$ (Exercise 1.13.5), we can approximate f to any desired accuracy $\varepsilon > 0$ in the L^1 norm by a smooth, compactly supported function $f_\varepsilon : \mathbf{R} \to \mathbf{C}$, thus

$$\int_{\mathbf{R}} |f(x) - f_\varepsilon(x)|\, dx \leq \varepsilon. \tag{2.21}$$

The point is that f_ε is much better behaved than f, and it is not difficult to show the analogue of the Riemann-Lebesgue lemma for f_ε. Indeed, being smooth and compactly supported, we can now

justifiably integrate by parts to obtain

$$\hat{f}_\varepsilon(\xi) = \frac{1}{2\pi i\xi} \int_{\mathbf{R}} f'_\varepsilon(x) e^{-2\pi i x\xi}\, dx$$

for any non-zero ξ, and it is now clear (since f' is bounded and compactly supported) that $\hat{f}_\varepsilon(\xi) \to 0$ as $\xi \to \infty$.

Now we need to take limits as $\varepsilon \to 0$. It will be enough to have $\hat{f}_\varepsilon$ converge uniformly to $\hat{f}$. But from (2.21) and the basic estimate

$$\sup_\xi |\hat{g}(\xi)| \leq \int_{\mathbf{R}} |g(x)|\, dx \tag{2.22}$$

(which is the single "hard analysis" ingredient in the proof of the lemma) applied to $g := f - f_\varepsilon$, we see (by the linearity of the Fourier transform) that

$$\sup_\xi |\hat{f}(\xi) - \hat{f}_\varepsilon(\xi)| \leq \varepsilon$$

and we obtain the desired uniform convergence.

Remark 2.7.2. The same argument also shows that $\hat{f}$ is continuous; we leave this as an exercise to the reader. See also Exercise 1.12.11 for the generalisation of this lemma to other locally compact abelian groups.

Remark 2.7.3. Example 2.7.1 is a model case of a much more general instance of the limiting argument: in order to prove a convergence or continuity theorem for all "rough" functions in a function space, it suffices to first prove convergence or continuity for a dense subclass of "smooth" functions, and combine that with some quantitative estimate in the function space (in this case, (2.22)) in order to justify the limiting argument. See Corollary 1.7.7 for an important example of this principle.

Example 2.7.4. The limiting argument in Example 2.7.1 relied on the linearity of the Fourier transform $f \mapsto \hat{f}$. But, with more effort, it is also possible to extend this type of argument to nonlinear settings. We will sketch (omitting several technical details, which can be found for instance in my PDE book [**Ta2006**]) a very typical instance. Consider a nonlinear PDE, e.g. the cubic nonlinear wave equation

$$-u_{tt} + u_{xx} = u^3 \tag{2.23}$$

where $u : \mathbf{R} \times \mathbf{R} \to \mathbf{R}$ is some scalar field, and the t and x subscripts denote differentiation of the field $u(t,x)$. If u is sufficiently smooth, and sufficiently decaying at spatial infinity, one can show that the energy

$$E(u)(t) := \int_{\mathbf{R}} \frac{1}{2}|u_t(t,x)|^2 + \frac{1}{2}|u_x(t,x)|^2 + \frac{1}{4}|u(t,x)|^4 \, dx \tag{2.24}$$

is conserved, thus $E(u)(t) = E(u)(0)$ for all t. Indeed, this can be formally justified by computing the derivative $\partial_t E(u)(t)$ by differentiating under the integral sign, integrating by parts, and then applying the PDE (2.23); we leave this as an exercise for the reader[7]. However, these justifications do require a fair amount of regularity on the solution u; for instance, requiring u to be three-times continuously differentiable in space and time, and compactly supported in space on each bounded time interval, would be sufficient to make the computations rigorous by applying "off the shelf" theorems about differentiation under the integration sign, etc.

But suppose one only has a much rougher solution, for instance an energy class solution which has finite energy (2.24), but for which higher derivatives of u need not exist in the classical sense[8]. Then it is difficult to justify the energy conservation law directly. However, it is still possible to obtain energy conservation by the limiting argument. Namely, one takes the energy class solution u at some initial time (e.g. $t = 0$) and approximates that initial data (the initial position $u(0)$ and initial data $u_t(0)$) by a much smoother (and compactly supported) choice $(u^{(\varepsilon)}(0), u_t^{(\varepsilon)}(0))$ of initial data, which converges back to $(u(0), u_t(0))$ in a suitable "energy topology" related to (2.24), which we will not define here (it is based on Sobolev spaces, which are discussed in Section 1.14). It then turns out (from the existence theory of the PDE (2.23)) that one can extend the smooth

[7] There are also more fancy ways to see why the energy is conserved, using Hamiltonian or Lagrangian mechanics or by the more general theory of stress-energy tensors, but we will not discuss these here.

[8] There is a non-trivial issue regarding how to make sense of the PDE (2.23) when u is only in the energy class, since the terms u_{tt} and u_{xx} do not then make sense classically, but there are standard ways to deal with this, e.g. using *weak derivatives*, see Section 1.13.

initial data $(u^{(\varepsilon)}(0), u_t^{(\varepsilon)}(0))$ to other times t, providing a smooth solution $u^{(\varepsilon)}$ to that data. For this solution, the energy conservation law $E(u^{(\varepsilon)})(t) = E(u^{(\varepsilon)})(0)$ can be justified.

Now we take limits as $\varepsilon \to 0$ (keeping t fixed). Since $(u^{(\varepsilon)}(0), u_t^{(\varepsilon)}(0))$ converges in the energy topology to $(u(0), u_t(0))$, and the energy functional E is continuous in this topology, $E(u^{(\varepsilon)})(0)$ converges to $E(u)(0)$. To conclude the argument, we will also need $E(u^{(\varepsilon)})(t)$ to converge to $E(u)(t)$, which will be possible if $(u^{(\varepsilon)}(t), u_t^{(\varepsilon)}(t))$ converges in the energy topology to $(u(t), u_t(t))$. Thus in turn follows from a fundamental fact (which requires a certain amount of effort to prove) about the PDE to (2.24), namely that it is well-posed in the energy class. This means that not only do solutions exist and are unique for initial data in the energy class, but they depend continuously on the initial data in the energy topology; small perturbations in the data lead to small perturbations in the solution, or more formally that the map $(u(0), u_t(0)) \to (u(t), u_t(t))$ from data to solution (say, at some fixed time t) is continuous in the energy topology. This final fact concludes the limiting argument and gives us the desired conservation law $E(u(t)) = E(u(0))$.

Remark 2.7.5. It is important that one have a suitable well-posedness theory in order to make the limiting argument work for rough solutions to a PDE; without such a well-posedness theory, it is possible for quantities which are formally conserved to cease being conserved when the solutions become too rough or otherwise "weak"; energy, for instance, could disappear into a singularity and not come back.

Example 2.7.6 (Maximum principle)**.** The maximum principle is a fundamental tool in elliptic and parabolic PDE (for example, it is used heavily in the proof of the Poincaré conjecture, discussed extensively in *Poincaré's legacies, Vol. II*). Here is a model example of this principle:

Proposition 2.7.7. *Let $u : \overline{\mathbf{D}} \to \mathbf{R}$ be a smooth harmonic function on the closed unit disk $\overline{\mathbf{D}} := \{(x,y) : x^2 + y^2 \leq 1\}$. If M is a bound such that $u(x,y) \leq M$ on the boundary $\partial\mathbf{D} := \{(x,y) : x^2 + y^2 = 1\}$. Then $u(x,y) \leq M$ on the interior as well.*

A naive attempt to prove Proposition 2.7.7 comes very close to working, and goes like this: suppose for contradiction that the proposition failed, thus u exceeds M somewhere in the interior of the disk. Since u is continuous, and the disk is compact, there must then be a point (x_0, y_0) in the interior of the disk where the maximum is attained. Undergraduate calculus then tells us that $u_{xx}(x_0, y_0)$ and $u_{yy}(x_0, y_0)$ are non-positive, which almost contradicts the harmonicity hypothesis $u_{xx} + u_{yy} = 0$. However, it is still possible that u_{xx} and u_{yy} both vanish at (x_0, y_0), so we don't yet get a contradiction.

But we can finish the proof by giving ourselves an epsilon of room. The trick is to work not with the function u directly, but with the modified function $u^{(\varepsilon)}(x, y) := u(x, y) + \varepsilon(x^2 + y^2)$, to boost the harmonicity into subharmonicity. Indeed, we have $u^{(\varepsilon)}_{xx} + u^{(\varepsilon)}_{yy} = 4\varepsilon > 0$. The preceding argument now shows that $u^{(\varepsilon)}$ cannot attain its maximum in the interior of the disk; since it is bounded by $M + \varepsilon$ on the boundary of the disk, we we conclude that $u^{(\varepsilon)}$ is bounded by $M + \varepsilon$ on the interior of the disk as well. Sending $\varepsilon \to 0$ we obtain the claim.

Remark 2.7.8. Of course, Proposition 2.7.7 can also be proven by much more direct means, for instance via the *Green's function* for the disk. However, the argument given is extremely robust and applies to a large class of both linear and nonlinear elliptic and parabolic equations, including those with rough variable coefficients.

Exercise 2.7.1. Use the maximum modulus principle to prove the *Phragmén-Lindelöf principle*: if f is complex analytic on the strip $\{z : 0 \leq \mathrm{Re}(z) \leq 1\}$, is bounded in magnitude by 1 on the boundary of this strip, and obeys a growth condition $|f(z)| \leq Ce^{|z|^C}$ on the interior of the strip, then show that f is bounded in magnitude by 1 throughout the strip. (*Hint*: multiply f by $e^{-\varepsilon z^m}$ for some even integer m.) See Section 1.11 for some applications of this principle to interpolation theory.

Example 2.7.9 (Manipulating generalised functions)**.** In PDE one is primarily interested in smooth (classical) solutions; but for a variety of reasons it is useful to also consider rougher solutions. Sometimes, these solutions are so rough that they are no longer functions,

but are measures, distributions (see Section 1.13), or some other concept of "generalised function" or "generalised solution". For instance, the fundamental solution to a PDE is typically just a distribution or measure, rather than a classical function. A typical example: a (sufficiently smooth) solution to the three-dimensional wave equation $-u_{tt} + \Delta u = 0$ with initial position $u(0,x) = 0$ and initial velocity $u_t(0,x) = g(x)$ is given by the classical formula

$$u(t) = tg * \sigma_t$$

for $t > 0$, where σ_t is the unique rotation-invariant probability measure on the sphere $S_t := \{(x,y,z) \in \mathbf{R}^3 : x^2 + y^2 + z^2 = t^2\}$ of radius t, or equivalently, the area element dS on that sphere divided by the surface area $4\pi t^2$ of that sphere. (The convolution $f * \mu$ of a smooth function f and a (compactly supported) finite measure μ is defined by $f * \mu(x) := \int f(x-y)\ d\mu(y)$; one can also use the distributional convolution defined in Section 1.13.)

For this and many other reasons, it is important to manipulate measures and distributions in various ways. For instance, in addition to convolving functions with measures, it is also useful to convolve measures with measures; the convolution $\mu * \nu$ of two finite measures on $\mathbf{R}^n$ is defined as the measure which assigns to each measurable set E in $\mathbf{R}^n$, the measure

$$\mu * \nu(E) := \int \int 1_E(x+y)\ d\mu(x) d\nu(y). \tag{2.25}$$

For sake of concreteness, let's focus on a specific question, namely to compute (or at least estimate) the measure $\sigma * \sigma$, where σ is the normalised rotation-invariant measure on the unit circle $\{x \in \mathbf{R}^2 : |x| = 1\}$. It turns out that while σ is not absolutely continuous with respect to Lebesgue measure m, the convolution is: $d(\sigma * \sigma) = f dm$ for some absolutely integrable function f on $\mathbf{R}^2$. But what is this function f? It certainly is possible to compute it from the definition (2.25), or by other methods (e.g. the Fourier transform), but I would like to give one approach to computing these sorts of expressions involving measures (or other generalised functions) based on epsilon regularisation, which requires a certain amount of geometric computation but which I find to be rather visual and conceptual, compared to more

algebraic approaches (e.g. based on Fourier transforms). The idea is to approximate a singular object, such as the singular measure σ, by a smoother object σ_ε, such as an absolutely continuous measure. For instance, one can approximate σ by

$$d\sigma_\varepsilon := \frac{1}{m(A_\varepsilon)} 1_{A_\varepsilon}\, dm$$

where $A_\varepsilon := \{x \in \mathbf{R}^2 : 1-\varepsilon \leq |x| \leq 1+\varepsilon\}$ is a thin annular neighbourhood of the unit circle. It is clear that σ_ε converges to σ in the *vague topology*, which implies that $\sigma_\varepsilon * \sigma_\varepsilon$ converges to $\sigma * \sigma$ in the vague topology also. Since

$$\sigma_\varepsilon * \sigma_\varepsilon = \frac{1}{m(A_\varepsilon)^2} 1_{A_\varepsilon} * 1_{A_\varepsilon}\, dm,$$

we will be able to understand the limit f by first considering the function

$$f_\varepsilon(x) := \frac{1}{m(A_\varepsilon)^2} 1_{A_\varepsilon} * 1_{A_\varepsilon}(x) = \frac{m(A_\varepsilon \cap (x - A_\varepsilon))}{m(A_\varepsilon)^2}$$

and then taking (weak) limits as $\varepsilon \to 0$ to recover f.

Up to constants, one can compute from elementary geometry that $m(A_\varepsilon)$ is comparable to ε, and $m(A_\varepsilon \cap (x - A_\varepsilon))$ vanishes for $|x| \geq 2+2\varepsilon$, and is comparable to $\varepsilon^2(2-|x|)^{-1/2}$ for $1 \leq |x| \leq 2-2\varepsilon$ (and of size $O(\varepsilon^{3/2})$ in the transition region $|x| = 2+O(\varepsilon)$) and is comparable to $\varepsilon^2|x|^{-1}$ for $\varepsilon \leq |x| \leq 1$ (and of size about $O(\varepsilon)$ when $|x| \leq \varepsilon$. (This is a good exercise for anyone who wants practice in quickly computing the orders of magnitude of geometric quantities such as areas; for such order of magnitude calculations, quick and dirty geometric methods tend to work better here than the more algebraic calculus methods you would have learned as an undergraduate.) The bounds here are strong enough to allow one to take limits and conclude what f looks like: it is comparable to $|x|^{-1}(2-|x|)^{-1/2} 1_{|x|\leq 2}$. And by being more careful with the computations of area, one can compute the exact formula for $f(x)$, though I will not do so here.

Remark 2.7.10. Epsilon regularisation also sheds light on why certain operations on measures or distributions are not permissible. For instance, squaring the Dirac delta function δ will not give a measure or distribution, because if one looks at the squares δ_ε^2 of some smoothed

out approximations δ_ε to the Dirac function (i.e. approximations to the identity), one sees that their masses go to infinity in the limit $\varepsilon \to 0$, and so cannot be integrated against test functions uniformly in ε. On the other hand, derivatives of the delta function, while no longer measures (the total variation of derivatives of δ_ε become unbounded), are at least still distributions (the integrals of derivatives of δ_ε against test functions remain convergent).

Notes. This article first appeared at `terrytao.wordpress.com/2009/02/28`. Thanks to Harald, nicolaennio, and RK for corrections.

The article was a submission to the *Tricki* (`www.tricki.org`), an online repository of mathematical tricks. A version of this article appears on that site at `www.tricki.org/article/Create_an_epsilon_of_room`.

Dima pointed out that a variant of the epsilon regularisation argument is used routinely in real algebraic geometry, when the underlying field $\mathbf{R}$ is extended to the field of real Puiseaux series in a parameter ε. After performing computations in this extension, one eventually sets ε to zero to recover results in the original real field.

2.8. Amenability

Recently, I have been studying the concept of amenability on groups. This concept can be defined in a "combinatorial" or "finitary" fashion, using *Følner sequences*, and also in a more "functional-analytic" or "infinitary" fashion, using invariant means. I wanted to get some practice passing back and forth between these two definitions, so I wrote down some notes on how to do this, and also how to take some facts about amenability that are usually proven in one setting, and prove them instead in the other.

2.8.1. Equivalent definitions of amenability. For simplicity I will restrict attention to countable groups G. Given any $f : G \to \mathbf{R}$ and $x \in G$, I define the left-translation $\tau_x f : G \to \mathbf{R}$ by the formula $\tau_x f(y) := f(x^{-1}y)$. Given $g : G \to \mathbf{R}$ as well, I define the inner product $\langle f, g \rangle := \sum_{x \in G} f(x)g(x)$ whenever the right-hand side is convergent.

All ℓ^p spaces are real-valued. The cardinality of a finite set A is denoted $|A|$. The symmetric difference of two sets A, B is denoted $A\Delta B$.

A *finite mean* is a non-negative, finitely supported function $\mu : G \to \mathbf{R}^+$ such that $\|\mu\|_{\ell^1(G)} = 1$. A *mean* is a non-negative linear functional $\lambda : \ell^\infty(G) \to \mathbf{R}$ such that $\lambda(1) = 1$. Note that every finite mean μ can be viewed as a mean λ_μ by the formula $\lambda_\mu(f) := \langle f, \mu\rangle$.

The following equivalences were established by Følner[**Fo1955**]:

Theorem 2.8.1. *Let G be a countable group. Then the following are equivalent:*

(i) *There exists a left-invariant mean $\lambda : \ell^\infty(G) \to \mathbf{R}$, i.e. mean such that $\lambda(\tau_x f) = \lambda(f)$ for all $f \in \ell^\infty(G)$ and $x \in G$.*

(ii) *For every finite set $S \subset G$ and every $\varepsilon > 0$, there exists a finite mean ν such that $\|\nu - \tau_x\nu\|_{\ell^1(G)} \leq \varepsilon$ for all $x \in S$.*

(iii) *For every finite set $S \subset G$ and every $\varepsilon > 0$, there exists a non-empty finite set $A \subset G$ such that $|(x \cdot A)\Delta A|/|A| \leq \varepsilon$ for all $x \in S$.*

(iv) *There exists a sequence A_n of non-empty finite sets such that $|x \cdot A_n \Delta A_n|/|A_n| \to 0$ as $n \to \infty$ for each $x \in G$. (Such a sequence is called a* Følner sequence*.)*

Proof. We shall use an argument of Namioka[**Na1964**].

(i) implies (ii): Suppose for contradiction that (ii) failed, then there exists S, ε such that $\|\nu - \tau_x\nu\|_{\ell^1(G)} > \varepsilon$ for all means ν and all $x \in S$. The set $\{(\nu - \tau_x\nu)_{x\in S} : \nu \in \ell^1(G)\}$ is then a convex set of $(\ell^1(G))^S$ that is bounded away from zero. Applying the Hahn-Banach separation theorem (Theorem 1.5.14), there thus exists a linear functional $\rho \in (\ell^1(G)^S)^*$ such that $\rho((\nu - \tau_x\nu)_{x\in S}) \geq 1$ for all means ν. Since $(\ell^1(G)^S)^* \equiv \ell^\infty(G)^S$, there thus exist $m_x \in \ell^\infty(G)$ for $x \in S$ such that $\sum_{x\in S}\langle \nu - \delta_x * \nu, m_x\rangle \geq 1$ for all means ν, thus $\langle \nu, \sum_{x\in S} m_x - \tau_{x^{-1}} m_x\rangle \geq 1$. Specialising ν to the Kronecker means δ_y we see that $\sum_{x\in S} m_x - \tau_{x^{-1}} m_x \geq 1$ pointwise. Applying the mean λ, we conclude that $\sum_{x\in S} \lambda(m_x) - \lambda(\tau_{x^{-1}} m_x) \geq 1$. But this contradicts the left-invariance of λ.

(ii) implies (iii): Fix S (which we can take to be non-empty), and let $\varepsilon > 0$ be a small quantity to be chosen later. By (ii) we can find a finite mean ν such that

$$\|\nu - \tau_x \nu\|_{\ell^1(G)} < \varepsilon/|S|$$

for all $x \in S$.

Using the layer-cake decomposition, we can write $\nu = \sum_{i=1}^k c_i 1_{E_i}$ for some nested non-empty sets $E_1 \supset E_2 \supset \ldots \supset E_k$ and some positive constants c_i. As ν is a mean, we have $\sum_{i=1}^k c_i |E_i| = 1$. On the other hand, observe that $|\nu - \tau_x \nu|$ is at least c_i on $(x \cdot E_i) \Delta E_i$. We conclude that

$$\sum_{i=1}^k c_i |(x \cdot E_i) \Delta E_i| \leq \frac{\varepsilon}{|S|} \sum_{i=1}^k c_i |E_i|$$

for all $x \in S$, and thus

$$\sum_{i=1}^k c_i \sum_{x \in S} |(x \cdot E_i) \Delta E_i| \leq \varepsilon \sum_{i=1}^k c_i |E_i|.$$

By the pigeonhole principle, there thus exists i such that

$$\sum_{x \in S} |(x \cdot E_i) \Delta E_i| \leq \varepsilon |E_i|$$

and the claim follows.

(iii) implies (iv): Write the countable group G as the increasing union of finite sets S_n and apply (iii) with $\varepsilon := 1/n$ and $S := S_n$ to create the set A_n.

(iv) implies (i): Use the Hahn-Banach theorem to select an infinite mean $\rho \in \ell^\infty(\mathbf{N})^* \backslash \ell^1(\mathbf{N})$, and define $\lambda(m) = \rho((\langle m, \frac{1}{|A_n|} 1_{A_n} \rangle)_{n \in \mathbf{N}})$. (Alternatively, one can define $\lambda(m)$ to be an *ultralimit* of the $\langle m, \frac{1}{|A_n|} 1_{A_n} \rangle$.) □

Any countable group obeying any (and hence all) of (i)-(iv) is called *amenable*.

Remark 2.8.2. The above equivalences are proven in a non-constructive manner, due to the use of the Hahn-Banach theorem (as well as the contradiction argument). Thus, for instance, it is not immediately

obvious how to convert an invariant mean into a Følner sequence, despite the above equivalences.

2.8.2. Examples of amenable groups. We give some model examples of amenable and non-amenable groups:

Proposition 2.8.3. *Every finite group is amenable.*

Proof. Trivial (either using invariant means or Følner sequences). □

Proposition 2.8.4. *The integers* $\mathbf{Z} = (\mathbf{Z}, +)$ *are are amenable.*

Proof. One can take the sets $A_N = \{1, \ldots, N\}$ as the Følner sequence, or an ultralimit as an invariant mean. □

Proposition 2.8.5. *The free group* F_2 *on two generators* e_1, e_2 *is not amenable.*

Proof. We first argue using invariant means. Suppose for contradiction that one had an invariant mean λ. Let $E_1, E_2, E_{-1}, E_{-2} \subset F_2$ be the set of all words beginning with e_1, e_2, e_1^{-1}, e_2^{-1} respectively. Observe that $E_2 \subset (e_1^{-1} \cdot E_1) \backslash E_1$, thus $\lambda(1_{E_2}) \leq \lambda(\tau_{e_1^{-1}} 1_{E_1}) - \lambda(1_{E_1})$. By invariance we conclude that $\lambda(1_{E_2}) = 0$; similarly for 1_{E_1}, $1_{E_{-1}}$, $1_{E_{-2}}$. Since the identity element clearly must have mean zero, we conclude that the mean λ is identically zero, which is absurd.

Now we argue using Følner sequences. If F_2 were amenable, then for any $\varepsilon > 0$ we could find a finite non-empty set A such that $x \cdot A$ differs from A by at most $\varepsilon|A|$ points for $x = e_1, e_2, e_1^{-1}, e_2^{-1}$. The set $e_1 \cdot (A \cap (E_2 \cup E_{-1} \cup E_{-2}))$ is contained in $e_1 \cdot A$ and in E_1, and so

$$|e_1 \cdot (A \backslash E_{-1})| \leq |A \cap E_1| + \varepsilon |A|,$$

and thus

$$|A| - |A \cap E_{-1}| \leq |A \cap E_1| + \varepsilon |A|.$$

Similarly for permutations. Summing up over all four permutations, we obtain

$$4|A| - |A| \leq |A| + 4\varepsilon |A|,$$

leading to a contradiction for ε small enough (any $\varepsilon < 1/2$ will do). □

Remark 2.8.6. The non-amenability of the free group is related to the *Banach-Tarski paradox* (see Section 2.2).

Now we generate some more amenable groups.

Proposition 2.8.7. *Let* $0 \to H \to G \to K \to 0$ *be a* short exact sequence *of countable groups (thus* H *can be identified with a normal subgroup of* G*, and* K *can be identified with* G/H*). If* H *and* K *are amenable, then* G *is amenable also.*

Proof. Using invariant means, there is a very short proof: given invariant means λ_H, λ_K for H, K, we can build an invariant mean λ_G for G by the formula

$$\lambda_G(f) := \lambda_K(F)$$

for any $f \in \ell^\infty(G)$, where $F : K \to \mathbf{R}$ is the function defined as $F(xH) := \lambda_H(f(x\cdot))$ for all cosets xH (note that the left-invariance of λ_H shows that the exact choice of coset representative x is irrelevant). (One can view λ_G as sort of a "product measure" of the λ_H and λ_K.)

Now we argue using Følner sequences instead. Let E_n, F_n be Følner sequences for H, K respectively. Let S be a finite subset of G, and let $\varepsilon > 0$. We would like to find a finite non-empty subset $A \subset G$ such that $|(x\cdot A)\backslash A| \leq \varepsilon|A|$ for all $x \in S$; this will demonstrate amenability. (Note that by taking S to be symmetric, we can replace $|(x \cdot A)\backslash A|$ with $|(x \cdot A)\Delta A|$ without difficulty.)

By taking n large enough, we can find F_n such that $\pi(x) \cdot F_n$ differs from F_n by at most $\varepsilon|F_n|/2$ elements for all $x \in S$, where $\pi : G \to K$ is the projection map. Now, let F'_n be a preimage of F_n in G. Let T be the set of all group elements $t \in K$ such that $S \cdot F'_n$ intersects $F'_n \cdot t$. Observe that T is finite. Thus, by taking m large enough, we can find E_m such that $t \cdot E_m$ differs from E_m by at most $\varepsilon|E_m|/2|T|$ points for all $t \in T$.

Now set $A := F'_n \cdot E_m = \{zy : y \in E_m, z \in F'_n\}$. Observe that the sets $z \cdot E_m$ for $z \in F'_n$ lie in disjoint cosets of H and so $|A| = |E_m||F'_n| = |E_m||F_n|$. Now take $x \in S$, and consider an element of $(x \cdot A)\backslash A$. This element must take the form xzy for some $y \in E_m$ and $z \in F'_n$. The coset of H that xzy lies in is given by $\pi(x)\pi(z)$. Suppose first that $\pi(x)\pi(z)$ lies outside of F_n. By construction, this occurs for

at most $\varepsilon|F_n|/2$ choices of z, leading to at most $\varepsilon|E_m||F_n|/2 = \varepsilon|A|/2$ elements in $(x \cdot A)\backslash A$.

Now suppose instead that $\pi(x)\pi(z)$ lies in F_n. Then we have $xz = z't$ for some $z' \in F'_n$ and $t \in T$, by construction of T, and so $xzy = z'ty$. But as xzy lies outside of A, ty must lie outside of E_m. But by construction of E_m, there are at most $\varepsilon|E_m|/2|T|$ possible choices of y that do this for each fixed x, t, leading to at most $\varepsilon|E_m||F_n|/2 = \varepsilon|A|/2$. We thus have $|(x \cdot A)\backslash A| \leq \varepsilon|A|$ as required. □

Proposition 2.8.8. *Let $G_1 \subset G_2 \subset \ldots$ be a sequence of countable amenable groups. Then $G := \bigcup_n G_n$ is amenable.*

Proof. We first use invariant means. An invariant mean on $\ell^\infty(G_n)$ induces a mean on $\ell^\infty(G)$ which is invariant with respect to translations by G_n. Taking an ultralimit of these means, we obtain the claim.

Now we use Følner sequences. Given any finite set $S \subset G$ and $\varepsilon > 0$, we have $S \subset G_n$ for some n. As G_n is amenable, we can find $A \subset G_n$ such that $|(x \cdot A)\Delta A| \leq \varepsilon|A|$ for all $x \in S$, and the claim follows. □

Proposition 2.8.9. *Every countable* virtually solvable *group G is amenable.*

Proof. By definition, every virtually solvable group contains a solvable group of finite index, and thus contains a normal solvable subgroup of finite index. (Note that every subgroup H of G of index I contains a normal subgroup of index at most $I!$, namely the stabiliser of the G action on G/H.) By Proposition 2.8.7 and Proposition 2.8.3, we may thus reduce to the case when G is solvable. By inducting on the derived length of this solvable group using Proposition 2.8.7 again, it suffices to verify this when the group is abelian. By Proposition 2.8.8, it suffices to verify this when the group is abelian and finitely generated. By Proposition 2.8.7 again, it suffices to verify this when the group is cyclic. But this follows from Proposition 2.8.3 and Proposition 2.8.4. □

Notes. This article first appeared at `terrytao.wordpress.com/2009/04/14`. Thanks to Orr for corrections.

Danny Calegari noted the application of amenability to that of obtaining asymptotic limit objects in dynamics (e.g. via the ergodic theorem for amenable groups). Jason Behrstock mentioned an amusing characterisation of amenability, as those groups which do not admit successful "Ponzi schemes" - schemes in which each group element passes on a bounded amount of money to its neighbours (in a Cayley graph) in such a way that everyone profits. There was some ensuing discussion as to the related question of whether amenable and non-amenable groups admit nontrivial bounded harmonic functions.

Chapter 3

Expository articles

3.1. An explicitly solvable nonlinear wave equation

As is well known, the linear one-dimensional wave equation

$$-\phi_{tt} + \phi_{xx} = 0, \tag{3.1}$$

where $\phi : \mathbf{R} \times \mathbf{R} \to \mathbf{R}$ is the unknown field (which, for simplicity, we assume to be smooth), can be solved explicitly; indeed, the general solution to (3.1) takes the form

$$\phi(t,x) = f(t+x) + g(t-x) \tag{3.2}$$

for some arbitrary (smooth) functions $f, g : \mathbf{R} \to \mathbf{R}$. (One can of course determine f and g once one specifies enough initial data or other boundary conditions, but this is not the focus of my post today.)

When one moves from linear wave equations to nonlinear wave equations, then in general one does not expect to have a closed-form solution such as (3.2). So I was pleasantly surprised recently while playing with the nonlinear wave equation

$$-\phi_{tt} + \phi_{xx} = e^{\phi}, \tag{3.3}$$

to discover that this equation can also be explicitly solved in closed form. (For the reason why I was interested in this equation, see [**Ta2010**].)

A posteriori, I now know the reason for this explicit solvability; (3.3) is the limiting case $a = 0, b \to -\infty$ of the more general equation

$$-\phi_{tt} + \phi_{xx} = e^{\phi+a} - e^{-\phi+b}$$

which (after applying the simple transformation $\phi = \frac{b-a}{2} + \psi(\sqrt{2}e^{\frac{a+b}{4}}t, \sqrt{2}e^{\frac{a+b}{4}}x)$) becomes the *sinh-Gordon equation*

$$-\psi_{tt} + \psi_{xx} = \sinh(\psi)$$

(a close cousin of the more famous *sine-Gordon equation* $-\phi_{tt} + \phi_{xx} = \sin(\phi)$), which is known to be completely integrable, and exactly solvable. However, I only realised this after the fact, and stumbled upon the explicit solution to (3.3) by much more classical and elementary means. I thought I might share the computations here, as I found them somewhat cute, and seem to serve as an example of how one

might go about finding explicit solutions to PDE in general; accordingly, I will take a rather pedestrian approach to describing the hunt for the solution, rather than presenting the shortest or slickest route to the answer.

After the initial publishing of this post, Patrick Dorey pointed out to me that (3.3) is extremely classical; it is known as Liouville's equation and was solved by Liouville [**Li1853**], with essentially the same solution as presented here.

3.1.1. Symmetries. To simplify the discussion let us ignore all issues of regularity, division by zero, taking square roots and logarithms of negative numbers, etc., and proceed for now in a purely formal fashion, pretending that all functions are smooth and lie in the domain of whatever algebraic operations are being performed. (It is not too difficult to go back after the fact and justify these formal computations, but I do not wish to focus on that aspect of the problem here.)

Although not strictly necessary for solving the equation (3.3), I find it convenient to bear in mind the various symmetries that (3.3) enjoys, as this provides a useful "reality check" to guard against errors (e.g. arriving at a class of solutions which is not invariant under the symmetries of the original equation). These symmetries are also useful to normalise various special families of solutions.

One easily sees that solutions to (3.3) are invariant under spacetime translations

$$\phi(t,x) \mapsto \phi(t-t_0, x-x_0) \tag{3.4}$$

and also spacetime reflections

$$\phi(t,x) \mapsto \phi(\pm t, \pm x). \tag{3.5}$$

Being relativistic, the equation is also invariant under Lorentz transformations

$$\phi(t,x) \mapsto \phi(\frac{t-vx}{\sqrt{1-v^2}}, \frac{x-vt}{\sqrt{1-v^2}}). \tag{3.6}$$

Finally, one has the scaling symmetry

$$\phi(t,x) \mapsto \phi(\lambda t, \lambda x) + 2\log\lambda. \tag{3.7}$$

3.1.2. Solution. Henceforth ϕ will be a solution to (3.3). In view of the linear explicit solution (3.2), it is natural to move to null coordinates

$$u = t + x, v = t - x,$$

thus

$$\partial_u = \frac{1}{2}(\partial_t + \partial_x); \partial_v = \frac{1}{2}(\partial_t - \partial_x)$$

and (3.3) becomes

$$\phi_{uv} = -\frac{1}{4}e^{\phi}. \tag{3.8}$$

The various symmetries (3.4)-(3.7) can of course be rephrased in terms of null coordinates in a straightforward manner. The Lorentz symmetry (3.6) simplifies particularly nicely in null coordinates, to

$$\phi(u, v) \mapsto \phi(\lambda u, \lambda^{-1} v). \tag{3.9}$$

Motivated by the general theory of stress-energy tensors of relativistic wave equations (of which (3.3) is a very simple example), we now look at the null energy densities ϕ_u^2, ϕ_v^2. For the linear wave equation (3.1) (or equivalently $\phi_{uv} = 0$), these null energy densities are transported in null directions:

$$\partial_v \phi_u^2 = 0; \partial_u \phi_v^2 = 0. \tag{3.10}$$

(One can also see this from the explicit solution (3.2).)

The above transport law isn't quite true for the nonlinear wave equation, of course, but we can hope to get some usable substitute. Let us just look at the first null energy ϕ_u^2 for now. By two applications of (3.10), this density obeys the transport equation

$$\begin{aligned} \partial_v \phi_u^2 &= 2\phi_u \phi_{uv} \\ &= -\frac{1}{2}\phi_u e^{\phi} \\ &= -\frac{1}{2}\partial_u(e^{\phi}) \\ &= 2\partial_u \phi_{uv} \\ &= \partial_v(2\phi_{uu}) \end{aligned}$$

and thus we have the pointwise conservation law

$$\partial_v(\phi_u^2 - 2\phi_{uu}) = 0$$

which implies that

$$-\frac{1}{2}\phi_{uu} + \frac{1}{4}\phi_u^2 = U(u) \tag{3.11}$$

for some function $U : \mathbf{R} \to \mathbf{R}$ depending only on u. Similarly we have

$$-\frac{1}{2}\phi_{vv} + \frac{1}{4}\phi_v^2 = V(v)$$

for some function $V : \mathbf{R} \to \mathbf{R}$ depending only on v.

For any fixed v, (11) is a nonlinear ODE in u. To solve it, we can first look at the homogeneous ODE

$$-\frac{1}{2}\phi_{uu} + \frac{1}{4}\phi_u^2 = 0. \tag{3.12}$$

Undergraduate ODE methods (e.g. separation of variables, after substituting $\psi := \phi_u$) soon reveal that the general solution to this ODE is given by $\phi(u) = -2\log(u + C) + D$ for arbitrary constants C, D (ignoring the issue of singularities or degeneracies for now). Equivalently, (3.12) is obeyed if and only if $e^{-\phi/2}$ is linear in u. Motivated by this, we become tempted to rewrite (3.11) in terms of $\Phi := e^{-\phi/2}$. One soon realises that

$$\partial_{uu}\Phi = (-\frac{1}{2}\phi_{uu} + \frac{1}{4}\phi_u^2)\Phi$$

and hence (3.11) becomes

$$(-\partial_{uu} + U(u))\Phi = 0, \tag{3.13}$$

thus Φ is a null (generalised) eigenfunction of the Schrodinger operator (or Hill operator) $-\partial_{uu} + U(u)$. If we let $a(u)$ and $b(u)$ be two linearly independent solutions to the ODE

$$-f_{uu} + Uf = 0, \tag{3.14}$$

we thus have

$$\Phi = a(u)c(v) + b(u)d(v) \tag{3.15}$$

for some functions c, d (which one easily verifies to be smooth, since ϕ, a, b are smooth and a, b are linearly independent). Meanwhile, by playing around with the second null energy density we have the counterpart to (3.14),

$$(-\partial_{vv} + V(v))\Phi = 0,$$

and hence (by linear independence of a, b) c, d must be solutions to the ODE

$$-g_{vv} + Vg = 0.$$

This would be a good time to pause and see whether our implications are reversible, i.e. whether any ϕ that obeys the relation (3.15) will solve (3.3) or (3.10). It is of course natural to first write (3.10) in terms of Φ. Since

$$\Phi_u = -\frac{1}{2}\phi_u\Phi; \Phi_v = -\frac{1}{2}\phi_v\Phi; \Phi_{uv} = (\frac{1}{4}\phi_u\phi_v - \frac{1}{2}\phi_{uv})\Phi$$

one soon sees that (3.10) is equivalent to

$$\Phi\Phi_{uv} = \Phi_u\Phi_v + \frac{1}{8}. \tag{3.16}$$

If we then insert the ansatz (3.15), we soon reformulate the above equation as

$$(a(u)b'(u) - b(u)a'(u))(c(v)d'(v) - d(v)c'(v)) = \frac{1}{8}.$$

It is at this time that one should remember the classical fact that if a, u are two solutions to the ODE (3.11), then the *Wronskian* $ab' - ba'$ is constant; similarly $cd' - dc'$ is constant. Putting this all together, we see that

Theorem 3.1.1. *A smooth function ϕ solves* (3.3) *if and only if we have the relation* (3.13) *for some functions a, b, c, d obeying the Wronskian conditions $ab' - ba' = \alpha$, $cd' - dc' = \beta$ for some constants α, β multiplying to $\frac{1}{8}$.*

Note that one can generate solutions to the Wronskian equation $ab' - ba' = \alpha$ by a variety of means, for instance by first choosing a arbitrarily and then rewriting the equation as $(b/a)' = \alpha/a^2$ to recover b. (This doesn't quite work at the locations when a vanishes, but there are a variety of ways to resolve that; as I said above, we are ignoring this issue for the purposes of this discussion.)

This is not the only way to express solutions. Factoring $a(u)d(v)$ (say) from (3.13), we see that Φ is the product of a solution $c(v)/d(v) + b(u)/a(u)$ to the linear wave equation, plus the exponential of a solution $\log a(u) + \log d(u)$ to the linear wave equation. Thus we may

write $\phi = F - 2\log G$, where F and G solve the linear wave equation. Inserting this back ansatz into (3.1) we obtain

$$2(-G_t^2 + G_x^2)/G^2 = e^F/G^2$$

and so we see that

$$\phi = \log \frac{2(-G_t^2 + G_x^2)}{G^2} = \log \frac{-8G_u G_v}{G^2} \tag{3.17}$$

for some solution G to the free wave equation, and conversely every expression of the form (3.17) can be verified to solve (3.1) (since $\log 2(-G_t^2 + G_x^2)$ does indeed solve the free wave equation, thanks to (3.2)). Inserting (3.2) into (3.17) we thus obtain the explicit solution

$$\phi = \log \frac{-8f'(t+x)g'(t-x)}{(f(t+x) + g(t-x))^2} \tag{3.18}$$

to (3.1), where f and g are arbitrary functions (recall that we are neglecting issues such as whether the quotient and the logarithm are well-defined).

I, for one, would not have expected the solution to take this form. But it is instructive to check that (3.18) does at least respect all the symmetries (3.4)-(3.7).

3.1.3. Some special solutions. If we set $U = V = 0$, then a, b, c, d are linear functions, and so Φ is affine-linear in u, v. One also checks that the uv term in Φ cannot vanish. After translating in u and v, we end up with the ansatz $\Phi(u, v) = c_1 + c_2 uv$ for some constants c_1, c_2; applying (3.16) we see that $c_1 c_2 = 1/8$, and by using the scaling symmetry (3.7) we may normalise e.g. $c_1 = 8, c_2 = 1$, and so we arrive at the (singular) solution

$$\phi = -2\log(8 + uv) = \log \frac{1}{(8 + t^2 - x^2)^2}. \tag{3.19}$$

To express this solution in the form (3.18), one can take $f(u) = \frac{8}{u}$ and $g(v) = v$; some other choices of f, g are also possible. (Determining the extent to which f, g are uniquely determined by ϕ in general can be established from a closer inspection of the previous arguments, and is left as an exercise.)

We can also look at what happens when ϕ is constant in space, i.e. it solves the ODE $-\phi_{tt} = e^\phi$. It is not hard to see that U and

V must be constant in this case, leading to a, b, c, d which are either trigonometric or exponential functions. This soon leads to the ansatz $\Phi = c_1 e^{\alpha t} + c_2 e^{-\alpha t}$ for some (possibly complex) constants c_1, c_2, α, thus $\phi = -2\log(c_1 e^{\alpha t} + c_2 e^{-\alpha t})$. By using the symmetries (3.4), (3.7) we can make $c_1 = c_2$ and specify α to be whatever we please, thus leading to the solutions $\phi = -2\log\cosh \alpha t + c_3$. Applying (3.1) we see that this is a solution as long as $e^{c_3} = 2\alpha^2$. For instance, we may fix $c_3 = 0$ and $\alpha = 1/\sqrt{2}$, leading to the solution

$$\phi = -2\log\cosh \frac{t}{\sqrt{2}}. \tag{3.20}$$

To express this solution in the form (3.18), one can take for instance $f(u) = e^{u/\sqrt{2}}$ and $g(v) = e^{-v/\sqrt{2}}$.

One can of course push around (3.19), (3.20) by the symmetries (3.4)-(3.7) to generate a few more special solutions.

Notes. This article first appeared at `terrytao.wordpress.com/2009/01/22`. Thanks to Jake K. for corrections.

There was some interesting discussion online regarding whether the heat equation had a natural relativistic counterpart, and more generally whether it was profitable to study non-relativistic equations via relativistic approximations.

3.2. Infinite fields, finite fields, and the Ax-Grothendieck theorem

Jean-Pierre Serre (whose papers are, of course, always worth reading) recently wrote a lovely article[**Se2009**] in which he describes several ways in which algebraic statements over fields of zero characteristic, such as $\mathbf{C}$, can be deduced from their positive characteristic counterparts such as F_{p^m}, despite the fact that there is no non-trivial field homomorphism between the two types of fields. In particular finitary tools, including such basic concepts as cardinality, can now be deployed to establish infinitary results. This leads to some simple and elegant proofs of non-trivial algebraic results which are not easy to establish by other means.

One deduction of this type is based on the idea that positive characteristic fields can partially *model* zero characteristic fields, and proceeds like this: if a certain algebraic statement failed over (say) $\mathbf{C}$, then there should be a "finitary algebraic" obstruction that "witnesses" this failure over $\mathbf{C}$. Because this obstruction is both finitary and algebraic, it must also be definable in some (large) finite characteristic, thus leading to a comparable failure over a finite characteristic field. Taking contrapositives, one obtains the claim.

Algebra is definitely not my own field of expertise, but it is interesting to note that similar themes have also come up in my own area of additive combinatorics (and more generally arithmetic combinatorics), because the combinatorics of addition and multiplication on finite sets is definitely of a "finitary algebraic" nature. For instance, a recent paper of Vu, Wood, and Wood[**VuWoWo2010**] establishes a finitary "Freiman-type" homomorphism from (finite subsets of) the complex numbers to large finite fields that allows them to pull back many results in arithmetic combinatorics in finite fields (e.g. the sum-product theorem) to the complex plane. Van Vu and I also used a similar trick in [**TaVu2007**] to control the singularity property of random sign matrices by first mapping them into finite fields in which cardinality arguments became available.) And I have a particular fondness for correspondences between finitary and infinitary mathematics; the correspondence Serre discusses is slightly different from the one I discuss for instance in Section 1.3 of *Structure and Randomness*, although there seems to be a common theme of "compactness" (or of model theory) tying these correspondences together.

As one of his examples, Serre cites one of my own favourite results in algebra, discovered independently by Ax[**Ax1968**] and by Grothendieck[**Gr1966**] (and then rediscovered many times since). Here is a special case of that theorem:

Theorem 3.2.1 (Ax-Grothendieck theorem, special case)**.** *Let $P : \mathbf{C}^n \to \mathbf{C}^n$ be a polynomial map from a complex vector space to itself. If P is injective, then P is bijective.*

The full version of the theorem allows one to replace $\mathbf{C}^n$ by an algebraic variety X over any algebraically closed field, and for P to be an morphism from the algebraic variety X to itself, but for simplicity

I will just discuss the above special case. This theorem is not at all obvious; it is not too difficult (see Lemma 3.2.6 below) to show that the *Jacobian* of P is non-degenerate, but this does not come close to solving the problem since one would then be faced with the notorious *Jacobian conjecture*. Also, the claim fails if "polynomial" is replaced by "holomorphic", due to the existence of *Fatou-Bieberbach domains*.

In this post I would like to give the proof of Theorem 3.2.1 based on finite fields as mentioned by Serre, as well as another elegant proof of Rudin[**Ru1995**] that combines algebra with some elementary complex variable methods. (There are several other proofs of this theorem and its generalisations, for instance a topological proof by Borel[**Bo1969**], which I will not discuss here.)

3.2.1. Proof via finite fields. The first observation is that the theorem is utterly trivial in the finite field case:

Theorem 3.2.2 (Ax-Grothendieck theorem in F)**.** *Let F be a finite field, and let $P : F^n \to F^n$ be a polynomial. If P is injective, then P is bijective.*

Proof. Any injection from a finite set to itself is necessarily bijective. (The hypothesis that P is a polynomial is not needed at this stage, but becomes crucial later on.) □

Next, we pass from a finite field F to its algebraic closure $\overline{F}$.

Theorem 3.2.3 (Ax-Grothendieck theorem in $\overline{F}$)**.** *Let F be a finite field, let $\overline{F}$ be its algebraic closure, and let $P : \overline{F}^n \to \overline{F}^n$ be a polynomial. If P is injective, then P is bijective.*

Proof. Our main tool here is *Hilbert's nullstellensatz*, which we interpret here as an assertion that if an algebraic problem is insoluble, then there exists a finitary algebraic obstruction that witnesses this lack of solution (see also Section 1.15 of *Structure and Randomness*). Specifically, suppose for contradiction that we can find a polynomial $P : \overline{F}^n \to \overline{F}^n$ which is injective but not surjective. Injectivity of P means that the algebraic system

$$P(x) = P(y); \quad x \neq y$$

has no solution over the algebraically closed field $\overline{F}$; by the nullstellensatz, this implies that there must exist an algebraic identity of the form

$$(P(x) - P(y)) \cdot Q(x,y) = (x-y)^r \tag{3.21}$$

for some $r \geq 1$ and some polynomial $Q : \overline{F}^n \times \overline{F}^n \to \overline{F}^n$ that specifically witnesses this lack of solvability. Similarly, lack of surjectivity means the existence of an $z_0 \in \overline{F}^n$ such that the algebraic system

$$P(x) = z_0$$

has no solution over the algebraically closed field $\overline{F}$; by another application of the nullstellensatz, there must exist an algebraic identity of the form

$$(P(x) - z_0) \cdot R(x) = 1 \tag{3.22}$$

for some polynomial $R : \overline{F}^n \to \overline{F}^n$ that specifically witnesses this lack of solvability.

Fix Q, z_0, R as above, and let k be the subfield of $\overline{F}$ generated by F and the coefficients of P, Q, z_0, R. Then we observe (thanks to our explicit witnesses (3.21), (3.22)) that the counterexample P descends from $\overline{F}$ to k; P is a polynomial from k^n to k^n which is injective but not surjective.

But k is finitely generated, and every element of k is algebraic over the finite field F, thus k is finite. But this contradicts Theorem 3.2.2. □

Remark 3.2.4. As pointed out to me by L. Spice, there is a simpler proof of Theorem 3.2.3 that avoids the nullstellensatz: one observes from Theorem 3.2.2 that P is bijective over any finite extension of F that contains all of the coefficients of P, and the claim then follows by taking limits.

The complex case $\mathbf{C}$ follows by a slight extension of the argument used to prove Theorem 3.2.3. Indeed, suppose for contradiction that there is a polynomial $P : \mathbf{C}^n \to \mathbf{C}^n$ which is injective but not surjective. As $\mathbf{C}$ is algebraically closed (the *fundamental theorem of algebra*), we may invoke the nullstellensatz as before and find witnesses (3.21), (3.22) for some Q, z_0, R.

Now let $k = Q[\mathcal{C}]$ be the subfield of $\mathbf{C}$ generated by the rationals $\mathbf{Q}$ and the coefficients $\mathcal{C}$ of P, Q, z_0, R. Then we can descend the counterexample to k. This time, k is not finite, but we can descend it to a finite field (and obtain the desired contradiction) by a number of methods. One approach, which is the one taken by Serre, is to quotient the ring $\mathbf{Z}[\mathcal{C}]$ generated by the above coefficients by a maximal ideal, observing that this quotient is necessarily a finite field. Another is to use a general mapping theorem of Vu, Wood, and Wood[**VuWoWo2010**]. We sketch the latter approach as follows. Being finitely generated, we know that k has a finite *transcendence basis* $\alpha_1, \ldots, \alpha_m$ over $\mathbf{Q}$. Applying the *primitive element theorem*, we can then express k as the finite extension of $\mathbf{Q}[\alpha_1, \ldots, \alpha_m]$ by an element β which is algebraic over $\mathbf{Q}[\alpha_1, \ldots, \alpha_m]$; all the coefficients $\mathcal{C}$ are thus rational combinations of $\alpha_1, \ldots, \alpha_m, \beta$. By rationalising, we can ensure that the denominators of the expressions of these coefficients are integers in $\mathbf{Z}[\alpha_1, \ldots, \alpha_m]$; dividing β by an appropriate power of the product of these denominators we may assume that the coefficients in $\mathcal{C}$ all lie in the commutative ring $\mathbf{Z}[\alpha_1, \ldots, \alpha_m, \beta]$, which can be identified with the commutative ring $\mathbf{Z}[a_1, \ldots, a_m, b]$ generated by formal indeterminates $a_1, \ldots, a_m, b$, quotiented by the ideal generated by the minimal polynomial $f \in \mathbf{Z}[a_1, \ldots, a_m, b]$ of β; the algebraic identities (3.21), (3.22) then transfer to this ring. Now pick a large prime p, and map $a_1, \ldots, a_m$ to random elements of F_p. With high probability, the image of f (which is now in $F_p[b]$) is non-degenerate; we can then map b to a root of this image in a finite extension of F_p. (In fact, by using the *Chebotarev density theorem* (or Frobenius density theorem), we can place b back in F_p for infinitely many primes p.) This descends the identities (3.21), (3.22) to this finite extension, as desired.

Remark 3.2.5. This argument can be generalised substantially; it can be used to show that any first-order sentence in the language of fields is true in all algebraically closed fields of characteristic zero if and only if it is true for all algebraically closed fields of sufficiently large characteristic. This result can be deduced from the famous result (proved by Tarski[**Ta1951**], and independently, in an equivalent formulation, by Chevalley) that the theory of algebraically

closed fields (in the language of rings) admits elimination of quantifiers. See for instance [**PCM**, Section IV.23.4]. There are also analogues for real closed fields, starting with the paper of Bialynicki-Birula and Rosenlicht[**BiRo1962**], with a general result established by Kurdyka[**Ku1999**]. Ax-Grothendieck type properties in other categories have been studied by Gromov[**Gr1999**], who calls this property "surjunctivity".

3.2.2. Rudin's proof. Now we give Rudin's proof, which does not use the nullstellensatz, instead relying on some Galois theory and the topological structure of $\mathbf{C}$. We first need a basic fact:

Lemma 3.2.6. *Let $\Omega \subset \mathbf{C}^n$ be an open set, and let $f : \Omega \to \mathbf{C}^n$ be an injective holomorphic map. Then the Jacobian of f is non-degenerate, i.e. $\det Df(z) \neq 0$ for all $z \in \Omega$.*

Actually, we only need the special case of this lemma when f is a polynomial.

Proof. We use an argument of Rosay[**Ro1982**]. For $n = 1$ the claim follows from Taylor expansion. Now suppose $n > 1$ and the claim is proven for $n - 1$. Suppose for contradiction that $\det Df(z_0) = 0$ for some $z_0 \in \Omega$. We claim that $Df(z_0)$ in fact vanishes entirely. If not, then we can find $1 \leq i, j \leq n$ such that $\frac{\partial}{\partial z_j} f_i(z_0) \neq 0$; by permuting we may take $i = j = 1$. We can also normalise $z_0 = f(z_0) = 0$. Then the map $h : z \mapsto (f_1(z), z_2, \ldots, z_n)$ is holomorphic with non-degenerate Jacobian at 0 and is thus locally invertible at 0. The map $f \circ h^{-1}$ is then holomorphic at 0 and preserves the z_1 coordinate, and thus descends to an injective holomorphic map on a neighbourhood of the origin $\mathbf{C}^{n-1}$, and so its Jacobian is non-degenerate by induction hypothesis, a contradiction.

We have just shown that the gradient of f vanishes on the zero set $\{\det Df = 0\}$, which is an analytic variety of codimension 1 (if f is polynomial, it is of course an algebraic variety). Thus f is locally constant on this variety, which contradicts injectivity and we are done. □

From this lemma and the inverse function theorem we have

Corollary 3.2.7. *Injective holomorphic maps from* $\mathbf{C}^n$ *to* $\mathbf{C}^n$ *are* open *(i.e. they map open sets to open sets).*

Now we can give Rudin's proof. Let $P : \mathbf{C}^n \to \mathbf{C}^n$ be an injective polynomial. We let k be the field generated by $\mathbf{Q}$ and the coefficients of P; thus P is definable over k. Let $k[z] = k[z_1, \ldots, z_n]$ be the extension of k by n indeterminates $z_1, \ldots, z_n$. Inside $k[z]$ we have the subfield $k[P(z)]$ generated by k and the components of $P(z)$.

We claim that $k[P(z)]$ is all of $k[z]$. For if this were not the case, we see from Galois theory that there is a non-trivial automorphism $\phi : k[z] \to k[z]$ that fixes $k[P(z)]$; in particular, there exists a non-trivial rational (over k) combination $Q(z)/R(z)$ of z such that $P(Q(z)/R(z)) = P(z)$. Now map z to a random complex number in $\mathbf{C}$, which will almost surely be transcendental over the countable field k; this explicitly demonstrates non-injectivity of P, a contradiction.

Since $k[P(z)] = k[z]$, there exists a rational function $Q_j(z)/R_j(z)$ over k for each $j = 1, \ldots, n$ such that $z_j = Q_j(P(z))/R_j(P(z))$. We may of course assume that Q_j, R_j have no common factors.

We have the polynomial identity $Q_j(P(z)) = z_j R_j(P(z))$. In particular, this implies that on the domain $P(\mathbf{C}^n) \subset \mathbf{C}^n$ (which is open by Corollary 3.2.7), the zero set of R_j is contained in the zero set of Q_j. But as Q_j and R_j have no common factors, this is impossible by elementary algebraic geometry; thus R_j is non-vanishing on $P(\mathbf{C}^n)$. Thus the polynomial $R_j \cdot P$ has no zeroes and is thus constant; we may then normalise so that $R_j \cdot P = 1$. Thus we now have $z = Q(P(z))$ for some polynomial Q, which implies that $w = P(Q(w))$ for all w in the open set $P(\mathbf{C}^n)$. But w and $P(Q(w))$ are both polynomials, and thus must agree on all of $\mathbf{C}^n$. Thus P is bijective as required.

Remark 3.2.8. Note that Rudin's proof gives the stronger statement that if a polynomial map from $\mathbf{C}^n$ to $\mathbf{C}^n$ is injective, then it is bijective and its inverse is also a polynomial.

Notes. This article first appeared at `terrytao.wordpress.com/2009/03/07`. Thanks to fdreher and Ricardo Menares for corrections.

Ricardo Menares and Terry Hughes also mentioned some alternate proofs and generalisations of the Ax-Grothendieck theorem.

3.3. Sailing into the wind, or faster than the wind

One of the more unintuitive facts about sailing is that it is possible to harness the power of the wind to sail in a direction against that of the wind or to sail with a speed faster than the wind itself, even when the water itself is calm. It is somewhat less known, but nevertheless true, that one can (in principle) do both at the same time - sail against the wind (even directly against the wind!) at speeds faster than the wind. This does not contradict any laws of physics, such as conservation of momentum or energy (basically because the reservoir of momentum and energy in the wind far outweighs the portion that will be transmitted to the sailboat), but it is certainly not obvious at first sight how it is to be done.

The key is to exploit all three dimensions of space when sailing. The most obvious dimension to exploit is the *windward/leeward* dimension - the direction that the wind velocity v_0 is oriented in. But if this is the only dimension one exploits, one can only sail up to the wind speed $|v_0|$ and no faster, and it is not possible to sail in the direction opposite to the wind.

Things get more interesting when one also exploits the *crosswind* dimension perpendicular to the wind velocity, in particular by tacking the sail. If one does this, then (in principle) it becomes possible to travel up to double the speed $|v_0|$ of wind, as we shall see below.

However, one still cannot sail against to the wind purely by tacking the sail. To do this, one needs to not just harness the power of the wind, but also that of the water *beneath* the sailboat, thus exploiting (barely) the third available dimension. By combining the use of a sail in the air with the use of sails in the water - better known as *keels*, *rudders*, and *hydrofoils* - one can now sail in certain directions against the wind, and at certain speeds. In most sailboats, one relies primarily on the keel, which lets one sail against the wind but not directly opposite it. But if one tacks the rudder or other hydrofoils as well as the sail, then in fact one can (in principle) sail in arbitrary directions (including those directly opposite to v_0), and in arbitrary speeds (even those much larger than $|v_0|$), although it is quite difficult

to actually achieve this in practice. It may seem odd that the water, which we are assuming to be calm (i.e. traveling at zero velocity) can be used to increase the range of available velocities and speeds for the sailboat, but we shall see shortly why this is the case.

If one makes several simplifying and idealised (and, admittedly, rather unrealistic in practice) assumptions in the underlying physics, then sailing can in fact be analysed by a simple two-dimensional geometric model which explains all of the above statements. In this post, I would like to describe this mathematical model and how it gives the conclusions stated above.

3.3.1. One-dimensional sailing. Let us first begin with the simplest case of one-dimensional sailing, in which the sailboat lies in a one-dimensional universe (which we describe mathematically by the real line $\mathbf{R}$). To begin with, we will ignore the friction effects of the water (one might imagine sailing on an iceboat rather than a sailing boat). We assume that the air is blowing at a constant velocity $v_0 \in \mathbf{R}$, which for sake of discussion we shall take to be positive. We also assume that one can do precisely two things with a sailboat: one can either *furl* the sail, in which case the wind does not propel the sailboat at all, or one can *unfurl* the sail, in order to exploit the force of the wind.

When the sail is furled, then (ignoring friction), the velocity v of the boat stays constant, as per *Newton's first law*. When instead the sail is unfurled, the motion is instead governed by *Newton's second law*, which among other things asserts that the velocity v of the boat will be altered in the direction of the net force exerted by the sail. This net force (which, in one dimension, is purely a *drag force*) is determined not by the true wind speed v_0 as measured by an observer at rest, but by the *apparent* wind speed $v_0 - v$ as experienced by the boat, as per the (Galilean) *principle of relativity*. (Indeed, Galileo himself supported this principle with a famous thought-experiment on a ship.) Thus, the sail can increase the velocity v when $v_0 - v$ is positive, and decrease it when $v_0 - v$ is negative. We can illustrate the effect of an unfurled sail by a vector field in velocity space (Figure 1).

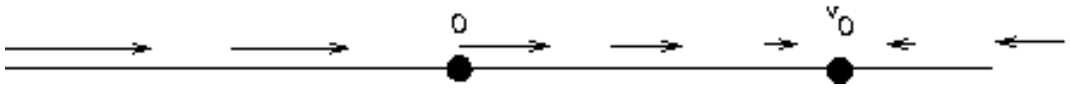

Figure 1. The effect of a sail in one dimension.

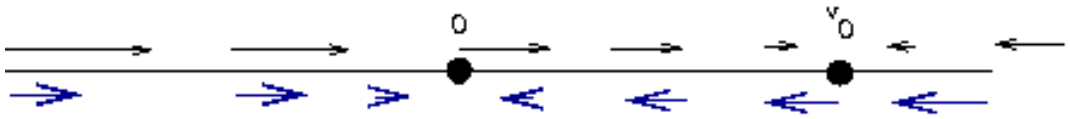

Figure 2. The effects of a sail and an anchor in one dimension.

The line here represents the space of all possible velocities v of a boat in this one-dimensional universe, including the rest velocity 0 and the wind velocity v_0. The vector field at any given velocity v represents the direction the velocity will move in if the sail is unfurled. We thus see that the effect of unfurling the sail will be to move the velocity of the sail towards v. Once one is at that speed, one is stuck there; neither furling nor unfurling the sail will affect one's velocity again in this frictionless model.

Now let's reinstate the role of the water. Let us use the crudest example of a water sail, namely an *anchor*. When the anchor is raised, we assume that we are back in the frictionless situation above; but when the anchor is dropped (so that it is dragging in the water), it exerts a force on the boat which is in the direction of the apparent velocity $0-v$ of the water with respect to the boat, and which (ideally) has a magnitude proportional to square of the apparent speed $|0-v|$, thanks to the drag equation. This gives a second vector field in velocity space that one is able to effect on the boat (displayed here as thick blue arrows); see Figure 2.

It is now apparent that by using either the sail or the anchor, one can reach any given velocity between 0 and v_0. However, once one is in this range, one cannot use the sail and anchor to move faster than v_0, or to move at a negative velocity.

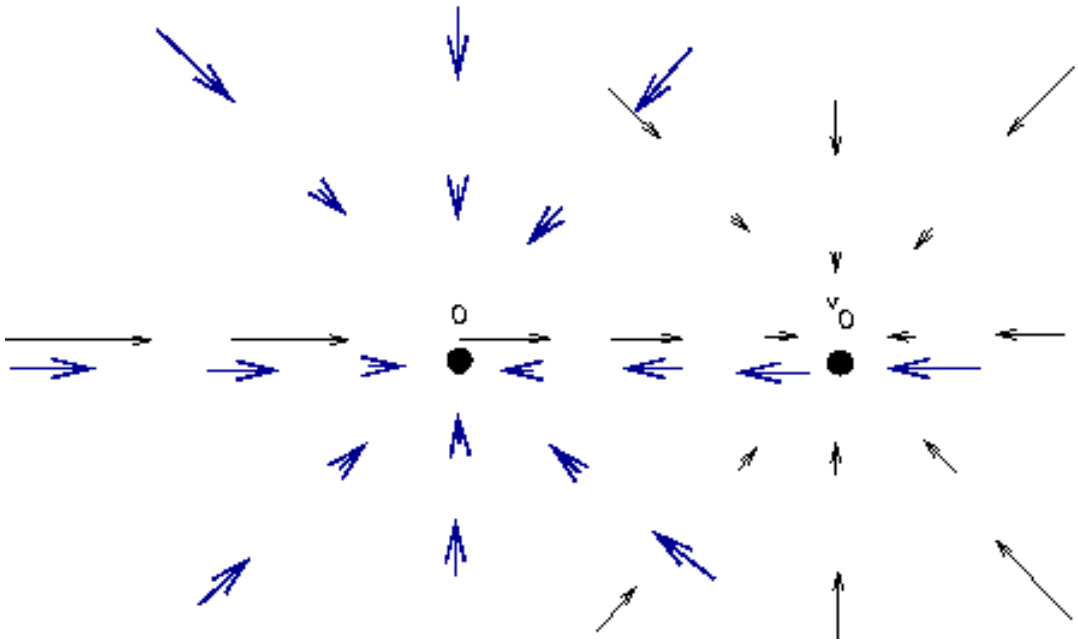

Figure 3. The effects of a pure-drag sail (black) and an anchor (blue) in two dimensions.

3.3.2. Two-dimensional sailing. Now let us sail in a two-dimensional plane $\mathbf{R}^2$, thus the wind velocity v_0 is now a vector in that plane. To begin with, let us again ignore the friction effects of the water (e.g. imagine one is ice yachting on a two-dimensional frozen lake).

With the square-rigged sails of the ancient era, which could only exploit drag, the net force exerted by an unfurled sail in two dimensions followed essentially the same law as in the one-dimensional case, i.e. the force was always proportional to the relative velocity $v_0 - v$ of the wind and the ship, thus leading to the black vector field in Figure 3.

We thus see that, starting from rest $v = 0$, the only thing one can do with such a sail is move the velocity v along the line segment from 0 to v_0, at which point one is stuck (unless one can exploit water friction, e.g. via an anchor, to move back down that line segment to 0). No crosswind velocity is possible at all with this type of sail.

With the invention of the curved sail, which redirects the (apparent) wind velocity $v_0 - v$ to another direction rather than stalling it to zero, it became possible for sails to provide a lift force[1] which is essentially perpendicular to the (apparent) wind velocity, in contrast to the drag force that is parallel to that velocity. (Not co-incidentally, such a sail has essentially the same aerofoil shape as an airplane wing, and one can also explain the lift force via *Bernoulli's principle*.)

By setting the sail in an appropriate direction, one can now use the lift force to adjust the velocity v of a sailboat in directions perpendicular to the apparent wind velocity $v_0 - v$, while using the drag force to adjust v in directions parallel to this apparent velocity; of course, one can also adjust the velocity in all intermediate directions by combining both drag and lift. This leads to the vector fields displayed in red in Figure 4.

Note that no matter how one orients the sail, the apparent wind speed $|v_0 - v|$ will decrease (or at best stay constant); this can also be seen from the law of conservation of energy in the reference frame of the wind. Thus, starting from rest, and using only the sail, one can only reach speeds in the circle centred at v_0 with radius $|v_0|$ (i.e. the circle in Figure 4); thus one cannot sail against the wind, but one can at least reach speeds of twice the wind speed, at least in principle[2].

Remark 3.3.1. If all one has to work with is the air sail(s), then one cannot do any better than what is depicted in Figure 4, no matter how complicated the rigging. This can be seen by looking at the law of conservation of energy in the reference frame of the wind. In that frame, the air is at rest and thus has zero kinetic energy, while the sailboat has kinetic energy $\frac{1}{2}m|v_0|^2$. The water in this frame has an enormous reservoir of kinetic energy, but if one is not allowed to interact with this water, then the kinetic energy of the boat cannot exceed $\frac{1}{2}m|v_0|^2$ in this frame, and so the boat velocity is limited to

[1]Despite the name, the lift force is not a vertical force in this context, but instead a horizontal one; in general, lift forces are basically perpendicular to the orientation of the aerofoil providing the lift. Unlike airplane wings, sails are vertically oriented, so the lift will be horizontal in this case.

[2]In practice, friction effects of air and water, such as wave making resistance, and the difficulty in forcing the sail to provide purely lift and no drag, mean that one cannot quite reach this limit, but it has still been possible to exceed the wind speed with this type of technique.

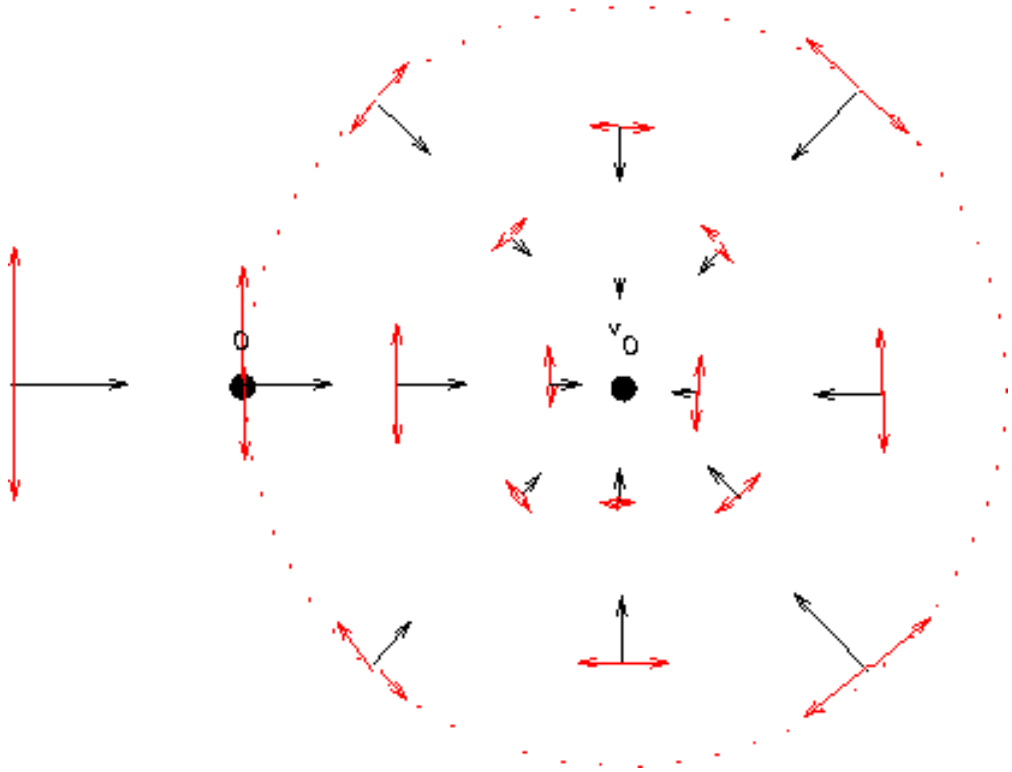

Figure 4. The effect of a pure-drag sail (black) and a pure-lift sail (red) in two dimensions. The disk enclosed by the dotted circle represents the velocities one can reach from these sails starting from the rest velocity $v = 0$.

the region inside the dotted circle. In particular, no arrangement of sails can give a negative drag force.

3.3.3. Three-dimensional sailing. Now we can turn to three-dimensional sailing, in which the sailboat is still largely confined to $\mathbf{R}^2$ but one can use both air sails and water sails as necessary to control the velocity v of the boat[3].

[3]Some boats do in fact exploit the third dimension more substantially than this, e.g. using sails to vertically lift the boat to reduce water drag, but we will not discuss these more advanced sailing techniques here.

As mentioned earlier, the crudest example of a water sail is an anchor, which, when dropped, exerts a pure drag force in the direction of $0-v$ on the boat; this is displayed as the blue vector field in Figure 3. Comparing this with Figure 4 (which is describing all the forces available from using the air sail) we see that such a device does not increase the range of velocities attainable from a boat starting at rest (although it does allow a boat moving with the wind to return to rest, as in the one-dimensional setting). Unsurprisingly, anchors are not used all that much for sailing in practice.

However, we can do better by using other water sails. For instance, the *keel* of a boat is essentially a water sail oriented in the direction of the boat (which in practice is kept close to parallel to v, e.g. by use of the rudder, else one would encounter substantial (and presumably unwanted) water drag and torque effects). The effect of the keel is to introduce significant resistance to any lateral movement of the boat. Ideally, the effect this has on the net force acting on the boat is that it should orthogonally project that force to be parallel to the direction of the boat (which, as stated before, is usually parallel to v). Applying this projection to the vector fields arising from the air sail, we obtain some new vector fields along which we can modify the boat's velocity; see Figure 5.

In particular, it becomes possible to sail against the wind, or faster than the wind, so long as one is moving at a non-trivial angle to the wind (i.e. v is not parallel to v_0 or $-v_0$).

What is going on here is as follows. By using lift instead of drag, and tacking the sail appropriately, one can make the force exerted by the sail be at any angle of up to 90° from the actual direction of apparent wind. By then using the keel, one can make the net force on the boat be at any angle up to 90° from the force exerted by the sail. Putting the two together, one can create a force on the boat at any angle up to 180° from the apparent wind speed - i.e. in any direction other than directly against the wind. (In practice, because it is impossible have a pure lift force free of drag, and because the keel does not perfectly eliminate all lateral forces, most sailboats can only move at angles up to about 135° or so from the apparent wind direction, though one can then create a net movement at larger angles

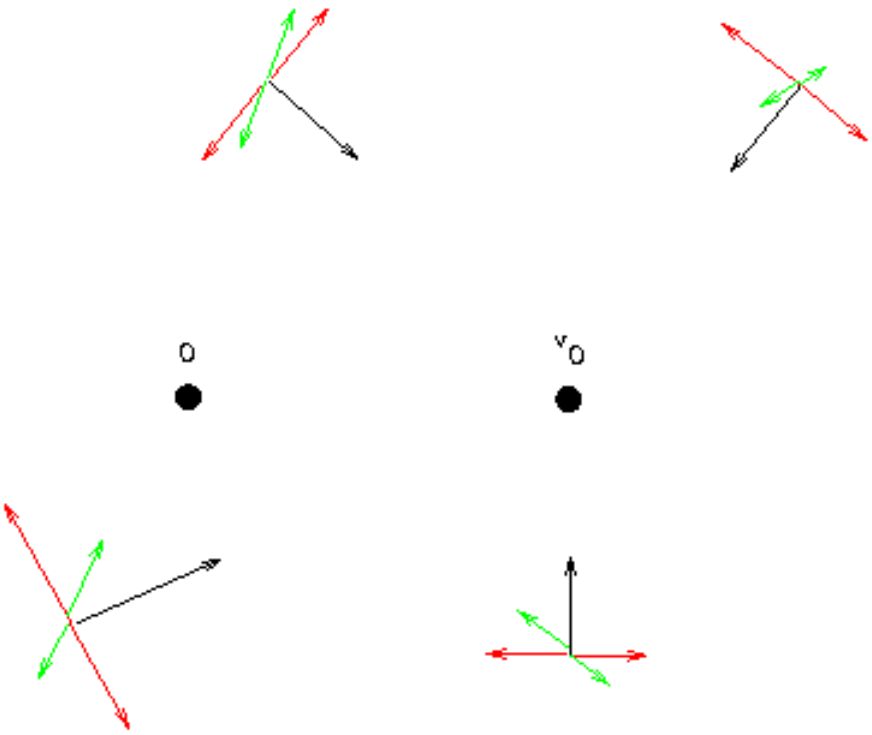

Figure 5. The effect of a pure-drag sail (black), a pure-lift sail (red), and a pure-lift sail combined with a keel (green). Note that one now has the ability to shift the velocity v away from both 0 and v_0 no matter how fast one is already traveling, so long as v is not collinear with 0 and v_0.

by tacking and beating. For similar reasons, water drag prevents one from using these methods to move too much faster than the wind speed.)

In theory, one can also sail at any desired speed and direction by combining the use of an air sail (or aerofoil) with the use of a water sail (or hydrofoil). While water is a rather different fluid from air in many respects (it is far denser, and virtually incompressible), one could in principle deploy hydrofoils to exert lift forces on a boat perpendicular to the apparent water velocity $0 - v$, much as an aerofoil can be used to exert lift forces on the boat perpendicular to the apparent wind

velocity $v_0 - v$. We saw in the previous section that if the effects of air resistance somehow be ignored, then one could use lift to alter the velocity v along a circle centred at the true wind speed v_0; similarly, if the effects of water resistance could also be ignored (e.g. by *planing*, which reduces, but does not completely eliminate, these effects), then one could alter the velocity v along a circle centred at the true water speed 0. By alternately using the aerofoil and hydrofoil, one could in principle reach arbitrarily large speeds and directions, as illustrated in Figure 6.

I do not know however if one could actually implement such a strategy with a physical sailing vessel. (Iceboats, however, have been known to reach speeds of up to six times the wind speed or more, though not exactly by the technique indicated in Figure 6. Thanks to kanyonman for this fact.)

It is reasonable (in light of results such as the Kutta-Joukowski theorem) to assume that the amount of lift provided by an aerofoil or hydrofoil is linearly proportional to the apparent wind speed or water speed. If so, then some basic trigonometry then reveals that (assuming negligible drag) one can use either of the above techniques to increase one's speed at what is essentially a constant rate; in particular, one can reach speeds of $n|v_0|$ for any $n > 0$ in time $O(n)$. On the other hand, as drag forces are quadratically proportional to apparent wind or water speed, one can decrease one's speed at an very rapid rate simply by dropping anchor; in fact one can drop speed from $n|v_0|$ to $|v_0|$ in bounded time $O(1)$ no matter how large n is! (This fact is the time-reversal of the well-known fact that the Riccati ODE $u' = u^2$ blows up in finite time.) These appear to be the best possible rates for acceleration or deceleration using only air and water sails, though I do not have a formal proof of this fact.

Notes. This article first appeared at `terrytao.wordpress.com/2009/03/23`.

Izabella Łaba pointed out several real-world sailing features not covered by the above simplified model, notably the interaction between multiple sails, and noted that the model was closer in many ways to windsurfing (or ice-sailing) than to traditional sailing.

Meichenl pointed out the relevance of the drag equation.

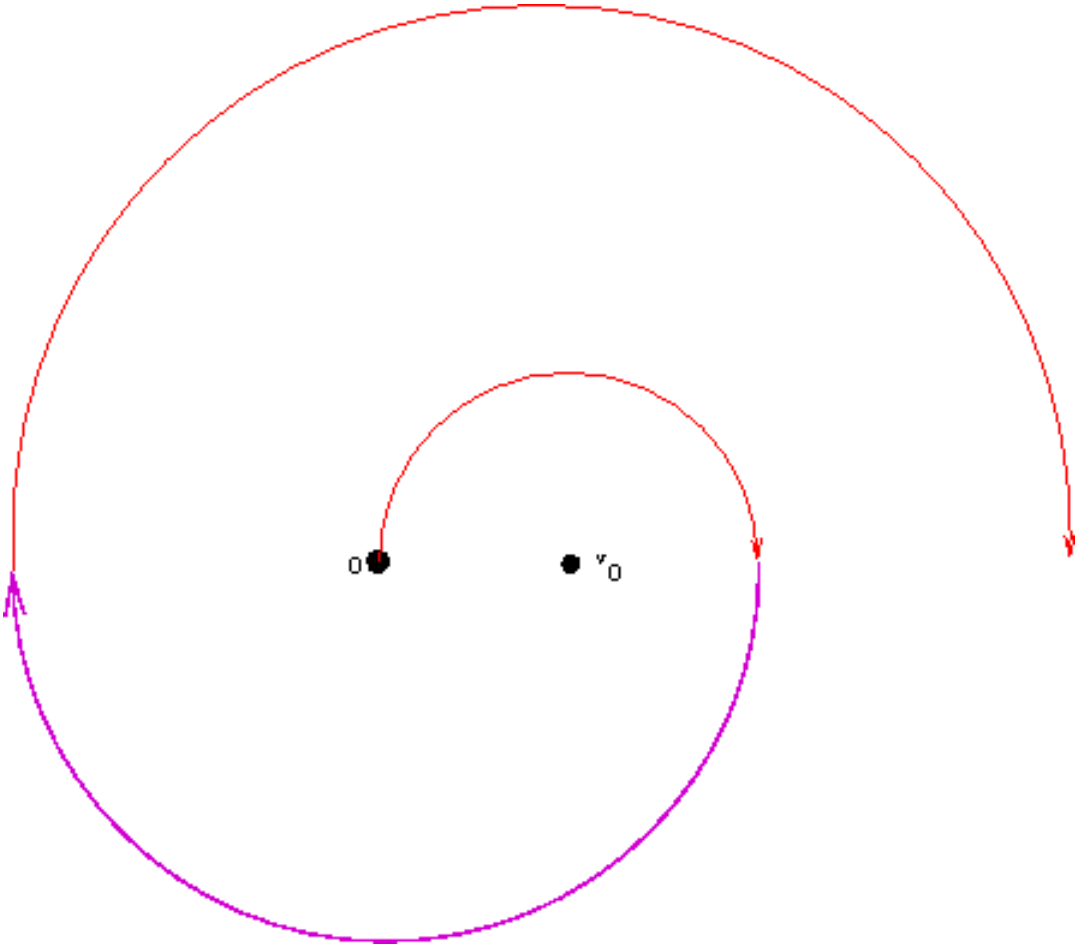

Figure 6. By alternating between a pure-lift aerofoil (red) and a pure-lift hydrofoil (purple), one can in principle reach arbitrarily large speeds in any direction.

3.4. The completeness and compactness theorems of first-order logic

The famous *Gödel completeness theorem* in logic (not to be confused with the even more famous *Gödel incompleteness theorem*) roughly states the following:

Theorem 3.4.1 (Gödel completeness theorem, informal statement)**.** *Let* Γ *be a* theory *(a* formal language $\mathcal{L}$*, together with a set of axioms, i.e.* sentences *assumed to be true), and let* ϕ *be a sentence in the formal language. Assume also that the language* $\mathcal{L}$ *has at most countably many symbols. Then the following are equivalent:*

(i) *(Syntactic consequence)* ϕ *can be deduced from the axioms in* Γ *by a finite number of applications of the* laws of deduction *in first order logic. (This property is abbreviated as* $\Gamma \vdash \phi$*.)*

(ii) *(Semantic consequence) Every* structure $\mathfrak{U}$ *which* satisfies *or* models Γ*, also satisfies* ϕ*. (This property is abbreviated as* $\Gamma \models \phi$*.)*

(iii) *(Semantic consequence for at most countable models) Every structure* $\mathfrak{U}$ *which is at most countable, and which models* Γ*, also satisfies* ϕ*.*

One can also formulate versions of the completeness theorem for languages with uncountably many symbols, but I will not do so here. One can also force other cardinalities on the model $\mathfrak{U}$ by using the *Löwenheim-Skolem theorem*.

To state this theorem even more informally, any (first-order) result which is true in *all* models of a theory, must be logically deducible from that theory, and vice versa. (For instance, any result which is true for all groups, must be deducible from the group axioms; any result which is true for all systems obeying *Peano arithmetic*, must be deducible from the Peano axioms; and so forth.) In fact, it suffices to check countable and finite models only; for instance, any first-order statement which is true for all finite or countable groups, is in fact true for all groups! Informally, a first-order language with only countably many symbols cannot "detect" whether a given structure is countably or uncountably infinite. Thus for instance even the *Zermelo-Frankel-Choice (ZFC)* axioms of set theory must have some at most countable model, even though one can use ZFC to prove the existence of uncountable sets; this is known as *Skolem's paradox*. (To resolve the paradox, one needs to carefully distinguish between an object in a set theory being "externally" countable in the structure that models that theory, and being "internally" countable within that theory.)

Of course, a theory Γ may contain *undecidable* statements ϕ - sentences which are neither provable nor disprovable in the theory. By the completeness theorem, this is equivalent to saying that ϕ is satisfied by some models of Γ but not by other models. Thus the completeness theorem is compatible with the incompleteness theorem: *recursively enumerable* theories such as Peano arithmetic are modeled by the natural numbers $\mathbf{N}$, but are also modeled by other structures also, and there are sentences satisfied by $\mathbf{N}$ which are not satisfied by other models of Peano arithmetic, and are thus undecidable within that arithmetic.

An important corollary of the completeness theorem is the *compactness theorem*:

Corollary 3.4.2 (Compactness theorem, informal statement)**.** *Let Γ be a first-order theory whose language has at most countably many symbols. Then the following are equivalent:*

(i) Γ *is* consistent*, i.e. it is not possible to logically deduce a contradiction from the axioms in* Γ*.*

(ii) Γ *is* satisfiable*, i.e. there exists a structure* $\mathfrak{U}$ *that models* Γ*.*

(iii) *There exists a structure* $\mathfrak{U}$ *which is at most countable, that models* Γ*.*

(iv) *Every finite subset* Γ' *of* Γ *is consistent.*

(v) *Every finite subset* Γ' *of* Γ *is satisfiable.*

(vi) *Every finite subset* Γ' *of* Γ *is satisfiable with an at most countable model.*

Indeed, the equivalence of (i)-(iii), or (iv)-(vi), follows directly from the completeness theorem, while the equivalence of (i) and (iv) follows from the fact that any logical deduction has finite length and so can involve at most finitely many of the axioms in Γ. (Again, the theorem can be generalised to uncountable languages, but the models become uncountable also.)

There is a consequence of the compactness theorem which more closely resembles the sequential concept of compactness. Given a sequence $\mathfrak{U}_1, \mathfrak{U}_2, \ldots$ be a sequence of structures for $\mathcal{L}$, and another

structure $\mathfrak{U}$ for $\mathcal{L}$, let us say that $\mathfrak{U}_n$ *converges elementarily* to $\mathfrak{U}$ if every sentence ϕ which is satisfied by $\mathfrak{U}$, is also satisfied by $\mathfrak{U}_n$ for sufficiently large n. (Replacing ϕ by its negation $\neg\phi$, we also see that every sentence that is not satisfied by $\mathfrak{U}$, is not satisfied by $\mathfrak{U}_n$ for sufficiently large n.) Note that the limit $\mathfrak{U}$ is only unique up to *elementary equivalence*. Clearly, if each of the $\mathfrak{U}_n$ models some theory Γ, then the limit $\mathfrak{U}$ will also; thus for instance the elementary limit of a sequence of groups is still a group, the elementary limit of a sequence of rings is still a ring, etc.

Corollary 3.4.3 (Sequential compactness theorem)**.** *Let $\mathcal{L}$ be a language with at most countably many symbols, and let $\mathfrak{U}_1, \mathfrak{U}_2, \ldots$ be a sequence of structures for $\mathcal{L}$. Then there exists a subsequence $\mathfrak{U}_{n_j}$ which converges elementarily to a limit $\mathfrak{U}$ which is at most countable.*

Proof. For each structure $\mathfrak{U}_n$, let $\mathrm{Th}(\mathfrak{U}_n)$ be the theory of that structure, i.e. the set of all sentences that are satisfied by that structure. One can view that theory as a point in $\{0,1\}^{\mathcal{S}}$, where $\mathcal{S}$ is the set of all sentences in the language $\mathcal{L}$. Since $\mathcal{L}$ has at most countably many symbols, $\mathcal{S}$ is at most countable, and so (by the sequential *Tychonoff theorem*) $\{0,1\}^{\mathcal{S}}$ is sequentially compact in the product topology. (This can also be seen directly by the usual *Arzelá-Ascoli* diagonalisation argument.) Thus we can find a subsequence $\mathrm{Th}(\mathfrak{U}_{n_j})$ which converges in the product topology to a limit theory $\Gamma \in \{0,1\}^{\mathcal{S}}$, thus every sentence in Γ is satisfied by $\mathfrak{U}_{n_j}$ for sufficiently large j (and every sentence not in Γ is not satisfied by $\mathfrak{U}_{n_j}$ for sufficiently large j). In particular, any finite subset of Γ is satisfiable, hence consistent; by the compactness theorem, Γ itself is therefore consistent, and has an at most countable model $\mathfrak{U}$. Also, each of the theories $\mathrm{Th}(\mathfrak{U}_{n_j})$ is clearly complete (given any sentence ϕ, either ϕ or $\neg\phi$ is in the theory), and so Γ is complete as well. One concludes that Γ is the theory of $\mathfrak{U}$, and hence $\mathfrak{U}$ is the elementary limit of the $\mathfrak{U}_{n_j}$ as claimed. □

Remark 3.4.4. It is also possible to state the compactness theorem using the topological notion of compactness, as follows: let X be the space of all structures of a given language $\mathcal{L}$, quotiented by elementary equivalence. One can define a topology on X by taking the sets $\{\mathfrak{U} \in X : \mathfrak{U} \models \phi\}$ as a sub-base, where ϕ ranges over all sentences.

Then the compactness theorem is equivalent to the assertion that X is topologically compact.

One can use the sequential compactness theorem to build a number of interesting "non-standard" models to various theories. For instance, consider the language $\mathcal{L}$ used by Peano arithmetic (which contains the operations $+, \times$ and the successor operation S, the relation $=$, and the constant 0), and adjoint a new constant N to create an expanded language $\mathcal{L} \cup \{N\}$. For each natural number $n \in \mathbf{N}$, let $\mathbf{N}_n$ be a structure for $\mathcal{L} \cup \{N\}$ which consists of the natural numbers $\mathbf{N}$ (with the usual interpretations of $+$, $\times$, etc.) and interprets the symbol N as the natural number n. By the compactness theorem, some subsequence of $\mathbf{N}_n$ must converge elementarily to a new structure $*\mathbf{N}$ of $\mathcal{L} \cup \{N\}$, which still models Peano arithmetic, but now has the additional property that $N > n$ for every (standard) natural number n; thus we have managed to create a non-standard model of Peano arithmetic which contains a non-standardly large number (one which is larger than every standard natural number).

The sequential compactness theorem also lets us construct infinitary limits of various sequences of finitary objects; for instance, one can construct infinite pseudo-finite fields as the elementary limits of sequences of finite fields. It also apepars to be related to a number of *correspondence principles* between finitary and infinitary objects, such as the Furstenberg correspondence principle between sets of integers and dynamical systems, or the more recent correspondence principles concerning graph limits.

In this article, I will review the proof of the completeness (and hence compactness) theorem. The material here is quite standard (I basically follow the usual proof of Henkin, and taking advantage of *Skolemisation*), but I wish to popularise the notion of an *elementary limit*, which is not particularly well-known[4].

[4]The closely related concept of an *ultraproduct* is better known, and can be used to prove most of the compactness theorem already, thanks to *Łos's theorem*, but I do not know how to use ultraproducts to ensure that the limiting model is countable. However, one can think (intuitively, at least), of the limit model $\mathfrak{U}$ in the above theorem as being the set of "constructible" elements of an ultraproduct of the $\mathfrak{U}_n$.

In order to emphasise the main ideas in the proof, I will gloss over some of the more technical details in the proofs, relying instead on informal arguments and examples at various points.

3.4.1. Completeness and compactness in propositional logic. The completeness and compactness theorems are results in first-order logic. But to motivate some of the ideas in proving these theorems, let us first consider the simpler case of *propositional logic*. The language $\mathcal{L}$ of a propositional logic consists of the following:

- A finite or infinite collection $A_1, A_2, A_3, \ldots$ of *propositional variables* - *atomic formulae* which could be true or false, depending on the interpretation;
- A collection of *logical connectives*, such as *conjunction* $\wedge$, *disjunction* $\vee$, *negation* $\neg$, or *implication* $\implies$. (The exact choice of which logical connectives to include in the language is to some extent a matter of taste.)
- Parentheses (in order to indicate the order of operations).

Of course, we assume that the symbols used for atomic formulae are distinct from those used for logical connectives, or for parentheses; we will implicitly make similar assumptions of this type in later sections without further comment.

Using this language, one can form *sentences* (or *formulae*) by some standard formal rules which I will not write down here. Typical examples of sentences in propositional logic are $A_1 \implies (A_2 \vee A_3)$, $(A_1 \wedge \neg A_1) \implies A_2$, and $(A_1 \wedge A_2) \vee (A_1 \wedge A_3)$. Each sentence is of finite length, and thus involves at most finitely many of the propositional variables. Observe that if $\mathcal{L}$ is at most countable, then there are at most countably many sentences.

The analogue of a structure in propositional logic is a *truth assignment*. A truth assignment $\mathfrak{U}$ for a propositional language $\mathcal{L}$ consists of a truth value $A_n^{\mathfrak{U}} \in \{\text{true}, \text{false}\}$ assigned to each propositional variable A_n. (Thus, for instance, if there are N propositional variables in the language, then there are 2^N possible truth assignments.) Once a truth assignment $\mathfrak{U}$ has assigned a truth value $A_n^{\mathfrak{U}}$ to each propositional variable A_n, it can then assign a truth value $\phi^{\mathfrak{U}}$ to any

other sentence ϕ in the language $\mathcal{L}$ by using the usual truth tables for conjunction, negation, etc.; we write $\mathfrak{U} \models \phi$ if $\mathfrak{U}$ assigns a true value to ϕ (and say that ϕ is *satisfied* by $\mathfrak{U}$), and $\mathfrak{U} \not\models \phi$ otherwise. Thus, for instance, if $A_1^{\mathfrak{U}} = \text{false}$ and $A_2^{\mathfrak{U}} = \text{true}$, then $\mathfrak{U} \models A_1 \vee A_2$ and $\mathfrak{U} \models A_1 \implies A_2$, but $\mathfrak{U} \not\models A_2 \implies A_1$. Some sentences, e.g. $A_1 \vee \neg A_1$, are true in every truth assignment; these are the (semantic) *tautologies*. At the other extreme, the negation of a tautology will of course be false in every truth assignment.

A *theory* Γ is a language $\mathcal{L}$, together with a (finite or infinite) collection of sentences (also called Γ) in that language. A truth assignment $\mathfrak{U}$ *satisfies* (or *models*) the theory Γ, and we write $\mathfrak{U} \models \Gamma$, if we have $\mathfrak{U} \models \phi$ for all $\phi \in \Gamma$. Thus, for instance, if $\mathfrak{U}$ is as in the preceding example and $\Gamma := \{A_1, A_1 \implies A_2\}$, then $\mathfrak{U} \models \Gamma$.

The analogue of the Gödel completeness theorem is then

Theorem 3.4.5 (Completeness theorem for propositional logic)**.** *Let Γ be a theory for a propositional language $\mathcal{L}$, and let ϕ be a sentence in $\mathcal{L}$. Then the following are equivalent:*

(i) *(Syntactic consequence) ϕ can be deduced from the axioms in Γ by a finite number of applications of the* laws of propositional logic*.*

(ii) *(Semantic consequence) Every truth assignment $\mathfrak{U}$ which satisfies (or models) Γ, also satisfies ϕ.*

One can list a complete set of laws of propositional logic used in (i), but we will not do so here.

To prove the completeness theorem, it suffices to show the following equivalent version.

Theorem 3.4.6 (Completeness theorem for propositional logic, again)**.** *Let Γ be a theory for a propositional language $\mathcal{L}$. Then the following are equivalent:*

(i) *Γ is* consistent*, i.e. it is not possible to logically deduce a contradiction from the axioms in Γ.*

(ii) *Γ is* satisfiable*, i.e. there exists a truth assignment $\mathfrak{U}$ that models Γ.*

Indeed, Theorem 3.4.5 follows from Theorem 3.4.6 by applying Theorem 3.4.6 to the theory $\Gamma \cup \{\neg\phi\}$ and taking contrapositives.

It remains to prove Theorem 3.4.6. It is easy to deduce (i) from (ii), because the laws of propositional logic are *sound*: given any truth assignment, it is easy to verify that these laws can only produce true conclusions given true hypotheses. The more interesting implication is to obtain (ii) from (i) - given a consistent theory Γ, one needs to produce a truth assignment that models that theory.

Let's first consider the case when the propositional language $\mathcal{L}$ is finite, so that there are only finitely many propositional variables $A_1, \ldots, A_N$. Then we can argue using the following "greedy algorithm".

- We begin with a consistent theory Γ.
- Observe that at least one of $\Gamma \cup \{A_1\}$ or $\Gamma \cup \{\neg A_1\}$ must be consistent. For if both $\Gamma \cup \{A_1\}$ and $\Gamma \cup \{\neg A_1\}$ led to a logical contradiction, then by the laws of logic one can show that Γ must also lead to a logical contradiction.
- If $\Gamma \cup \{A_1\}$ is consistent, we set $A_1^{\mathfrak{U}} := \text{true}$ and $\Gamma_1 := \Gamma \cup \{A_1\}$; otherwise, we set $A_1^{\mathfrak{U}} := \text{false}$ and $\Gamma_1 := \Gamma \cup \{\neg A_1\}$.
- Γ_1 is consistent, so arguing as before we know that at least one of $\Gamma_1 \cup \{A_2\}$ or $\Gamma_1 \cup \{\neg A_2\}$ must be consistent. If the former is consistent, we set $A_2^{\mathfrak{U}} := \text{true}$ and $\Gamma_2 := \Gamma_1 \cup \{A_2\}$; otherwise set $A_2^{\mathfrak{U}} := \text{false}$ and $\Gamma_2 := \Gamma_1 \cup \{\neg A_2\}$.
- We continue in this fashion, eventually ending up with a consistent theory Γ_N containing Γ, and a complete truth assignment $\mathfrak{U}$ such that $A_n \in \Gamma_N$ whenever $1 \leq n \leq N$ is such that $A_n^{\mathfrak{U}} = \text{true}$, and such that $\neg A_n \in \Gamma_N$ whenever $1 \leq n \leq N$ is such that $A_n^{\mathfrak{U}} = \text{false}$.
- From the laws of logic and an induction argument, one then sees that if ϕ is any sentence with $\phi^{\mathfrak{U}} = \text{true}$, then ϕ is a logical consequence of Γ_N, and hence (since Γ_N is consistent) $\neg\phi$ is not a consequence of Γ_N. Taking contrapositives, we see that $\phi^{\mathfrak{U}} = \text{false}$ whenever $\neg\phi$ is a consequence of Γ_N; replacing ϕ by $\neg\phi$ we conclude that $\mathfrak{U}$ satisfies every sentence in Γ_N, and the claim follows.

Remark 3.4.7. The above argument shows in particular that any finite theory either has a model or a proof of a contradictory statement (such as $A \wedge \neg A$). Actually producing a model if it exists, though, is essentially the infamous *satisfiability problem*, which is known to be NP-complete, and thus (if $P \neq NP$) would require super-polynomial time to execute.

The case of an infinite language can be obtained by combining the above argument with *Zorn's lemma* (or *transfinite induction* and the axiom of choice, if the set of propositional variables happens to be well-ordered). Alternatively, one can proceed by establishing

Theorem 3.4.8 (Compactness theorem for propositional logic)**.** *Let Γ be a theory for a propositional language $\mathcal{L}$. Then the following are equivalent:*

(i) *Γ is satisfiable.*

(ii) *Every finite subset Γ' of Γ is satisfiable.*

It is easy to see that Theorem 3.4.8 will allow us to use the finite case of Theorem 3.4.6 to deduce the infinite case, so it remains to prove Theorem 3.4.8. The implication of (ii) from (i) is trivial; the interesting implication is the converse.

Observe that there is a one-to-one correspondence between truth assignments $\mathfrak{U}$ and elements of the product space $\{0,1\}^{\mathcal{A}}$, where $\mathcal{A}$ is the set of propositional variables. For every sentence ϕ, let $F_\phi \subset \{0,1\}^{\mathcal{A}}$ be the collection of all truth assignments that satisfy ϕ; observe that this is a closed (and open) subset of $\{0,1\}^{\mathcal{A}}$ in the product topology (basically because ϕ only involves finitely many of the propositional variables). If every finite subset Γ' of Γ is satisfiable, then $\bigcup_{\phi\in\Gamma'} F_\phi$ is non-empty; thus the family $(F_\phi)_{\phi\in\Gamma}$ of closed sets enjoys the *finite intersection property*. On the other hand, from *Tychonoff's theorem*, $\{0,1\}^{\mathcal{A}}$ is compact. We conclude that $\bigcap_{\phi\in\Gamma} F_\phi$ is non-empty, and the claim follows.

Remark 3.4.9. While Tychonoff's theorem in full generality is equivalent to the axiom of choice, it is possible to prove the compactness theorem using a weaker version of this axiom, namely the *ultrafilter*

lemma. In fact, the compactness theorem is logically equivalent to this lemma.

3.4.2. Zeroth-order logic. Propositional logic is far too limited a language to do much mathematics. Let's make the language a bit more expressive, by adding constants, operations, relations, and (optionally) the equals sign; however, we refrain at this stage from adding variables or quantifiers, making this a *zeroth-order logic* rather than a first-order one.

A zeroth-order language $\mathcal{L}$ consists of the following objects:

- A (finite or infinite) collection $A_1, A_2, A_3, \ldots$ of propositional variables;
- A collection $R_1, R_2, R_3, \ldots$ of *relations* (or *predicates*), with each R_i having an *arity* (or *valence*) $a[R_i]$ (e.g. unary relation, binary relation, etc.);
- A collection $c_1, c_2, c_3, \ldots$ of constants;
- A collection $f_1, f_2, f_3, \ldots$ of *operators* (or functions), with each operator f_i having an arity $a[f_i]$ (e.g. unary operator, binary operator, etc.);
- Logical connectives;
- Parentheses;
- Optionally, the equals sign $=$.

For instance, a zeroth-order language for arithmetic on the natural numbers might include the constants $0, 1, 2, \ldots$, the binary relations $<, \leq, >, \geq$, the binary operations $+, \times$, the unary successor operation S, and the equals sign $=$. A zeroth-order language for studying all groups generated by six elements might include six generators $a_1, \ldots, a_6$ and the identity element e as constants, as well as the binary operation $\cdot$ of group multiplication and the unary operation $()^{-1}$ of group inversion, together with the equals sign $=$. And so forth.

Note that one could shorten the description of such languages by viewing propositional variables as relations of arity zero, and similarly viewing constants as operators of arity zero, but I find it conceptually clearer to leave these two operations separate, at least initially. As

we shall see shortly, one can also essentially eliminate the privileged role of the equals sign = by treating it as just another binary relation, which happens to have some standard axioms[5] attached to it.

By combining constants and operators together in the usual fashion one can create *terms*; for instance, in the zeroth-order language for arithmetic, $3+(4\times5)$ is a term. By inserting terms into a predicate or relation (or the equals sign =), or using a propositional variable, one obtains an *atomic formula*; thus for instance $3 + (4 \times 5) > 25$ is an atomic formula. By combining atomic formulae using logical connectives, one obtains a sentence (or *formula*); thus for instance $((4 \times 5) > 22) \implies (3 + (4 \times 5) > 25)$ is a sentence.

In order to assign meaning to sentences, we need the notion of a *structure* $\mathfrak{U}$ for a zeroth-order language $\mathcal{L}$. A structure consists of the following objects:

- A *domain of discourse* (or *universe of discourse*) $\mathrm{Dom}(\mathfrak{U})$;
- An assignment of a value $c_n^{\mathfrak{U}} \in \mathrm{Dom}(\mathfrak{U})$ to every constant c_n;
- An assignment of a function $f_n^{\mathfrak{U}} : \mathrm{Dom}(\mathfrak{U})^{a[f_n]} \to \mathrm{Dom}(\mathfrak{U})$ to every operation f_n;
- An assignment of a truth value $A_n^{\mathfrak{U}} \in \{\mathrm{true}, \mathrm{false}\}$ to every propositional variable A_n;
- An assignment of a function $R_n^{\mathfrak{U}} : \mathrm{Dom}(\mathfrak{U})^{a[R_n]}\{\mathrm{true}, \mathrm{false}\}$ to every relation R_n.

For instance, if $\mathcal{L}$ is the language of groups with six generators discussed above, then a structure $\mathfrak{U}$ would consist of a set $G = \mathrm{Dom}(\mathfrak{U})$, seven elements $a_1^{\mathfrak{U}}, \ldots, a_6^{\mathfrak{U}}, e^{\mathfrak{U}} \in G$ in that set, a binary operation $\cdot^{\mathfrak{U}} : G \times G \to G$, and a unary operation $(()^{-1})^{\mathfrak{U}} : G \to G$. At present, no group-type properties are assumed on these operations; the structure here is little more than a *magma* at this point.

Every sentence ϕ in a zeroth-order language $\mathcal{L}$ can be interpreted in a structure $\mathfrak{U}$ for that language to give a truth value $\phi^{\mathfrak{U}} \in \{\mathrm{true}, \mathrm{false}\}$, simply by substituting all symbols α in the language

[5]Namely, that equality is reflexive, transitive, and symmetric, and can be substituted in any expression to create an equal expression, or in any formula to create an equivalent formula.

with their interpreted counterparts $\alpha^{\mathfrak{U}}$ (note that the equals sign $=$ does not need any additional data in order to be interpreted). For instance, the sentence $a_1 \cdot a_2 = a_3$ is true in $\mathfrak{U}$ if $a_1^{\mathfrak{U}} \cdot^{\mathfrak{U}} a_2^{\mathfrak{U}} = a_3^{\mathfrak{U}}$. Similarly, every term t in the language can be interpreted to give a value $t^{\mathfrak{U}}$ in the domain of discourse $\mathrm{Dom}(\mathfrak{U})$.

As before, a theory is a collection of sentences; we can define satisfiability $\mathfrak{U} \models \phi$, $\mathfrak{U} \models \Gamma$ of a sentence ϕ or a theory Γ by a structure $\mathfrak{U}$ just as in the previous section. For instance, to describe groups with at most six generators in the language $\mathcal{L}$, one might use the theory Γ which consists of all the group axioms, specialised to terms, e.g. Γ would contain the associativity axioms $t_1 \cdot (t_2 \cdot t_3) = (t_1 \cdot t_2) \cdot t_3$ for all choices of terms t_1, t_2, t_3. (Note that this theory is not quite strong enough to capture the concept of a structure $\mathfrak{U}$ being a group generated by six elements, because the domain of $\mathfrak{U}$ may contain some "inaccessible" elements which are not the interpretation of any term in $\mathcal{L}$, but without the universal quantifier, there is not much we can do in zeroth-order logic to say anything about those elements, and so this is pretty much the best we can do with this limited logic.)

Now we can state the completeness theorem:

Theorem 3.4.10 (Completeness theorem for zeroth-order logic)**.** *Let Γ be a theory for a zeroth-order language $\mathcal{L}$, and let ϕ be a sentence in $\mathcal{L}$. Then the following are equivalent:*

(i) *(Syntactic consequence) ϕ can be deduced from the axioms in Γ by a finite number of applications of the laws of zeroth-order logic (i.e. all the* laws of first-order logic *that do not involve variables or quantifiers).*

(ii) *(Semantic consequence) Every truth assignment $\mathfrak{U}$ which satisfies (or models) Γ, also satisfies ϕ.*

To prove this theorem, it suffices as before to show that every consistent theory Γ in a zeroth-order logic is satisfiable, and conversely. The converse implication is again straightforward (the laws of zeroth-order logic are easily seen to be sound); the main task is to show the forward direction, i.e.

Proposition 3.4.11. *Let Γ be a consistent zeroth-order theory. Then Γ has at least one model.*

Proof. It is convenient to begin by eliminating the equality symbol from the language. Suppose we have already proven Proposition 3.4.11 has already been shown for languages without the equality symbol. Then we claim that the proposition also holds for languages with the equality symbol. Indeed, given a consistent theory Γ in a language $\mathcal{L}$ with equality, we can form a companion theory Γ' in the language $\mathcal{L}'$ formed by removing the equality symbol from $\mathcal{L}$ and replacing it with a new binary relation $='$, by taking all the sentences in Γ and replacing $=$ by $='$, and then adding in all the axioms of equality (with $=$ replaced by $='$) to Γ'. Thus, for instance, one would add the transitivity axioms $(x =' y) \wedge (y =' z) \implies (x =' z)$ to Γ for each triple of terms x, y, z, as well as substitution axioms such as $(x =' y) \implies (B(x,z) =' B(y,z))$ for any terms x, y, z and binary functions B. It is straightforward to verify that if Γ is consistent, then Γ' is also consistent, because any contradiction derived in Γ' can be translated to a contradiction derived in Γ simply by replacing $='$ with $=$ throughout and using the axioms of equality. By hypothesis, we conclude that Γ' has some model $\mathfrak{U}'$. By the axioms of equality, the interpretation $(=')^{\mathfrak{U}'}$ of $='$ in this model is then an *equivalence relation* on the domain $\mathrm{Dom}(\mathfrak{U}')$ of $\mathfrak{U}'$. One can also remove from the domain of $\mathfrak{U}'$ any element which is not of the form $t^{\mathfrak{U}'}$ for some term t, as such "inaccessible" elements will not influence the satisfiability of Γ'. We can then define a structure $\mathfrak{U}$ for the original language $\mathfrak{L}$ by *quotienting* the domain of $\mathfrak{U}'$ by the equivalence relation $='$, and also quotienting all the interpretations of the relations and operations of $\mathcal{L}$; the axioms of equality ensure that this quotienting is possible, and that the quotiented structure $\mathfrak{U}$ satisfies $\mathcal{L}$; we omit the details.

Henceforth we assume that $\mathcal{L}$ does not contain the equality sign. We will then choose a "tautological" domain of discourse $\mathrm{Dom}(\mathfrak{U})$, by setting this domain to be nothing more than the collection of all terms in the language $\mathcal{L}$. For instance, in the language of groups on six generators, the domain $\mathrm{Dom}(\mathfrak{U})$ is basically the free magma (with "inverse") on six generators plus an "identity", consisting of terms such as $(a_1 \cdot a_2)^{-1} \cdot a_1$, $(e \cdot a_3) \cdot ((a_4)^{-1})^{-1}$, etc. With this choice of domain, there is an obvious "tautological" interpretation of constants ($c^{\mathfrak{U}} := c$) and operations (e.g. $B^{\mathfrak{U}}(t_1, t_2) := B(t_1, t_2)^{\mathfrak{U}}$ for

binary operations B and terms t_1, t_2), which leads to every term t being interpreted as itself: $t^{\mathfrak{U}} = t$.

It remains to figure out how to interpret the propositional variables $A_1, A_2, \ldots$ and relations $R_1, R_2, \ldots$. Actually, one can replace each relation with an equivalent collection of new propositional variables by substituting in all possible terms in the relation. For instance, if one has a binary relation $R(,)$, one can replace this single relation symbol in the language by a (possibly infinite) collection of propositional variables $R(t_1, t_2)$, one for each pair of terms t_1, t_2, leading to a new (and probably much larger) language $\tilde{L}$ without any relation symbols. It is not hard to see that if theory Γ is consistent in $\mathcal{L}$, then the theory $\tilde{\Gamma}$ in $\tilde{L}$ formed by interpreting all atomic formulae such as $R(t_1, t_2)$ as propositional variables is also consistent. If $\tilde{\Gamma}$ has a model $\tilde{\mathfrak{U}}$ with the tautological domain of discourse, it is not hard to see that this can be converted to a model $\mathfrak{U}$ of Γ with the same domain by defining the interpretation $R^{\mathfrak{U}}$ of relations R in the obvious manner.

So now we may assume that there are no relation symbols, so that Γ now consists entirely of propositional sentences involving the propositional variables. But the claim now follows from the completeness theorem in propositional logic. □

Remark 3.4.12. The above proof can be viewed as a combination of the completeness theorem in propositional logic and the familiar procedure in algebra of constructing an algebraic object (e.g. a group) that obeys various relations, by starting with the free version of that object (e.g. a free group) and then quotienting out by the equivalence relation generated by those relations.

Remark 3.4.13. Observe that if $\mathfrak{L}$ is at most countable, then the structures $\mathfrak{U}$ constructed by the above procedure are at most countable (because the set of terms is at most countable, and quotienting by an equivalence relation cannot increase the cardinality). Thus we see (as in Theorem 3.4.1 or Corollary 3.4.2) that if a zeroth-order theory in an at most countable language is satisfiable, then it is in fact satisfiable with an at most countable model.

From the completeness theorem for zeroth-order logic and the above remark we obtain the compactness theorem for zeroth-order logic, which is formulated exactly as in Corollary 3.4.2.

3.4.3. First-order logic. We are now ready to study languages which are expressive enough to do some serious mathematics, namely the languages of first-order logic, which are formed from zeroth-order logics by adding variables and quantifiers. (There are *higher-order logics* as well, but unfortunately the completeness and compactness theorems typically fail for these, and they will not be discussed here.)

A language $\mathcal{L}$ for a first-order logic consists of the following:

- A (finite or infinite) collection $A_1, A_2, A_3, \ldots$ of propositional variables;
- A collection $R_1, R_2, R_3, \ldots$ of relations, with each R_i having an arity $a[R_i]$;
- A collection $c_1, c_2, c_3, \ldots$ of constants;
- A collection $f_1, f_2, f_3, \ldots$ of operators, with each f_i having an arity $a[f_i]$;
- A collection $x_1, x_2, x_3, \ldots$ of variables;
- Logical connectives;
- The *quantifiers* $\forall, \exists$;
- Parentheses;
- Optionally, the equals sign $=$.

For instance, the language for Peano arithmetic includes a constant 0, a unary operator S, binary operators $+, \times$, the equals sign $=$, and a countably infinite number of variables $x_1, x_2, \ldots$.

By combining constants, variables and operators together one creates terms; by inserting terms into predicates or relations, or using propositional variables, one obtains atomic formulae. These atomic formulae can contain a number of free variables. Using logical connectives as well as quantifiers to bind any or all of these variables, one obtains *well-formed formulae*; a formula with no free variables is a *sentence*. Thus, for instance, $\forall x_2 : x_1 + x_2 = x_2 + x_1$ is a well-formed formula, and $\forall x_1 \forall x_2 : x_1 + x_2 = x_2 + x_1$ is a sentence.

A structure $\mathfrak{U}$ for a first-order language $\mathfrak{L}$ is exactly the same concept as for a zeroth-order language: a domain of discourse, together with an interpretation of all the constants, operators, propositional variables, and relations in the language. Given a structure $\mathfrak{U}$, one can interpret terms t with no free variables as elements $t^{\mathfrak{U}}$ of $\mathrm{Dom}(\mathfrak{U})$, and interpret sentences ϕ as truth values $\phi^{\mathfrak{U}} \in \{\mathrm{true}, \mathrm{false}\}$, in the standard fashion.

A theory is, once again, a collection of sentences in the first-order language $\mathcal{L}$; one can define what it means for a structure to satisfy a sentence or a theory just as before.

Remark 3.4.14. In most fields of mathematics, one wishes to discuss several types of objects (e.g. numbers, sets, points, group elements, etc.) at once. For this, one would prefer to use a *typed* language, in which variables, constants, and functions take values in one type of object, and relations and functions take only certain types of objects as input. However, one can easily model a typed theory using a typeless theory by the trick of adding some additional unary predicates to capture type (e.g. $N(x)$ to indicate the assertion "x is a natural number", $S(x)$ to indicate the assertion "x is a set", etc.) and modifying the axioms of the theory being considered accordingly. (For instance, in a language involving both natural numbers and other mathematical objects, one might impose a new closure axiom $\forall x \forall y : N(x) \wedge N(y) \implies N(x+y)$, and axioms such as the commutativity axiom $\forall x \forall y : x+y = y+x$ would need to be modified to $\forall x \forall y : N(x) \wedge N(y) \implies x+y = y+x$.) It is a tedious but routine matter to show that the completeness and compactness theorems for typeless first-order logic imply analogous results for typed first-order logic; we omit the details.

To prove the completeness (and hence compactness) theorem, it suffices as before to show that

Proposition 3.4.15. *Let Γ be a consistent first-order theory, with an at most countable language $\mathcal{L}$. Then Γ has at least one model $\mathcal{U}$, which is also at most countable.*

We shall prove this result using a series of reductions. Firstly, we can mimic the arguments in the zeroth-order case and reduce to the

case when $\mathcal{L}$ does not contain the equality symbol. (We no longer need to restrict the domain of discourse to those elements which can be interpreted by terms, because the universal quantifier $\forall$ is now available for use when stating the axioms of equality.) Henceforth we shall assume that the equality symbol is not present in the language.

Next, by using the laws of first-order logic to push all quantifiers in the sentences in Γ to the beginning (e.g. replacing $(\forall x : P(x)) \wedge (\forall y : Q(y))$ with $\forall x \forall y : P(x) \wedge Q(y)$) one may assume that all sentences in Γ are in *prenex normal form*, i.e. they consist of a "matrix" of quantifiers, followed by an *quantifier-free formula* - a well-formed formula with no quantifiers. For instance, $\forall x \exists y \forall z \exists w : P(x, y, z, w)$ is in prenex normal form, where $P(x, y, z, w)$ is an quantifier-free formula with four free variables x, y, z, w.

Now we will start removing the existential quantifiers $\exists$ from the sentences in Γ. Let's begin with a simple case, when Γ contains a sentence of the form $\exists x : P(x)$ for some quantifier-free formula of one free variable x. Then one can eliminate the existential quantifier by introducing a *witness*, or more precisely adjoining a new constant c to the language $\mathcal{L}$ and replacing the statement $\exists x : P(x)$ with the statement $P(c)$, giving rise to a new theory Γ' in a new language Λ'. The consistency of Γ easily implies the consistency of Γ', while any at most countable model for Γ' can be easily converted to an at most countable model for Γ (by "forgetting" the symbol c). (In fact, Γ' is a *conservative extension* of Γ.) We can apply this reduction simultaneously to all sentences of the form $\exists x : P(x)$ in Γ (thus potentially expanding the collection of constants in the language by a countable amount).

The same argument works for any sentence in prenex normal form in which all the existential quantifiers are to the left of the universal quantifiers, e.g. $\exists x \exists y \forall z \forall w : P(x, y, z, w)$; this statement requires two constants to separately witness x and y, but otherwise one proceeds much as in the previous paragraph. But what about if one or more of the existential quantifiers is buried behind a universal quantifier? The trick is then to use *Skolemisation*. We illustrate this with the simplest case of this type, namely that of a sentence $\forall x \exists y : P(x, y)$. Here, one cannot use a constant witness for y. But this is no problem:

one simply introduces a witness that depends on x. More precisely, one adjoins a new unary operator c to the language $\mathcal{L}$ and replaces the statement $\forall x \exists y : P(x,y)$ by $\forall x : P(x, c(x))$, creating a new theory Γ' in a new language Λ'. One can again show (though this is not entirely trivial) that the consistency of Γ implies the consistency of Γ', and that every countable model for Γ' can be converted to a countable model for Γ (again by "forgetting" c). So one can eliminate the existential quantifier from this sentence also. Similar methods work for any other prenex normal form; for instance with the sentence

$$\forall x \exists y \forall z \exists w : P(x,y,z,w)$$

one can obtain a conservative extension of that theory by introducing a unary operator c and a binary operator d and replacing the above sentence with

$$\forall x \forall z : P(x, c(x), z, d(x,z)).$$

One can show that one can perform Skolemisation on all the sentences in Γ simultaneously, which has the effect of eliminating all existential quantifiers from Γ while still keeping the language $\mathcal{L}$ at most countable (since Γ is at most countable). (Intuitively, what is going on here is that we are interpreting all existential axioms in the theory as implicitly defining functions, which we then explicitly formalise as a new symbol in the language. For instance, if we had some theory of sets which contained the *axiom of choice* (every family of non-empty sets $(X_\alpha)_{\alpha \in A}$ admits a *choice function* $f : A \to \bigcup_{\alpha \in A} X_\alpha$), then we can Skolemise this by introducing a "choice function function" $\mathcal{F} : (X_\alpha)_{\alpha \in A} \mapsto \mathcal{F}((X_\alpha)_{\alpha \in A})$ that witnessed this axiom to the language. Note that we do not need uniqueness in the existential claim in order to be able to perform Skolemisation.)

After performing Skolemisation and adding all the witnesses to the language, we are reduced to the case in which all the sentences in Γ are in fact universal statements, i.e. of the form $\forall x_1 \ldots \forall x_k : P(x_1, \ldots, x_k)$, where $P(x_1, \ldots, x_k)$ is an quantifier-free formula of k free variables. In this case one can repeat the zeroth-order arguments, selecting a structure $\mathfrak{U}$ whose domain of discourse is the tautological one, indexed by all the terms with no free variables (in particular, this structure will be countable). One can then replace each first-order

statement $\forall x_1 \dots \forall x_k : P(x_1, \dots, x_k)$ in Γ by the family of zeroth-order statements $P(t_1, \dots, t_k)$, where $t_1, \dots, t_k$ ranges over all terms with no free variables, thus creating a zeroth-order theory Γ_0. As Γ is consistent, Γ_0 is also, so by the zeroth-order theory, we can find a model $\mathfrak{U}$ for Γ_0 with the tautological domain of discourse, and it is clear that this structure will also be a model for the original theory Γ. The proof of the completeness theorem (and thus the compactness theorem) is now complete.

In summary: to create a countable model from a consistent first-order theory, one first replaces the equality sign $=$ (if any) by a binary relation $='$, then uses Skolemisation to make all implied functions and operations explicit elements of the language. Next, one makes the zeroth-order terms of the new language the domain of discourse, applies a greedy algorithm to decide the truth assignment of all zeroth-order sentences, and then finally quotients out by the equivalence relation given by $='$ to recover the countable model.

Remark 3.4.16. I find the use of Skolemisation to greatly clarify, at a conceptual level, the proof of the completeness theorem. However, at a technical level it does make things more complicated: in particular, showing that the Skolemisation of a consistent theory is still consistent does require some non-trivial effort (one has to take all arguments involving the Skolem function $c()$, and replace every occurence of $c()$ by a "virtual" function, defined implicitly using existential quantifiers). On the other hand, this fact is easy to prove once one already *has* the completeness theorem, though we of course cannot formally take advantage of this while trying to *prove* that theorem!

The more traditional Henkin approach is based instead on adding a large number of constant witnesses, one for every existential statement: roughly speaking, for each existential sentence $\exists x : P(x)$ in the language, one adds a new constant c to the language and inserts an axiom $(\exists x : P(x)) \implies P(c)$ to the theory; it is easier to show that this preserves consistency than it is with a more general Skolemisation. Unfortunately, every time one adds a constant to the language, one increases the number of existential sentences for which one needs to perform this type of witnessing, but it turns out that after applying

this procedure a countable number of times, one can get to the point where every existential sentence is automatically witnessed by some constant. This has the same ultimate effect as Skolemisation, namely one can convert sentences containing existential quantifiers to ones which are purely universal, and so the rest of the proof is much the same as the proof described above. On the other hand, the Henkin approach avoids the axiom of choice (though one still must use the ultrafilter lemma, of course).

Notes. This article first appeared at `terrytao.wordpress.com/2009/04/10`. Thanks to Carson Chow, Ernie Cohen, Eric, John Goodrick, and anonymous commenters for corrections.

3.5. Talagrand's concentration inequality

In the theory of discrete random matrices (e.g. matrices whose entries are random signs ± 1), one often encounters the problem of understanding the distribution of the random variable $\mathrm{dist}(X, V)$, where $X = (x_1, \ldots, x_n) \in \{-1, +1\}^n$ is an n-dimensional random sign vector (so X is uniformly distributed in the discrete cube $\{-1, +1\}^n$), and V is some d-dimensional subspace of $\mathbf{R}^n$ for some $0 \leq d \leq n$.

It is not hard to compute the second moment of this random variable. Indeed, if $P = (p_{ij})_{1 \leq i,j \leq n}$ denotes the orthogonal projection matrix from $\mathbf{R}^n$ to the orthogonal complement $V^\perp$ of V, then one observes that

$$\mathrm{dist}(X, V)^2 = X \cdot PX = \sum_{i=1}^n \sum_{j=1}^n x_i x_j p_{ij}$$

and so upon taking expectations we see that

$$\mathbf{E}\,\mathrm{dist}(X, V)^2 = \sum_{i=1}^n p_{ii} = \mathrm{tr}\, P = n - d \tag{3.23}$$

since P is a rank $n-d$ orthogonal projection. So we expect $\mathrm{dist}(X, V)$ to be about $\sqrt{n-d}$ on the average.

In fact, one has sharp concentration around this value, in the sense that $\mathrm{dist}(X, V) = \sqrt{n-d} + O(1)$ with high probability. More precisely, we have

Proposition 3.5.1 (Large deviation inequality)**.** *For any $t > 0$, one has*

$$\mathbf{P}(|\operatorname{dist}(X, V) - \sqrt{n-d}| \geq t) \leq C \exp(-ct^2)$$

for some absolute constants $C, c > 0$.

In fact the constants C, c are very civilised; for large t one can basically take $C = 4$ and $c = 1/16$, for instance. This type of concentration, particularly for subspaces V of moderately large codimension[6] $n - d$, is fundamental to much of my work on random matrices with Van Vu, starting with our first paper[**TaVu2006**] (in which this proposition first appears).

Proposition 3.5.1 is an easy consequence of the second moment computation and *Talagrand's inequality*[**Ta1996**], which among other things provides a sharp concentration result for convex Lipschitz functions on the cube $\{-1, +1\}^n$; since $\operatorname{dist}(x, V)$ is indeed a convex Lipschitz function, this inequality can be applied immediately. The proof of Talagrand's inequality is short and can be found in several textbooks (e.g. [**AlSp2008**]), but I thought I would reproduce the argument here (specialised to the convex case), mostly to force myself to learn the proof properly. Note the concentration of $O(1)$ obtained by Talagrand's inequality is much stronger than what one would get from more elementary tools such as *Azuma's inequality* or *McDiarmid's inequality*, which would only give concentration of about $O(\sqrt{n})$ or so (which is in fact trivial, since the cube $\{-1, +1\}^n$ has diameter $2\sqrt{n}$); the point is that Talagrand's inequality is very effective at exploiting the convexity of the problem, as well as the Lipschitz nature of the function in all directions, whereas Azuma's inequality can only easily take advantage of the Lipschitz nature of the function in coordinate directions. On the other hand, Azuma's inequality works just as well if the ℓ^2 metric is replaced with the larger ℓ^1 metric, and one can conclude that the ℓ^1 distance between X and V concentrates around its median to a width $O(\sqrt{n})$, which is a more non-trivial fact than the ℓ^2 concentration bound given by that inequality. (The computation

[6] For subspaces of small codimension (such as hyperplanes) one has to use other tools to get good results, such as *inverse Littlewood-Offord theory* or the *Berry-Esséen central limit theorem*, but that is another story.

of the median of the ℓ^1 distance is more complicated than for the ℓ^2 distance, though, and depends on the orientation of V.)

Remark 3.5.2. If one makes the coordinates of X iid Gaussian variables $x_i \equiv N(0,1)$ rather than random signs, then Proposition 3.5.1 is much easier to prove; the probability distribution of a Gaussian vector is rotation-invariant, so one can rotate V to be, say, $\mathbf{R}^d$, at which point $\mathrm{dist}(X,V)^2$ is clearly the sum of $n-d$ independent squares of Gaussians (i.e. a *chi-square distribution*), and the claim follows from direct computation (or one can use the *Chernoff inequality*). The gaussian counterpart of Talagrand's inequality is more classical, being essentially due to Lévy, and will also be discussed later in this post.

3.5.1. Concentration on the cube. Proposition 3.5.1 follows easily from the following statement, that asserts that if a convex set $A \subset \mathbf{R}^n$ occupies a non-trivial fraction of the cube $\{-1,+1\}^n$, then the neighbourhood $A_t := \{x \in \mathbf{R}^n : \mathrm{dist}(x,A) \leq t\}$ will occupy almost all of the cube for $t \gg 1$:

Proposition 3.5.3 (Talagrand's concentration inequality)**.** *Let A be a convex set in $\mathbf{R}^d$. Then*

$$\mathbf{P}(X \in A)\mathbf{P}(X \notin A_t) \leq \exp(-ct^2)$$

for all $t > 0$ and some absolute constant $c > 0$, where $X \in \{-1,+1\}^n$ is chosen uniformly from $\{-1,+1\}^n$.

Remark 3.5.4. It is crucial that A is convex here. If instead A is, say, the set of all points in $\{-1,+1\}^n$ with fewer than $n/2 - \sqrt{n}$ $+1$'s, then $\mathbf{P}(X \in A)$ is comparable to 1, but $\mathbf{P}(X \notin A_t)$ only starts decaying once $t \gg \sqrt{n}$, rather than $t \gg 1$. Indeed, it is not hard to show that Proposition 3.5.3 implies the variant

$$\mathbf{P}(X \in A)\mathbf{P}(X \notin A_t) \leq \exp(-ct^2/n)$$

for non-convex A (by restricting A to $\{-1,+1\}^n$ and then passing from A to the convex hull, noting that distances to A on $\{-1,+1\}^n$ may be contracted by a factor of $O(\sqrt{n})$ by this latter process); this inequality can also be easily deduced from *Azuma's inequality*.

To apply this proposition to the situation at hand, observe that if A is the cylindrical region $\{x \in \mathbf{R}^n : \mathrm{dist}(x,V) \leq r\}$ for some r, then A is convex and A_t is contained in $\{x \in \mathbf{R}^n : \mathrm{dist}(x,V) \leq r+t\}$. Thus

$$\mathbf{P}(\mathrm{dist}(X,V) \leq r)\mathbf{P}(\mathrm{dist}(X,V) > r+t) \leq \exp(-ct^2).$$

Applying this with $r := M$ or $r := M-t$, where M is the median value of $\mathrm{dist}(X,V)$, one soon obtains concentration around the median:

$$\mathbf{P}(|\mathrm{dist}(X,V) - M| > t) \leq 4\exp(-ct^2).$$

This is only compatible with (3.23) if $M = \sqrt{n-d} + O(1)$, and the claim follows.

To prove Proposition 3.5.3, we use the exponential moment method. Indeed, it suffices by Markov's inequality to show that

$$\mathbf{P}(X \in A)\mathbf{E}\exp(c\,\mathrm{dist}(X,A)^2) \leq 1 \tag{3.24}$$

for a sufficiently small absolute constant $c > 0$ (in fact one can take $c = 1/16$).

We prove (3.24) by an induction on the dimension n. The claim is trivial for $n = 0$, so suppose $n \geq 1$ and the claim has already been proven for $n-1$.

Let us write $X = (X', x_n)$ for $x_n = \pm 1$. For each $t \in \mathbf{R}$, we introduce the slice $A_t := \{x' \in \mathbf{R}^{n-1} : (x',t) \in A\}$, then A_t is convex. We now try to bound the left-hand side of (3.24) in terms of X', A_t rather than X, A. Clearly

$$\mathbf{P}(X \in A) = \frac{1}{2}[\mathbf{P}(X' \in A_{-1}) + \mathbf{P}(X' \in A_{+1})].$$

By symmetry we may assume that $\mathbf{P}(X' \in A_{+1}) \geq \mathbf{P}(X' \in A_{-1})$, thus we may write

$$\mathbf{P}(X' \in A_{\pm 1}) = p(1 \pm q) \tag{3.25}$$

where $p := \mathbf{P}(X \in A)$ and $0 \leq q \leq 1$.

Now we look at $\mathrm{dist}(X,A)^2$. For $t = \pm 1$, let $Y_t \in \mathbf{R}^{n-1}$ be the closest point of (the closure of) A_t to X', thus

$$|X' - Y_t| = \mathrm{dist}(X', A_t). \tag{3.26}$$

Let $0 \leq \lambda \leq 1$ be chosen later; then the point $(1-\lambda)(Y_{x_n}, x_n) + \lambda(Y_{-x_n}, -x_n)$ lies in A by convexity, and so

$$\operatorname{dist}(X, A) \leq |(1-\lambda)(Y_{x_n}, x_n) + \lambda(Y_{-x_n}, -x_n) - (X', x_n)|.$$

Squaring this and using Pythagoras, one obtains

$$\operatorname{dist}(X, A)^2 \leq 4\lambda^2 + |(1-\lambda)(X' - Y_{x_n}) + \lambda(X' - Y_{-x_n})|^2.$$

As we will shortly be exponentiating the left-hand side, we need to linearise the right-hand side. Accordingly, we will exploit the convexity of the function $x \mapsto |x|^2$ to bound

$$|(1-\lambda)(X - Y_{x_n}) + \lambda(X - Y_{-x_n})|^2 \leq$$
$$(1-\lambda)|X' - Y_{x_n}|^2 + \lambda|X' - Y_{-x_n}|^2$$

and thus by (3.26)

$$\operatorname{dist}(X, A)^2 \leq 4\lambda^2 + (1-\lambda)\operatorname{dist}(X', A_{x_n})^2 + \lambda \operatorname{dist}(X', A_{-x_n})^2.$$

We exponentiate this and take expectations in X' (holding x_n fixed for now) to get

$$\mathbf{E}_{X'} e^{c \operatorname{dist}(X,A)^2} \leq e^{4c\lambda^2} \mathbf{E}_{X'} (e^{c \operatorname{dist}(X', A_{x_n})^2})^{1-\lambda} (e^{c \operatorname{dist}(X', A_{-x_n})^2})^{\lambda}.$$

Meanwhile, from the induction hypothesis and (3.25) we have

$$\mathbf{E}_{X'} e^{c \operatorname{dist}(X', A_{x_n})^2} \leq \frac{1}{p(1+x_n q)}$$

and similarly for A_{-x_n}. By Hölder's inequality, we conclude

$$\mathbf{E}_{X'} e^{c \operatorname{dist}(X,A)^2} \leq e^{4c\lambda^2} \frac{1}{p(1+x_n q)^{1-\lambda}(1-x_n q)^{\lambda}}.$$

For $x_n = +1$, the optimal choice of λ here is 0, obtaining

$$\mathbf{E}_{X'} e^{c \operatorname{dist}(X,A)^2} = \frac{1}{p(1+q)};$$

for $x_n = -1$, the optimal choice of λ is to be determined. Averaging, we obtain

$$\mathbf{E}_X e^{c \operatorname{dist}(X,A)^2} = \frac{1}{2}\left[\frac{1}{p(1+q)} + e^{4c\lambda^2} \frac{1}{p(1-q)^{1-\lambda}(1+q)^{\lambda}}\right]$$

so to establish (3.24), it suffices to pick $0 \leq \lambda \leq 1$ such that

$$\frac{1}{1+q} + e^{4c\lambda^2} \frac{1}{(1-q)^{1-\lambda}(1+q)^{\lambda}} \leq 2.$$

If q is bounded away from zero, then by choosing $\lambda = 1$ we would obtain the claim if c is small enough, so we may take q to be small. But then a Taylor expansion allows us to conclude if we take λ to be a constant multiple of q, and again pick c to be small enough. The point is that $\lambda = 0$ already almost works up to errors of $O(q^2)$, and increasing λ from zero to a small non-zero quantity will decrease the LHS by about $O(\lambda q) - O(c\lambda^2)$.

By optimising everything using first-year calculus, one eventually gets the constant $c = 1/16$ claimed earlier.

Remark 3.5.5. Talagrand's inequality is in fact far more general than this; it applies to arbitrary products of probability spaces, rather than just to $\{-1,+1\}^n$, and to non-convex A, but the notion of distance needed to define A_t becomes more complicated; the proof of the inequality, though, is essentially the same. Besides its applicability to convex Lipschitz functions, Talagrand's inequality is also very useful for controlling combinatorial Lipschitz functions F which are "locally certifiable" in the sense that whenever $F(x)$ is larger than some threshold t, then there exist some bounded number $f(t)$ of coefficients of x which "certify" this fact (in the sense that $F(y) \geq t$ for any other y which agrees with x on these coefficients). See e.g. [**AlSp2008**] for a more precise statement and some applications.

3.5.2. Gaussian concentration. As mentioned earlier, there are analogous results when the uniform distribution on the cube $\{-1,+1\}^n$ are replaced by other distributions, such as the n-dimensional Gaussian distribution. In fact, in this case convexity is not needed:

Proposition 3.5.6 (Gaussian concentration inequality)**.** *Let A be a measurable set in $\mathbf{R}^d$. Then*

$$\mathbf{P}(X \in A)\mathbf{P}(X \notin A_t) \leq \exp(-ct^2)$$

for all $t > 0$ and some absolute constant $c > 0$, where $X \equiv N(0,1)^n$ is a random Gaussian vector.

This inequality can be deduced from *Lévy's classical concentration of measure inequality for the sphere* (with the optimal constant), but we will give an alternate proof due to Maurey and Pisier. It suffices to prove the following variant of Proposition 3.5.6:

Proposition 3.5.7 (Gaussian concentration inequality for Lipschitz functions)**.** *Let $f : \mathbf{R}^d \to \mathbf{R}$ be a function which is Lipschitz with constant 1 (i.e. $|f(x) - f(y)| \leq |x - y|$ for all $x, y \in \mathbf{R}^d$). Then for any t we have*

$$\mathbf{P}(|f(X) - \mathbf{E}f(X)| \geq t) \leq \exp(-ct^2)$$

for all $t > 0$ and some absolute constant $c > 0$, where $X \equiv N(0,1)^n$ is a random variable.

Indeed, if one sets $f(x) := \mathrm{dist}(x, A)$ one can soon deduce Proposition 3.5.6 from Proposition 3.5.7.

Informally, Proposition 3.5.7 asserts that Lipschitz functions of Gaussian variables concentrate as if they were Gaussian themselves; for comparison, Talagrand's inequality implies that *convex* Lipschitz functions of *Bernoulli* variables concentrate as if they were Gaussian.

Now we prove Proposition 3.5.7. By the *epsilon regularisation argument* (Section 2.7) we may take f to be smooth, and so by the Lipschitz property we have

$$|\nabla f(x)| \leq 1 \tag{3.27}$$

for all x. By subtracting off the mean we may assume $\mathbf{E}f = 0$. By replacing f with $-f$ if necessary it suffices to control the upper tail probability $\mathbf{P}(f(X) \geq t)$ for $t > 0$.

We again use the exponential moment method. It suffices to show that

$$\mathbf{E}\exp(tf(X)) \leq \exp(Ct^2)$$

for some absolute constant C.

Now we use a variant of the *square and rearrange* trick. Let Y be an independent copy of X. Since $\mathbf{E}f(Y) = 0$, we see from *Jensen's inequality* that $\mathbf{E}\exp(-tf(Y)) \geq 1$, and so

$$\mathbf{E}\exp(tf(X)) \leq \mathbf{E}\exp(t(f(X) - f(Y))).$$

With an eye to exploiting (3.27), one might seek to use the fundamental theorem of calculus to write

$$f(X) - f(Y) = \int_0^1 \frac{d}{d\lambda} f((1-\lambda)Y + \lambda X)\, d\lambda.$$

But actually it turns out to be smarter to use a circular arc of integration, rather than a line segment:

$$f(X) - f(Y) = \int_0^{\pi/2} \frac{d}{d\theta} f(Y\cos\theta + X\sin\theta)\ d\theta.$$

The reason for this is that $X_\theta := Y\cos\theta + X\sin\theta$ is another gaussian random variable equivalent to $N(0,1)^n$, as is its derivative $X'_\theta := -Y\sin\theta + X\cos\theta$; furthermore, and crucially, these two random variables are *independent*.

To exploit this, we first use Jensen's inequality to bound

$$\exp(t(f(X) - f(Y))) \leq \frac{\pi}{2} \int_0^{\pi/2} \exp(\frac{2t}{\pi} \frac{d}{d\theta} f(X_\theta))\ d\theta.$$

Applying the chain rule and taking expectations, we have

$$\mathbf{E}\exp(t(f(X) - f(Y))) \leq \frac{\pi}{2} \int_0^{\pi/2} \mathbf{E}\exp(\frac{2t}{\pi} \nabla f(X_\theta) \cdot X'_\theta)\ d\theta.$$

Let us condition X_θ to be fixed, then $X'_\theta \equiv N(0,1)^n$; applying (3.27), we conclude that $\frac{2t}{\pi}\nabla f(X_\theta)\cdot X'_\theta$ is normally distributed with standard deviation at most $\frac{2t}{\pi}$. As such we have

$$\mathbf{E}\exp(\frac{2t}{\pi}\nabla f(X_\theta)\cdot X'_\theta) \leq \exp(Ct)$$

for some absolute constant C; integrating out the conditioning on X_θ we obtain the claim.

Notes. This article first appeared at `terrytao.wordpress.com/2009/06/09`. Thanks to Oded and vedadi for corrections.

3.6. The Szemerédi-Trotter theorem and the cell decomposition

The celebrated *Szemerédi-Trotter theorem* gives a bound for the set of incidences $I(P,L) := \{(p,\ell) \in P \times L : p \in \ell\}$ between a finite set of points P and a finite set of lines L in the Euclidean plane $\mathbf{R}^2$. Specifically, the bound is

$$|I(P,L)| \ll |P|^{2/3}|L|^{2/3} + |P| + |L| \tag{3.28}$$

where we use the asymptotic notation $X \ll Y$ or $X = O(Y)$ to denote the statement that $X \leq CY$ for some absolute constant C.

In particular, the number of incidences between n points and n lines is $O(n^{4/3})$. This bound is sharp; consider for instance the discrete box $P := \{(a,b) \in \mathbf{Z}^2 : 1 \leq a \leq N; 1 \leq b \leq 2N^2\}$ with L being the collection of lines $\{(x, mx+b) : m, b \in \mathbf{Z}, 1 \leq m \leq N, 1 \leq b \leq N^2\}$. One easily verifies that $|P| = 2N^3$, $|L| = N^3$, and $|I(P,L)| = N^4$, showing that (3.28) is essentially sharp in the case $|P| \sim |L|$; one can concoct similar examples for other regimes of $|P|$ and $|L|$.

On the other hand, if one replaces the Euclidean plane $\mathbf{R}^2$ by a finite field geometry F^2, where F is a finite field, then the estimate (3.28) is false. For instance, if P is the entire plane F^2, and L is the set of all lines in F^2, then $|P|, |L|$ are both comparable to $|F|^2$, but $|I(P,L)|$ is comparable to $|F|^3$, thus violating (3.28) when $|F|$ is large. Thus any proof of the Szemerédi-Trotter theorem must use a special property of the Euclidean plane which is not enjoyed by finite field geometries. In particular, this strongly suggests that one cannot rely purely on algebra and combinatorics to prove (3.28); one must also use some Euclidean geometry or topology as well.

Nowadays, the slickest proof of the Szemerédi-Trotter theorem is via the crossing number inequality (as discussed in Section 1.10 of *Structure and Randomness*), which ultimately relies on *Euler's famous formula* $|V| - |E| + |F| = 2$; thus in this argument it is *topology* which is the feature of Euclidean space which one is exploiting, and which is not present in the finite field setting. In this article, though, I would like to mention a different proof (closer in spirit to the original proof of Szemerédi-Trotter [**SzTr1983**], and closer still to the later paper [**ClEdGuShWe1990**]), based on the method of *cell decomposition*, which has proven to be a very flexible method in combinatorial incidence geometry. Here, the distinctive feature of Euclidean geometry one is exploiting is *convexity*, which again has no finite field analogue.

Roughly speaking, the idea is this. Using nothing more than the axiom that two points determine at most one line, one can obtain the bound

$$|I(P,L)| \ll |P||L|^{1/2} + |L|, \tag{3.29}$$

which is inferior to (3.28). (On the other hand, this estimate works in both Euclidean and finite field geometries, and is sharp in the latter case, as shown by the example given earlier.) Dually, the axiom that two lines determine at most one point gives the bound

$$|I(P,L)| \ll |L||P|^{1/2} + |P| \tag{3.30}$$

(or alternatively, one can use projective duality to interchange points and lines and deduce (3.30) from (3.29)).

An inspection of the proof of (3.29) shows that it is only expected to be sharp when the *bushes* $L_p := \{\ell \in L : \ell \ni p\}$ associated to each point $p \in P$ behave like "independent" subsets of L, so that there is no significant correlation between the bush L_p of one point and the bush of another point L_q.

However, in Euclidean space, we have the phenomenon that the bush of a point L_p is influenced by the region of space that p lies in. Clearly, if p lies in a set Ω (e.g. a convex polygon), then the only lines $\ell \in L$ that can contribute to L_p are those lines which pass through Ω. If Ω is a small convex region of space, one expects only a fraction of the lines in L to actually pass through Ω. As such, if p and q both lie in Ω, then L_p and L_q are compressed inside a smaller subset of L, namely the set of lines passing through Ω, and so should be more likely to intersect than if they were independent. This should lead to an improvement to (3.29) (and indeed, as we shall see below, ultimately leads to (3.28)).

More formally, the argument proceeds by applying the following lemma:

Lemma 3.6.1 (Cell decomposition)**.** *Let L be a finite collection of lines in $\mathbf{R}^2$, and let $r \geq 1$. Then it is possible to find a set R of $O(r)$ lines in the plane (which may or may not be in L), which subdivide $\mathbf{R}^2$ into $O(r^2)$ convex regions (or* cells*), such that the interior of each such cell is incident to at most $O(|L|/r)$ lines.*

The deduction of (3.28) from (3.29), (3.30) and Lemma 3.6.1 is very quick. Firstly we may assume we are in the range

$$|L|^{1/2} \ll |P| \ll |L|^2 \tag{3.31}$$

otherwise the bound (3.28) follows already from either (3.29) or (3.30) and some high-school algebra.

Let $r \geq 1$ be a parameter to be optimised later. We apply the cell decomposition to subdivide $\mathbf{R}^2$ into $O(r^2)$ open convex regions, plus a family R of $O(r)$ lines. Each of the $O(r^2)$ convex regions Ω has only $O(|L|/r)$ lines through it, and so by (3.29) contributes $O(|P \cap \Omega||L|^{1/2}/r^{1/2} + |L|/r)$ incidences. Meanwhile, on each of the lines ℓ in R used to perform this decomposition, there are at most $|L|$ transverse incidences (because each line in L distinct from ℓ can intersect ℓ at most once), plus all the incidences along ℓ itself. Putting all this together, one obtains

$$|I(P,L)| \leq |I(P, L \cap R)| + O(|P||L|^{1/2}/r^{1/2} + |L|r).$$

We optimise this by selecting $r \sim |P|^{2/3}/|L|^{1/3}$; from (3.31) we can ensure that $r \leq |L|/2$, so that $|L \cap R| \leq |L|/2$. One then obtains

$$|I(P,L)| \leq |I(P, L \cap R)| + O(|P|^{2/3}|L|^{2/3}).$$

We can iterate away the $L \cap R$ error (halving the number of lines each time) and sum the resulting geometric series to obtain (3.28).

It remains to prove (3.6.1). If one subdivides $\mathbf{R}^2$ using r arbitrary lines, one creates at most $O(r^2)$ cells (because each new line intersects the existing lines at most once, and so can create at most $O(r)$ distinct cells), and for a similar reason, every line in L visits at most r of these regions, and so by double counting one expects $O(|L|/r)$ lines per cell "on the average". The key difficulty is then to get $O(|L|/r)$ lines through *every* cell, not just on the average. It turns out that a probabilistic argument will almost work, but with a logarithmic loss (thus having $O(|L| \log |L|/r)$ lines per cell rather than $O(|L|/r)$); but with a little more work one can then iterate away this loss also. The arguments here are loosely based on those of [**ClEdGuShWe1990**]; a related (deterministic) decomposition also appears in [**SzTr1983**]. But I wish to focus here on the probabilistic approach.

It is also worth noting that the original (somewhat complicated) argument of Szemerédi-Trotter has been adapted to establish the analogue of (3.28) in the complex plane $\mathbf{C}^2$ by Toth[**To2005**], while the other known proofs of Szemerédi-Trotter, so far, have not been able to be extended to this setting (the Euler characteristic argument clearly

breaks down, as does any proof based on using lines to divide planes into half-spaces). So all three proofs have their advantages and disadvantages.

3.6.1. The trivial incidence estimate. We first give a quick proof of the trivial incidence bound (3.29). We have

$$|I(P,L)| = \sum_{\ell \in L} |P \cap \ell|$$

and thus by Cauchy-Schwarz

$$\sum_{\ell \in L} |P \cap \ell|^2 \geq \frac{|I(P,L)|^2}{|L|}.$$

On the other hand, observe that

$$\sum_{\ell \in L} |P \cap \ell|^2 - |P \cap \ell| = |\{(p,q,\ell) \in P \times P \times L : p \neq q; p, q \in \ell\}.$$

Because two distinct points p, q are incident to at most one line, the right-hand side is at most $|P|^2$, thus

$$\sum_{\ell \in L} |P \cap \ell|^2 \leq |I(P,L)| + |P|^2.$$

Comparing this with the Cauchy-Schwarz bound and using a little high-school algebra we obtain (3.29). A dual argument (swapping the role of lines and points) give (3.30).

A more informal proof of (3.29) can be given as follows. Suppose for contradiction that $|I(P,L)|$ was much larger than $|P||L|^{1/2} + |L|$. Since $|I(P,L)| = \sum_{p \in P} |L_p|$, this implies that that the $|L_p|$ are much larger than $|L|^{1/2}$ on the average. By the *birthday paradox*, one then expects two randomly chosen L_p, L_q to intersect in at least two places ℓ, ℓ'; but this would mean that two lines intersect in two points, a contradiction. The use of Cauchy-Schwarz in the rigorous argument given above can thus be viewed as an assertion that the average intersection of L_p and L_q is at least as large as what random chance predicts.

As mentioned in the introduction, we now see (intuitively, at least) that if nearby p, q are such that L_p, L_q are drawn from a smaller pool of lines than L, then their intersection is likely to be higher, and so one should be able to improve upon (3.29).

3.6.2. The probabilistic bound. Now we start proving Lemma 3.6.1. We can assume that $r < |L|$, since the claim is trivial otherwise (we just use all the lines in L to subdivide the plane, and there are no lines left in L to intersect any of the cells). Similarly we may assume that $r > 1$, and that $|L|$ is large. We can also perturb all the lines slightly and assume that the lines are in general position (no three are concurrent), as the general claim then follows from a limiting argument (note that this may send some of the cells to become empty). (Of course, the Szemerédi-Trotter theorem is quite easy under the assumption of general position, but this theorem is not our current objective right now.)

We use the *probabilistic method*, i.e. we construct R by some random recipe and aim to show that the conclusion of the lemma holds with positive probaility.

The most obvious approach would be to choose the r lines R at random from L, thus each line $\ell \in L$ has a probability of $r/|L|$ of lying in R. Actually, for technical reasons it is slightly better to use a Bernoulli process to select R, thus each line $\ell \in L$ lies in R with an *independent* probability of $r/|L|$. This can cause R to occasionally have size much larger than r, but this probability can be easily controlled (e.g. using the *Chernoff inequality*). So with high probability, R consists of $O(r)$ lines, which therefore carve out $O(r^2)$ cell. The remaining task is to show that each cell is incident to at most $O(|L|/r)$ lines from L.

Observe that each cell is a (possibly unbounded) polygon, whose edges come from lines in R. Note that (except in the degenerate case when R consists of at most one line, which we can ignore) any line ℓ which meets a cell in R, must intersect at least one of the edges of R. If we pretend for the moment that all cells have a bounded number of edges, it would then suffice to show that each edge of each cell was incident to $O(|L|/r)$ lines.

Let's see how this would go. Suppose that one line $\ell \in L$ was picked for the set R, and consider all the other lines in L that intersect ℓ; there are $O(|L|)$ of these lines ℓ', which (by the general position hypothesis) intersect ℓ at distinct points $\ell \cap \ell'$ on the line. If one of

these lines ℓ' intersecting ℓ is also selected for R, then the corresponding point $\ell \cap \ell'$ will become a vertex of one of the cells (indeed, it will be the vertex of four cells). Thus each of these points on ℓ has an independent probability of $r/|L|$ of becoming a vertex for a cell.

Now consider m consecutive such points on ℓ. The probability that they all fail to be chosen as cell vertices is $(1 - r/|L|)^m$; if $m = k|L|/r$, then this probability is $O(\exp(-k))$. Thus runs of much more than $|L|/r$ points without vertices are unlikely. In particular, setting $k = 100 \log |L|$, we see that the probability that any given $100|L| \log |L|/r$ consecutive points on any given line ℓ are skipped is $O(|L|^{-100})$. By the union bound, we thus see that with probability $1 - O(|L|^{-98})$, that *every* line ℓ has at most $O(|L| \log |L|/r)$ points between any two adjacent vertices. Or in other words, the edge of every cell is incident to at most $O(|L| \log |L|/r)$ lines from L. This yields Lemma 3.6.1 except for two things: firstly, the logarithmic loss of $O(\log |L|)$, and secondly, the assumption that each cell had only a bounded number of edges.

To fix the latter problem, we will have to modify the construction of R, allowing the use of some lines outside of L. First, we randomly rotate the plane so that none of the lines in L are vertical. Then we do the following modified construction: we select $O(r)$ lines from L as before, creating $O(r^2)$ cells, some of which may have a very large number of edges. But then for each cell, and each vertex in that cell, we draw a vertical line segment from that vertex (in either the up or down direction) to bisect the cell into two pieces. (If the vertex is on the far left or far right of the cell, we do nothing.) Doing this once for each vertex, we see that we have subdivided each of the old cells into a number of new cells, each of which have at most four sides (two vertical sides, and two non-vertical sides). So we have now achieved a bounded number of sides per cell. But what about the number of such cells? Well, each vertex of each cell is responsible for at most two subdivisions of one cell into two, and the number of such vertices is at most $O(r^2)$ (as they are formed by intersecting two lines from the original selection of $O(r)$ lines together), so the total number of cells is still $O(r^2)$.

Is it still true that each edge of each cell is incident to $O(|L| \log |L|/r)$ lines in L? We have already proven this (with high probability) for all the old edges - the ones that were formed from lines in L. But there are now some new edges, caused by dropping a vertical line segment from the intersection of two lines in L. But it is not hard to see that one can use much the same argument as before to see that with high probability, each of these line segments is incident to at most $O(|L| \log |L|/r)$ lines in L as desired.

Finally, we have to get rid of the logarithm. An inspection of the above arguments (and a use of the first moment method) reveals the following refinement: for any $k \geq 1$, there are expected to be at most $O(\exp(-k)r^2)$ cells which are incident to more than $Ck|L|/r$ lines, where C is an absolute constant. This is already enough to improve the $O(|L| \log |L|/r)$ bound slightly to $O(|L| \log r/r)$. But one can do even better by using Lemma 3.6.1 as an induction hypothesis, i.e. assume that for any smaller set L' of lines with $|L'| < |L|$, and any $r' \geq 1$, one can partition L' into at most $C_1(r')^2$ cells using at most $C_0 r'$ lines such that each cell is incident to at most $C_2|L'|/r'$ lines, where C_1, C_2, C_3 are absolute constants. (When using induction, asymptotic notation becomes quite dangerous to use, and it is much safer to start writing out the constants explicitly. To close the induction, one has to end up with the same constants C_0, C_1, C_2 as one started with.) For each k between C_2/C and $O(\log r)$ which is a power of two, one can apply the induction hypothesis to all the cells which are incident to between $Ck|L|/r$ and $2Ck|L|/r$ (with L' set equal to the lines in L incident to this cell, and r' set comparable to $2Ck$), and sum up (using the fact that $\sum_k k^2 \exp(-k)$ converges, especially if k is restricted to powers of two) to close the induction if the constants C_0, C_1, C_2 are chosen properly; we leave the details as an exercise.

Notes. This article first appeared at `terrytao.wordpress.com/2009/06/12`. Thanks to Oded and vedadi for corrections.

Jozsef Solymosi noted that there is still no good characterisation of the point-line configurations for which the Szemerédi-Trotter theorem is close to sharp; such a characterisation may well lead to improvements to a variety of bounds which are currently proven using this theorem.

Jordan Ellenberg raised the interesting possibility of using algebraic methods to attack the finite field analogue of this problem.

3.7. Benford's law, Zipf's law, and the Pareto distribution

A remarkable phenomenon in probability theory is that of *universality* - that many seemingly unrelated probability distributions, which ostensibly involve large numbers of unknown parameters, can end up converging to a universal law that may only depend on a small handful of parameters. One of the most famous examples of the universality phenomenon is the *central limit theorem*; another rich source of examples comes from *random matrix theory*, which is one of the areas of my own research.

Analogous universality phenomena also show up in *empirical* distributions - the distributions of a statistic X from a large population of "real-world" objects. Examples include *Benford's law*, *Zipf's law*, and the *Pareto distribution* (of which the *Pareto principle* or *80-20 law* is a special case). These laws govern the asymptotic distribution of many statistics X which

(i) take values as positive numbers;

(ii) range over many different orders of magnitude;

(iii) arise from a complicated combination of largely independent factors (with different samples of X arising from different independent factors); and

(iv) have not been artificially rounded, truncated, or otherwise constrained in size.

Examples here include the population of countries or cities, the frequency of occurrence of words in a language, the mass of astronomical objects, or the net worth of individuals or corporations. The laws are then as follows:

- **Benford's law:** For $k = 1, \ldots, 9$, the proportion of X whose first digit is k is approximately $\log_{10} \frac{k+1}{k}$. Thus, for instance, X should have a first digit of 1 about 30% of the time, but a first digit of 9 only about 5% of the time.

- **Zipf's law:** The n^{th} largest value of X should obey an approximate power law, i.e. it should be approximately $Cn^{-\alpha}$ for the first few $n = 1, 2, 3, \ldots$ and some parameters $C, \alpha > 0$. In many cases, α is close to 1.
- **Pareto distribution:** The proportion of X with at least m digits (before the decimal point), where m is above the median number of digits, should obey an approximate exponential law, i.e. be approximately of the form $c10^{-m/\alpha}$ for some $c, \alpha > 0$. Again, in many cases α is close to 1.

Benford's law and Pareto distribution are stated here for base 10, which is what we are most familiar with, but the laws hold for any base (after replacing all the occurrences of 10 in the above laws with the new base, of course). The laws tend to break down if the hypotheses (i)-(iv) are dropped. For instance, if the statistic X concentrates around its mean (as opposed to being spread over many orders of magnitude), then the *normal distribution* tends to be a much better model (as indicated by such results as the central limit theorem). If instead the various samples of the statistics are highly correlated with each other, then other laws can arise (for instance, the eigenvalues of a random matrix, as well as many empirically observed matrices, are correlated to each other, with the behaviour of the largest eigenvalues being governed by laws such as the *Tracy-Widom law* rather than Zipf's law, and the bulk distribution being governed by laws such as the *semicircular law* rather than the normal or Pareto distributions).

To illustrate these laws, let us take as a data set the populations of 235 countries and regions of the world in 2007[7]. This is a relatively small sample (cf. Section 1.9 of *Poincaré's Legacies, Vol. I*), but is already enough to discern these laws in action. For instance, here is how the data set tracks with Benford's law (rounded to three significant figures):

[7]This data was taken from the CIA world factbook at `http://www.umsl.edu/services/govdocs/wofact2007/index.html`; I have put the raw data at `http://spreadsheets.google.com/pub?key=rj_3TkLJrrVuvOXkijCHelQ&output=html`.

k	Countries	Number	Benford
1	Angola, Anguilla, Aruba, Bangladesh, Belgium, Botswana, Brazil, Burkina Faso, Cambodia, Cameroon, Chad, Chile, China, Christmas Island, Cook Islands, Cuba, Czech Republic, Ecuador, Estonia, Gabon, (The) Gambia, Greece, Guam, Guatemala, Guinea-Bissau, India, Japan, Kazakhstan, Kiribati, Malawi, Mali, Mauritius, Mexico, (Federated States of) Micronesia, Nauru, Netherlands, Niger, Nigeria, Niue, Pakistan, Portugal, Russia, Rwanda, Saint Lucia, Saint Vincent and the Grenadines, Senegal, Serbia, Swaziland, Syria, Timor-Leste (East-Timor), Tokelau, Tonga, Trinidad and Tobago, Tunisia, Tuvalu, (U.S.) Virgin Islands, Wallis and Futuna, Zambia, Zimbabwe	59 (25.1%)	71 (30.1%)
2	Armenia, Australia, Barbados, British Virgin Islands, Cote d'Ivoire, French Polynesia, Ghana, Gibraltar, Indonesia, Iraq, Jamaica, (North) Korea, Kosovo, Kuwait, Latvia, Lesotho, Macedonia, Madagascar, Malaysia, Mayotte, Mongolia, Mozambique, Namibia, Nepal, Netherlands Antilles, New Caledonia, Norfolk Island, Palau, Peru, Romania, Saint Martin, Samoa, San Marino, Sao Tome and Principe, Saudi Arabia, Slovenia, Sri Lanka, Svalbard, Taiwan, Turks and Caicos Islands, Uzbekistan, Vanuatu, Venezuela, Yemen	44 (18.7%)	41 (17.6%)
3	Afghanistan, Albania, Algeria, (The) Bahamas, Belize, Brunei, Canada, (Rep. of the) Congo, Falkland Islands, Iceland, Kenya, Lebanon, Liberia, Liechtenstein, Lithuania, Maldives, Mauritania, Monaco, Morocco, Oman, (Occupied) Palestinian Territory, Panama, Poland, Puerto Rico, Saint Kitts and Nevis, Uganda, United States of America, Uruguay, Western Sahara	29 (12.3%)	29 (12.5%)
4	Argentina, Bosnia and Herzegovina, Burma (Myanmar), Cape Verde, Cayman Islands, Central African Republic, Colombia, Costa Rica, Croatia, Faroe Islands, Georgia, Ireland, (South) Korea, Luxembourg, Malta, Moldova, New Zealand, Norway, Pitcairn Islands, Singapore, South Africa, Spain, Sudan, Suriname, Tanzania, Ukraine, United Arab Emirates	27 (11.4%)	22 (9.7%)
5	(Macao SAR) China, Cocos Islands, Denmark, Djibouti, Eritrea, Finland, Greenland, Italy, Kyrgyzstan, Montserrat, Nicaragua, Papua New Guinea, Slovakia, Solomon Islands, Togo, Turkmenistan	16 (6.8%)	19 (7.9%)
6	American Samoa, Bermuda, Bhutan, (Dem. Rep. of the) Congo, Equatorial Guinea, France, Guernsey, Iran, Jordan, Laos, Libya, Marshall Islands, Montenegro, Paraguay, Sierra Leone, Thailand, United Kingdom	17 (7.2%)	16 (6.7%)
7	Bahrain, Bulgaria, (Hong Kong SAR) China, Comoros, Cyprus, Dominica, El Salvador, Guyana, Honduras, Israel, (Isle of) Man, Saint Barthelemy, Saint Helena, Saint Pierre and Miquelon, Switzerland, Tajikistan, Turkey	17 (7.2%)	14 (5.8%)
8	Andorra, Antigua and Barbuda, Austria, Azerbaijan, Benin, Burundi, Egypt, Ethiopia, Germany, Haiti, Holy See (Vatican City), Northern Mariana Islands, Qatar, Seychelles, Vietnam	15 (6.4%)	12 (5.1%)
9	Belarus, Bolivia, Dominican Republic, Fiji, Grenada, Guinea, Hungary, Jersey, Philippines, Somalia, Sweden	11 (4.5%)	11 (4.6%)

Here is how the same data tracks Zipf's law for the first twenty values of n, with the parameters $C \approx 1.28 \times 10^9$ and $\alpha \approx 1.03$ (selected by log-linear regression), again rounding to three significant figures:

n	Country	Population	Zipf prediction	Error
1	China	1,330,000,000	1,280,000,000	+4.1%
2	India	1,150,000,000	626,000,000	+83.5%
3	USA	304,000,000	412,000,000	−26.3%
4	Indonesia	238,000,000	307,000,000	−22.5%
5	Brazil	196,000,000	244,000,000	−19.4%
6	Pakistan	173,000,000	202,000,000	−14.4%
7	Bangladesh	154,000,000	172,000,000	−10.9%
8	Nigeria	146,000,000	150,000,000	−2.6%
9	Russia	141,000,000	133,000,000	+5.8%
10	Japan	128,000,000	120,000,000	+6.7%
11	Mexico	110,000,000	108,000,000	+1.7%
12	Philippines	96,100,000	98,900,000	−2.9%
13	Vietnam	86,100,000	91,100,000	−5.4%
14	Ethiopia	82,600,000	84,400,000	−2.1%
15	Germany	82,400,000	78,600,000	+4.8%
16	Egypt	81,700,000	73,500,000	+11.1%
17	Turkey	71,900,000	69,100,000	+4.1%
18	Congo	66,500,000	65,100,000	+2.2%
19	Iran	65,900,000	61,600,000	+6.9%
20	Thailand	65,500,000	58,400,000	+12.1%

As one sees, Zipf's law is not particularly precise at the extreme edge of the statistics (when n is very small), but becomes reasonably accurate (given the small sample size, and given that we are fitting twenty data points using only two parameters) for moderate sizes of n.

This data set has too few scales in base 10 to illustrate the Pareto distribution effectively - over half of the country populations are either seven or eight digits in that base. But if we instead work in base 2, then country populations range in a decent number of scales (the majority of countries have population between 2^{23} and 2^{32}), and we begin to see the law emerge, where m is now the number of digits in binary, the best-fit parameters are $\alpha \approx 1.18$ and $c \approx 1.7 \times 2^{26}/235$:

m	Countries with m-binary-digit populations	Number	Pareto
31	China, India	2	1
30		2	2
29	United States of America	3	5
28	Indonesia, Brazil, Pakistan, Bangladesh, Nigeria, Russia	9	8
27	Japan, Mexico, Philippines, Vietnam, Ethiopia, Germany , Egypt, Turkey	17	15
26	(Dem. Rep. of the) Congo, Iran, Thailand, France, United Kingdom, Italy, South Africa, (South) Korea, Burma (Myanmar), Ukraine, Colombia, Spain, Argentina, Sudan, Tanzania, Poland, Kenya, Morocco, Algeria	36	27
25	Canada, Afghanistan, Uganda, Nepal, Peru, Iraq, Saudi Arabia, Uzbekistan, Venezuela, Malaysia, (North) Korea, Ghana, Yemen, Taiwan, Romania, Mozambique, Sri Lanka, Australia, Cote d'Ivoire, Madagascar, Syria, Cameroon	58	49
24	Netherlands, Chile, Kazakhstan, Burkina Faso, Cambodia, Malawi, Ecuador, Niger, Guatemala, Senegal, Angola, Mali, Zambia, Cuba, Zimbabwe, Greece, Portugal, Belgium, Tunisia, Czech Republic, Rwanda, Serbia, Chad, Hungary, Guinea, Belarus, Somalia, Dominican Republic, Bolivia, Sweden, Haiti, Burundi, Benin	91	88
23	Austria, Azerbaijan, Honduras, Switzerland, Bulgaria, Tajikistan, Israel, El Salvador, (Hong Kong SAR) China, Paraguay, Laos, Sierra Leone, Jordan, Libya, Papua New Guinea, Togo, Nicaragua, Eritrea, Denmark, Slovakia, Kyrgyzstan, Finland, Turkmenistan, Norway, Georgia, United Arab Emirates, Singapore, Bosnia and Herzegovina, Croatia, Central African Republic, Moldova, Costa Rica	123	159

Thus, with each new scale, the number of countries introduced increases by a factor of a little less than 2, on the average. This approximate doubling of countries with each new scale begins to falter at about the population 2^{23} (i.e. at around 4 million), for the simple reason that one has begun to run out of countries. (Note that the median-population country in this set, Singapore, has a population with 23 binary digits.)

These laws are not merely interesting statistical curiosities; for instance, Benford's law is often used to help detect fraudulent statistics (such as those arising from accounting fraud), as many such statistics are invented by choosing digits at random, and will therefore deviate significantly from Benford's law. (This is nicely discussed in [**Ma1999**].) In a somewhat analogous spirit, Zipf's law and the Pareto distribution can be used to mathematically test various models of real-world systems (e.g. formation of astronomical objects, accumulation of wealth, population growth of countries, etc.), without necessarily having to fit all the parameters of that model with the actual data.

Being empirically observed phenomena rather than abstract mathematical facts, Benford's law, Zipf's law, and the Pareto distribution cannot be "proved" the same way a mathematical theorem can be proved. However, one can still *support* these laws mathematically in a number of ways, for instance showing how these laws are compatible with each other, and with other plausible hypotheses on the source of the data. In this post I would like to describe a number of ways (both technical and non-technical) in which one can do this; these arguments do not fully explain these laws (in particular, the empirical fact that the exponent α in Zipf's law or the Pareto distribution is often close to 1 is still quite a mysterious phenomenon), and do not always have the same universal range of applicability as these laws seem to have, but I hope that they do demonstrate that these laws are not completely arbitrary, and ought to have a satisfactory basis of mathematical support.

3.7.1. Scale invariance. One consistency check that is enjoyed by all of these laws is that of *scale invariance* - they are invariant under rescalings of the data (for instance, by changing the units).

For example, suppose for sake of argument that the country populations X of the world in 2007 obey Benford's law, thus for instance about 30.7% of the countries have population with first digit 1, 17.6% have population with first digit 2, and so forth. Now, imagine that several decades in the future, say in 2067, all of the countries in the world double their population, from X to a new population $\tilde{X} := 2X$. (This makes the somewhat implausible assumption that growth rates

are uniform across all countries; I will talk about what happens when one omits this hypothesis later.) To further simplify the experiment, suppose that no countries are created or dissolved during this time period. What happens to Benford's law when passing from X to $\tilde{X}$?

The key observation here, of course, is that the first digit of X is linked to the first digit of $\tilde{X} = 2X$. If, for instance, the first digit of X is 1, then the first digit of $\tilde{X}$ is either 2 or 3; conversely, if the first digit of $\tilde{X}$ is 2 or 3, then the first digit of X is 1. As a consequence, the proportion of X's with first digit 1 is equal to the proportion of $\tilde{X}$'s with first digit 2, plus the proportion of $\tilde{X}$'s with first digit 3. This is consistent with Benford's law holding for both X and $\tilde{X}$, since

$$\log_{10} \frac{2}{1} = \log_{10} \frac{3}{2} + \log_{10} \frac{4}{3} (= \log_{10} \frac{4}{2})$$

(or numerically, $30.7\% = 17.6\% + 12.5\%$ after rounding). Indeed one can check the other digit ranges also and that conclude that Benford's law for X is compatible with Benford's law for $\tilde{X}$; to pick a contrasting example, a uniformly distributed model in which each digit from 1 to 9 is the first digit of X occurs with probability 1/9 totally fails to be preserved under doubling.

One can be even more precise. Observe (through telescoping series) that Benford's law implies that

$$\mathbf{P}(\alpha 10^n \leq X < \beta 10^n \text{ for some integer } n) = \log_{10} \frac{\beta}{\alpha} \tag{3.32}$$

for all integers $1 \leq \alpha \leq \beta < 10$, where the left-hand side denotes the proportion of data for which X lies between $\alpha 10^n$ and $\beta 10^n$ for some integer n. Suppose now that we generalise Benford's law to the *continuous Benford's law*, which asserts that (3.32) is true for all *real numbers* $1 \leq \alpha \leq \beta < 10$. Then it is not hard to show that a statistic X obeys the continuous Benford's law if and only if its dilate $\tilde{X} = 2X$ does, and similarly with 2 replaced by any other constant growth factor. (This is easiest seen by observing that (3.32) is equivalent to asserting that the fractional part of $\log_{10} X$ is uniformly distributed.) In fact, the continuous Benford law is the *only* distribution for the quantities on the left-hand side of (3.32) with this scale-invariance property; this fact is a special case of the general fact that Haar measures are unique (see Section 1.12).

It is also easy to see that Zipf's law and the Pareto distribution also enjoy this sort of scale-invariance property, as long as one generalises the Pareto distribution

$$\mathbf{P}(X \geq 10^m) = c10^{-m/\alpha} \tag{3.33}$$

from integer m to real m, just as with Benford's law. Once one does that, one can phrase the Pareto distribution law independently of any base as

$$\mathbf{P}(X \geq x) = cx^{-1/\alpha} \tag{3.34}$$

for any x much larger than the median value of X, at which point the scale-invariance is easily seen.

One may object that the above thought-experiment was too idealised, because it assumed uniform growth rates for all the statistics at once. What happens if there are non-uniform growth rates? To keep the computations simple, let us consider the following toy model, where we take the same 2007 population statistics X as before, and assume that half of the countries (the "high-growth" countries) will experience a population doubling by 2067, while the other half (the "zero-growth" countries) will keep their population constant, thus the 2067 population statistic $\tilde{X}$ is equal to $2X$ half the time and X half the time. (We will assume that our sample sizes are large enough that the *law of large numbers* kicks in, and we will therefore ignore issues such as what happens to this "half the time" if the number of samples is odd.) Furthermore, we make the plausible but crucial assumption that the event that a country is a high-growth or a zero-growth country is *independent* of the first digit of the 2007 population; thus, for instance, a country whose population begins with 3 is assumed to be just as likely to be high-growth as one whose population begins with 7.

Now let's have a look again at the proportion of countries whose 2067 population $\tilde{X}$ begins with either 2 or 3. There are exactly two ways in which a country can fall into this category: either it is a zero-growth country whose 2007 population X also began with either 2 or 3, or ot was a high-growth country whose population in 2007 began with 1. Since all countries have a probability 1/2 of being high-growth regardless of the first digit of their population, we conclude

the identity

$$\mathbf{P}(\tilde{X} \text{ has first digit } 2,3) = \frac{1}{2}\mathbf{P}(X \text{ has first digit } 2,3) \tag{3.35}$$

$$+\frac{1}{2}\mathbf{P}(X \text{ has first digit } 1)$$

which is once again compatible with Benford's law for $\tilde{X}$ since

$$\log_{10}\frac{4}{2} = \frac{1}{2}\log_{10}\frac{4}{2} + \frac{1}{2}\log\frac{2}{1}.$$

More generally, it is not hard to show that if X obeys the continuous Benford's law (3.32), and one multiplies X by some positive multiplier Y which is independent of the first digit of X (and, *a fortiori*, is independent of the fractional part of $\log_{10} X$), one obtains another quantity $\tilde{X} = XY$ which also obeys the continuous Benford's law. (Indeed, we have already seen this to be the case when Y is a deterministic constant, and the case when Y is random then follows simply by conditioning Y to be fixed.)

In particular, we see an absorptive property of Benford's law: if X obeys Benford's law, and Y is any positive statistic independent of X, then the product $\tilde{X} = XY$ also obeys Benford's law - *even if Y did not obey this law*. Thus, if a statistic is the product of many independent factors, then it only requires a single factor to obey Benford's law in order for the whole product to obey the law also. For instance, the population of a country is the product of its area and its population density. Assuming that the population density of a country is independent of the area of that country (which is not a completely reasonable assumption, but let us take it for the sake of argument), then we see that Benford's law for the population would follow if just one of the area or population density obeyed this law. It is also clear that Benford's law is the only distribution with this absorptive property (if there was another law with this property, what would happen if one multiplied a statistic with that law with an independent statistic with Benford's law?). Thus we begin to get a glimpse as to why Benford's law is universal for quantities which are the product of many separate factors, in a manner that no other law could be.

As an example: for any given number N, the uniform distribution from 1 to N does not obey Benford's law; for instance, if one picks a random number from 1 to $999,999$ then each digit from 1 to 9 appears as the first digit with an equal probability of $1/9$ each. However, if N is not fixed, but instead obeys Benford's law, then a random number selected from 1 to N also obeys Benford's law (ignoring for now the distinction between continuous and discrete distributions), as it can be viewed as the product of N with an independent random number selected from between 0 and 1.

Actually, one can say something even stronger than the absorption property. Suppose that the continuous Benford's law (3.32) for a statistic X did not hold exactly, but instead held with some accuracy $\varepsilon > 0$, thus

$$\begin{aligned} \log_{10} \frac{\beta}{\alpha} - \varepsilon \leq \mathbf{P}(\alpha 10^n \leq X < \beta 10^n \text{ for some integer } n) \\ \leq \log_{10} \frac{\beta}{\alpha} + \varepsilon \end{aligned} \tag{3.36}$$

for all $1 \leq \alpha \leq \beta < 10$. Then it is not hard to see that any dilated statistic, such as $\tilde{X} = 2X$, or more generally $\tilde{X} = XY$ for any fixed deterministic Y, also obeys (3.36) with exactly the same accuracy ε. But now suppose one uses a variable multiplier; for instance, suppose one uses the model discussed earlier in which $\tilde{X}$ is equal to $2X$ half the time and X half the time. Then the relationship between the distribution of the first digit of $\tilde{X}$ and the first digit of X is given by formulae such as (3.35). Now, in the right-hand side of (3.35), each of the two terms $\mathbf{P}(X \text{ has first digit } 2,3)$ and $\mathbf{P}(X \text{ has first digit } 1)$ differs from the Benford's law predictions of $\log_{10} \frac{4}{2}$ and $\log_{10} \frac{2}{1}$ respectively by at most ε. Since the left-hand side of (3.35) is the average of these two terms, it also differs from the Benford law prediction by at most ε. But the averaging opens up an opportunity for cancelling; for instance, an overestimate of $+\varepsilon$ for $\mathbf{P}(X \text{ has first digit } 2,3)$ could cancel an underestimate of $-\varepsilon$ for $\mathbf{P}(X \text{ has first digit } 1)$ to produce a spot-on prediction for $\tilde{X}$. Thus we see that variable multipliers (or variable growth rates) not only preserve Benford's law, but in fact *stabilise* it by averaging out the errors. In fact, if one started with a distribution which did not initially obey Benford's law, and then started applying some variable (and independent) growth rates

to the various samples in the distribution, then under reasonable assumptions one can show that the resulting distribution will converge to Benford's law over time. This helps explain the universality of Benford's law for statistics such as populations, for which the independent variable growth law is not so unreasonable (at least, until the population hits some maximum capacity threshold).

Note that the independence property is crucial; if for instance population growth always slowed down for some inexplicable reason to a crawl whenever the first digit of the population was 6, then there would be a noticeable deviation from Benford's law, particularly in digits 6 and 7, due to this growth bottleneck. But this is not a particularly plausible scenario (being somewhat analogous to *Maxwell's demon* in thermodynamics).

The above analysis can also be carried over to some extent to the Pareto distribution and Zipf's law; if a statistic X obeys these laws approximately, then after multiplying by an independent variable Y, the product $\tilde{X} = XY$ will obey the same laws with equal or higher accuracy, so long as Y is small compared to the number of scales that X typically ranges over. (One needs a restriction such as this because the Pareto distribution and Zipf's law must break down below the median. Also, Zipf's law loses its stability at the very extreme end of the distribution, because there are no longer enough samples for the law of large numbers to kick in; this is consistent with the empirical observation that Zipf's law tends to break down *in extremis*.) These laws are also stable under other multiplicative processes, for instance if some fraction of the samples in X spontaneously split into two smaller pieces, or conversely if two samples in X spontaneously merge into one; as before, the key is that the occurrence of these events should be independent of the actual size of the objects being split. If one considers a generalisation of the Pareto or Zipf law in which the exponent α is not fixed, but varies with n or k, then the effect of these sorts of multiplicative changes is to blur and average together the various values of α, thus "flattening" the α curve over time and making the distribution approach Zipf's law and/or the Pareto distribution. This helps explain why α eventually becomes constant;

however, I do not have a good explanation as to why α is often close to 1.

3.7.2. Compatibility between laws. Another mathematical line of support for Benford's law, Zipf's law, and the Pareto distribution are that the laws are highly compatible with each other. For instance, Zipf's law and the Pareto distribution are formally equivalent: if there are N samples of X, then applying (3.34) with x equal to the n^{th} largest value X_n of X gives

$$\frac{n}{N} = \mathbf{P}(X \geq X_n) = cX_n^{-1/\alpha}$$

which implies Zipf's law $X_n = Cn^{-\alpha}$ with $C := (Nc)^\alpha$. Conversely one can deduce the Pareto distribution from Zipf's law. These deductions are only formal in nature, because the Pareto distribution can only hold exactly for continuous distributions, whereas Zipf's law only makes sense for discrete distributions, but one can generate more rigorous variants of these deductions without much difficulty.

In some literature, Zipf's law is applied primarily near the extreme edge of the distribution (e.g. the top 0.1% of the sample space), whereas the Pareto distribution in regions closer to the bulk (e.g. between the top 0.1% and and top 50%). But this is mostly a difference of degree rather than of kind, though in some cases (such as with the example of the 2007 country populations data set) the exponent α for the Pareto distribtion in the bulk can differ slightly from the exponent for Zipf's law at the extreme edge.

The relationship between Zipf's law or the Pareto distribution and Benford's law is more subtle. For instance Benford's law predicts that the proportion of X with initial digit 1 should equal the proportion of X with initial digit 2 or 3. But if one formally uses the Pareto distribution (3.34) to compare those X between 10^m and 2×10^m, and those X between 2×10^m and 4×10^m, it seems that the former is larger by a factor of $2^{1/\alpha}$, which upon summing by m appears inconsistent with Benford's law (unless α is extremely large). A similar inconsistency is revealed if one uses Zipf's law instead.

However, the fallacy here is that the Pareto distribution (or Zipf's law) does not apply on the entire range of X, but only on the upper

tail region when X is significantly higher than the median; it is a law for the *outliers* of X only. In contrast, Benford's law concerns the behaviour of *typical* values of X; the behaviour of the top 0.1% is of negligible significance to Benford's law, though it is of prime importance for Zipf's law and the Pareto distribution. Thus the two laws describe different components of the distribution and thus complement each other. Roughly speaking, Benford's law asserts that the bulk distribution of $\log_{10} X$ is locally uniform at unit scales, while the Pareto distribution (or Zipf's law) asserts that the tail distribution of $\log_{10} X$ decays exponentially. Note that Benford's law only describes the fine-scale behaviour of the bulk distribution; the coarse-scale distribution can be a variety of distributions (e.g. log-gaussian).

Notes. This article first appeared at `terrytao.wordpress.com/2009/07/03`. Thanks to Kevin O'Bryant for corrections. Several other derivations of Benford's law and the Pareto distribution, such as those relying on max-entropy principles, were also discussed in the comments.

3.8. Selberg's limit theorem for the Riemann zeta function on the critical line

The *Riemann zeta function* $\zeta(s)$, defined for $\mathrm{Re}(s) > 1$ by

$$\zeta(s) := \sum_{n=1}^{\infty} \frac{1}{n^s} \tag{3.37}$$

and then continued *meromorphically* to other values of s by *analytic continuation*, is a fundamentally important function in analytic number theory, as it is connected to the primes $p = 2, 3, 5, \ldots$ via the *Euler product formula*

$$\zeta(s) = \prod_p (1 - \frac{1}{p^s})^{-1} \tag{3.38}$$

(for $\mathrm{Re}(s) > 1$, at least), where p ranges over primes. (The equivalence between (3.37) and (3.38) is essentially the *generating function* version of the *fundamental theorem of arithmetic*.) The function ζ has a pole at 1 and a number of zeroes ρ. A formal application of the

factor theorem gives

$$\zeta(s) = \frac{1}{s-1} \prod_{\rho} (s-\rho) \times \dots \tag{3.39}$$

where ρ ranges over zeroes of ζ, and we will be vague about what the $\dots$ factor is, how to make sense of the infinite product, and exactly which zeroes of ζ are involved in the product. Equating (3.38) and (3.39) and taking logarithms gives the formal identity

$$-\log \zeta(s) = \sum_p \log(1 - \frac{1}{p^s}) = \log(s-1) - \sum_\rho \log(s-\rho) + \dots; \tag{3.40}$$

using the Taylor expansion

$$\log(1 - \frac{1}{p^s}) = -\frac{1}{p^s} - \frac{1}{2p^{2s}} - \frac{1}{3p^{3s}} - \dots \tag{3.41}$$

and differentiating the above identity in s yields the formal identity

$$-\frac{\zeta'(s)}{\zeta(s)} = \sum_n \frac{\Lambda(n)}{n^s} = \frac{1}{s-1} - \sum_\rho \frac{1}{s-\rho} + \dots \tag{3.42}$$

where $\Lambda(n)$ is the *von Mangoldt function*, defined to be $\log p$ when n is a power of a prime p, and zero otherwise. Thus we see that the behaviour of the primes (as encoded by the von Mangoldt function) is intimately tied to the distribution of the zeroes ρ. For instance, if we knew that the zeroes were far away from the axis $\mathrm{Re}(s) = 1$, then we would heuristically have

$$\sum_n \frac{\Lambda(n)}{n^{1+it}} \approx \frac{1}{it}$$

for real t. On the other hand, the integral test suggests that

$$\sum_n \frac{1}{n^{1+it}} \approx \frac{1}{it}$$

and thus we see that $\frac{\Lambda(n)}{n}$ and $\frac{1}{n}$ have essentially the same (multiplicative) *Fourier transform*:

$$\sum_n \frac{\Lambda(n)}{n^{1+it}} \approx \sum_n \frac{1}{n^{1+it}}.$$

Inverting the Fourier transform (or performing a contour integral closely related to the inverse Fourier transform), one is led to the *prime number theorem*

$$\sum_{n \leq x} \Lambda(n) \approx \sum_{n \leq x} 1.$$

In fact, the standard proof of the prime number theorem basically proceeds by making all of the above formal arguments precise and rigorous.

Unfortunately, we don't know as much about the zeroes ρ of the zeta function (and hence, about the ζ function itself) as we would like. The *Riemann hypothesis* (RH) asserts that all the zeroes (except for the "trivial" zeroes at the negative even numbers) lie on the *critical line* $\mathrm{Re}(s) = 1/2$; this hypothesis would make the error terms in the above proof of the prime number theorem significantly more accurate. Furthermore, the stronger *GUE hypothesis* asserts in addition to RH that the local distribution of these zeroes on the critical line should behave like the local distribution of the eigenvalues of a random matrix drawn from the *gaussian unitary ensemble* (GUE). I will not give a precise formulation of this hypothesis here, except to say that the adjective "local" in the context of distribution of zeroes ρ means something like "at scale $O(1/\log T)$ when $\mathrm{Im}(s) = O(T)$".

Nevertheless, we do know some reasonably non-trivial facts about the zeroes ρ and the zeta function ζ, either unconditionally, or assuming RH (or GUE). Firstly, there are no zeroes for $\mathrm{Re}(s) > 1$ (as one can already see from the convergence of the Euler product (3.38) in this case) or for $\mathrm{Re}(s) = 1$ (this is trickier, relying on (3.42) and the elementary observation that

$$\mathrm{Re}(3\frac{\Lambda(n)}{n^\sigma} + 4\frac{\Lambda(n)}{n^{\sigma+it}} + \frac{\Lambda(n)}{n^{\sigma+2it}}) = 2\frac{\Lambda(n)}{n^\sigma}(1+\cos(t\log n))^2$$

is non-negative for $\sigma > 1$ and $t \in \mathbf{R}$); from the *functional equation*

$$\pi^{-s/2}\Gamma(s/2)\zeta(s) = \pi^{-(1-s)/2}\Gamma((1-s)/2)\zeta(1-s)$$

(which can be viewed as a consequence of the *Poisson summation formula*, see e.g. Section 1.5 of *Poincaré's Legacies, Vol. I*) we know that there are no zeroes for $\mathrm{Re}(s) \leq 0$ either (except for the trivial zeroes at negative even integers, corresponding to the poles of the

Gamma function). Thus all the non-trivial zeroes lie in the *critical strip* $0 < \mathrm{Re}(s) < 1$.

We also know that there are infinitely many non-trivial zeroes, and can approximately count how many zeroes there are in any large bounded region of the critical strip. For instance, for large T, the number of zeroes ρ in this strip with $\mathrm{Im}(\rho) = T + O(1)$ is $O(\log T)$. This can be seen by applying (3.42) to $s = 2+iT$ (say); the trivial zeroes at the negative integers end up giving a contribution of $O(\log T)$ to this sum (this is a heavily disguised variant of *Stirling's formula*, as one can view the trivial zeroes as essentially being poles of the Gamma function), while the $\frac{1}{s-1}$ and $\ldots$ terms end up being negligible (of size $O(1)$), while each non-trivial zero ρ contributes a term which has a non-negative real part, and furthermore has size comparable to 1 if $\mathrm{Im}(\rho) = T + O(1)$. (Here I am glossing over a technical renormalisation needed to make the infinite series in (3.42) converge properly.) Meanwhile, the left-hand side of (3.42) is absolutely convergent for $s = 2+iT$ and of size $O(1)$, and the claim follows. A more refined version of this argument shows that the number of non-trivial zeroes with $0 \leq \mathrm{Im}(\rho) \leq T$ is $\frac{T}{2\pi} \log \frac{T}{2\pi} - \frac{T}{2\pi} + O(\log T)$, but we will not need this more precise formula here. (A fair fraction - at least 40%, in fact - of these zeroes are known to lie on the critical line; see [**Co1989**].)

Another thing that we happen to know is how the *magnitude* $|\zeta(1/2 + it)|$ of the zeta function is distributed as $t \to \infty$; it turns out to be *log-normally* distributed with log-variance about $\frac{1}{2} \log \log t$. More precisely, we have the following result of Selberg:

Theorem 3.8.1. *Let T be a large number, and let t be chosen uniformly at random from between T and $2T$ (say). Then the distribution of $\frac{1}{\sqrt{\frac{1}{2}\log\log T}} \log |\zeta(1/2 + it)|$ converges (*in distribution*) to the normal distribution $N(0, 1)$.*

To put it more informally, $\log |\zeta(1/2+it)|$ behaves like $\sqrt{\frac{1}{2}\log\log t} \times N(0,1)$ plus lower order terms for "typical" large values of t. (Zeroes ρ of ζ are, of course, certainly not typical, but one can show that one can usually stay away from these zeroes.) In fact, Selberg showed a slightly more precise result, namely that for any fixed $k \geq 1$, the k^{th}

moment of $\frac{1}{\sqrt{\frac{1}{2}\log\log T}}\log|\zeta(1/2+it)|$ converges to the k^{th} moment of $N(0,1)$.

Remarkably, Selberg's result does not need RH or GUE, though it is certainly consistent with such hypotheses. (For instance, the determinant of a GUE matrix asymptotically obeys a remarkably similar log-normal law to that given by Selberg's theorem.) Indeed, the net effect of these hypotheses only affects some error terms in $\log|\zeta(1/2+it)|$ of magnitude $O(1)$, and are thus asymptotically negligible compared to the main term, which has magnitude about $O(\sqrt{\log\log T})$. So Selberg's result, while very pretty, manages to finesse the question of what the zeroes ρ of ζ are actually doing - he makes the primes do most of the work, rather than the zeroes.

Selberg never actually published the above result, but it is reproduced in a number of places (e.g. in [**Jo1986**] or [**La1996**]). As with many other results in analytic number theory, the actual details of the proof can get somewhat technical; but I would like to record here (partly for my own benefit) an informal sketch of some of the main ideas in the argument.

3.8.1. Informal overview of argument. The first step is to get a usable (approximate) formula for $\log|\zeta(s)|$. On the one hand, from the second part of (3.40) one has

$$-\log|\zeta(s)| = \log|s-1| - \sum_\rho \log|s-\rho| + \dots. \tag{3.43}$$

This formula turns out not to be directly useful because it requires one to know much more about the distribution of the zeroes ρ than we currently possess. On the other hand, from the first part of (3.40) and (3.41) one also has the formula

$$\log|\zeta(s)| = \sum_p \operatorname{Re}\frac{1}{p^s} + \dots. \tag{3.44}$$

This formula also turns out not to be directly useful, because it requires one to know much more about the distribution of the primes p than we currently possess.

However, it turns out that we can "split the difference" between (3.43), (3.44), and get a formula for $\log|\zeta(s)|$ which involves some

zeroes ρ and some primes p, in a manner that one can control them both. Roughly speaking, the formula looks like this[8]:

(3.45)
$$\log|\zeta(s)| = \sum_{p\leq T^\varepsilon} \mathrm{Re}\,\frac{1}{p^s} + O(\sum_{\rho=s+O(1/\log T)} 1 + |\log\frac{|s-\rho|}{1/\log T}|) + \dots$$

for $s = 1/2 + it$ and $t = O(T)$, where ε is a small parameter that we can choose (e.g. $\varepsilon = 0.01$); thus we have localised the prime sum to the primes p of size $O(T^{O(1)})$, and the zero sum to those zeroes at a distance $O(1/\log T)$ from s.

It turns out that all of these expressions can be controlled. The error term coming from the zeroes (as well as the $\dots$ error term) turn out to be of size $O(1)$ for most values of t, so are a lower order term. (As mentioned before, it is this error term that would be better controlled if one had RH or GUE, but this is not necessary to establish Selberg's result.) The main term is the one coming from the primes.

We can heuristically argue as follows. The expression $X_p := \mathrm{Re}\,\frac{1}{p^s} = \frac{1}{\sqrt{p}}\cos(t\log p)$, for t ranging between T and $2T$, is a random variable of mean zero and variance approximately $\frac{1}{2p}$ (if $p \leq T^\varepsilon$ and ε is small). Making the heuristic assumption that the X_p behave as if they were independent, the *central limit theorem* then suggests that the sum $\sum_{p\leq T^\varepsilon} X_p$ should behave like a normal distribution of mean zero and variance $\sum_{p\leq T^\varepsilon}\frac{1}{2p}$. But the claim now follows from the classical estimate

$$\sum_{p\leq x}\frac{1}{p} = \log\log x + O(1)$$

(which follows from the prime number theorem, but can also be deduced from the formula (3.44) for $s = 1 + O(1/\log x)$, using the fact that ζ has a simple pole at 1).

To summarise, there are three main tasks to establish Selberg's theorem:

(1) Establish a formula along the lines of (3.45);

[8]This is an oversimplification; there is a "tail" coming from those zeroes that are more distant from s than $O(1/\log T)$, and also one has to smooth out the sum in p a little bit, and allow the implied constants in the $O()$ notation to depend on ε, but let us ignore these technical issues here, as well as the issue of what exactly is hiding in the $\dots$ error.

(2) Show that the error terms arising from zeroes are $O(1)$ on the average;

(3) Justify the central limit calculation for $\sum_p X_p$.

I'll now briefly talk (informally) about each of the three steps in turn.

3.8.2. The explicit formula. To get a formula such as (3.45), the basic strategy is to take a suitable average of the formula (3.43) and the formula (3.44). Traditionally, this is done by contour integration; however, I prefer (perhaps idiosyncratically) to take a more Fourier-analytic perspective, using convolutions rather than contour integrals. (The two approaches are largely equivalent, though.) The basic point is that the imaginary part $\mathrm{Im}(\rho)$ of the zeroes inhabits the same space as the imaginary part $t = \mathrm{Im}(s)$ of the s variable, which in turn is the Fourier analytic dual of the variable that the logarithm $\log p$ of the primes p live in; this can be seen by writing (3.43), (3.44) in a more Fourier-like manner[9] as

$$\sum_\rho \log|1/2 + it - \rho| + \ldots = \mathrm{Re} \sum_p \frac{1}{\sqrt{p}} e^{-it\log p} + \ldots.$$

The uncertainty principle then predicts that localising $\log p$ to the scale $O(\log T^\varepsilon)$ should result in blurring out the zeroes ρ at scale $O(1/\log T^\varepsilon)$, which is where (3.45) is going to come from.

Let's see how this idea works in practice. We consider a convolution of the form

$$\int_{\mathbf{R}} \log|\zeta(s + \frac{iy}{\log T^\varepsilon})|\psi(y)\, dy \tag{3.46}$$

where ψ is some bump function with total mass 1; informally, this is $\log|\zeta(s)|$ averaged out in the vertical direction at scale $O(1/\log T^\varepsilon) = O(1/\log T)$ (we allow implied constants to depend on ε).

We can express (3.46) in two different ways, one using (3.43), and one using (3.44). Let's look at (3.43) first. If one modifies s by $O(1/\log T)$, then the quantity $\log|s-\rho|$ doesn't fluctuate very much,

[9] These sorts of Fourier-analytic connections are often summarised by the slogan "the zeroes of the zeta function are the music of the primes".

unless ρ is within $O(1/\log T)$ of s, in which case it can move by about $O(1 + \log \frac{|s-\rho|}{1/\log T})$. As a consequence, we see that

$$\int_{\mathbf{R}} \log|s + \frac{iy}{\log T^\varepsilon} - \rho|\psi(y)\ dy \approx \log|s-\rho|$$

when $|\rho - s| \gg 1/\log T$, and

$$\int_{\mathbf{R}} \log|s + \frac{iy}{\log T^\varepsilon} - \rho|\psi(y)\ dy = \log|s-\rho| + O(1 + \log \frac{|s-\rho|}{1/\log T}).$$

The quantity $\log|s-1|$ also doesn't move very much by this shift (we are assuming the imaginary part of s to be large). Inserting these facts into (3.43), we thus see that (3.46) is (heuristically) equal to

$$\log|\zeta(s)| + \sum_{\rho = s + O(1/\log T)} O(1 + \log \frac{|s-\rho|}{1/\log T}) + \ldots. \tag{3.47}$$

Now let's compute (3.46) using (3.44) instead. Writing $s = 1/2 + it$, we express (3.46) as

$$\sum_p \mathrm{Re}\, \frac{1}{p^s} \int_{\mathbf{R}} e^{-iy \log p/\log T^\varepsilon} \psi(y)\ dy + \ldots.$$

Introducing the Fourier transform $\hat{\psi}(\xi) := \int_{\mathbf{R}} e^{-iy\xi}\psi(y)\ dy$ of ψ, one can write this as

$$\sum_p \mathrm{Re}\, \frac{1}{p^s} \hat{\psi}(\log p/\log T^\varepsilon) + \ldots.$$

Now we took ψ to be a bump function, so its Fourier transform should also be like a bump function (or perhaps a Schwartz function). As a first approximation, one can thus think of $\hat{\psi}$ as a smoothed truncation to the region $\{\xi : \xi = O(1)\}$, thus the $\hat{\psi}(\log p/\log T^\varepsilon)$ weight is morally restricting p to the region $p \leq T^\varepsilon$. Thus we (morally) can express (3.46) as

$$\sum_{p \leq T^\varepsilon} \mathrm{Re}\, \frac{1}{p^s} + \ldots.$$

Comparing this with the other formula (3.47) we have for (3.46), we obtain (3.45) as required (formally, at least).

3.8.3. Controlling the zeroes. Next, we want to show that the quantity

$$\sum_{\rho=s+O(1/\log T)} 1 + |\log \frac{|s-\rho|}{1/\log T}|$$

is $O(1)$ on the average, when $s = 1/2 + it$ and t is chosen uniformly at random from T to $2T$.

For this, we can use the *first moment method*. For each zero ρ, let I_ρ be the random variable which equals $1 + |\log \frac{|s-\rho|}{1/\log T}|$ when $\rho = s + O(1/\log T)$ and zero otherwise, thus we are trying to control the expectation of $\sum_\rho I_\rho$. The only zeroes which are relevant are those which are of size $O(T)$, and we know that there are $O(T \log T)$ of these (indeed, we have an even more precise formula, as remarked earlier). On the other hand, a randomly chosen s has a probability of $O(1/T \log T)$ of falling within $O(1/\log T)$ of ρ, and so we expect each I_ρ to have an expected value of $O(1/T \log T)$. (The logarithmic factor in the definition of I_ρ turns out not to be an issue, basically because $\log x$ is locally integrable.) By linearity of expectation, we conclude that $\sum_\rho I_\rho$ has expectation $O(T \log T) \times O(1/T \log T) = O(1)$, and the claim follows.

Remark 3.8.2. One can actually do a bit better than this, showing that higher order moments of $\sum_\rho I_\rho$ are also $O(1)$, by using a variant of (3.45) together with the moment bounds in the next section; but we will not need that refinement here.

3.8.4. The central limit theorem. Finally, we have to show that $\sum_{p \leq T^\varepsilon} X_p$ behaves like a normal distribution, as predicted by the central limit theorem heuristic. The key is to show that the X_p behave "as if" they were jointly independent. In particular, as the X_p all have mean zero, one would like to show that products such as

$$X_{p_1} \dots X_{p_k} \tag{3.48}$$

have a negligible expectation as long as at least one of the primes in $p_1, \dots, p_k$ occurs at most once. Once one has this (as well as a similar formula for the case when all primes appear at least twice), one can then do a standard moment computation of the k^{th} moment $(\sum_{p \leq T^\varepsilon} X_p)^k$ and verify that this moment then matches the answer

predicted by the central limit theorem, which by standard arguments (involving the *Weierstrass approximation theorem*) is enough to establish the distributional law. Note that to get close to the normal distribution by a fixed amount of accuracy, it suffices to control a bounded number of moments, which ultimately means that we can treat k as being bounded, $k = O(1)$.

If we expand out the product (3.48), we get

$$\frac{1}{\sqrt{p_1} \cdots \sqrt{p_k}} \cos(t \log p_1) \ldots \cos(t \log p_k).$$

Using the product formula for cosines (or *Euler's formula*), the product of cosines here can be expressed as a linear combination of cosines $\cos(t\xi)$, where the frequency ξ takes the form

$$\xi = \pm \log p_1 \pm \log p_2 \ldots \pm \log p_k.$$

Thus, ξ is the logarithm of a rational number, whose numerator and denominator are the product of some of the $p_1, \ldots, p_k$. Since all the p_j are at most T^ε, we see that the numerator and denominator here are at most $T^{k\varepsilon}$.

Now for the punchline. If there is a prime in $p_1, \ldots, p_k$ that appears only once, then the numerator and denominator cannot fully cancel, by the *fundamental theorem of arithmetic*. Thus ξ cannot be 0. Furthermore, since the denominator is at most $T^{k\varepsilon}$, we see that ξ must stay away from 0 by a distance of about $1/T^{k\varepsilon}$ or more, and so $\cos(t\xi)$ has a wavelength of at most $O(T^{k\varepsilon})$. On the other hand, t ranges between T and $2T$. If k is fixed and ε is small enough (much smaller than $1/k$), we thus see that the average value of $\cos(t\xi)$ between T and $2T$ is close to zero, and so (3.48) does indeed have negligible expectation as claimed. (A similar argument lets one compute the expectation of (3.48) when all primes appear at least twice.)

Remark 3.8.3. A famous theorem of Erdös and Kac[**ErKa1940**] gives a normal distribution for the number of prime factors of a large number n, with mean $\log\log n$ and variance $\log\log n$. One can view Selberg's theorem as a sort of Fourier-analytic variant of the Erdös-Kac theorem.

Remark 3.8.4. The Fourier-like correspondence between zeroes of the zeta function and primes can be used to convert statements about

zeroes, such as the Riemann hypothesis and the GUE hypothesis, into equivalent statements about primes. For instance, the Riemann hypothesis is equivalent to having the square root error term

$$\sum_{x \leq n \leq x+y} \Lambda(n) = y + O_\varepsilon(y^{1/2+\varepsilon})$$

in the prime number theorem holding asymptotically as $x \to \infty$ for all $\varepsilon > 0$ and all intervals $[x, x+y]$ which are *large* in the sense that y is comparable to x. Meanwhile, the pair correlation conjecture (the simplest component of the GUE hypothesis) is equivalent (on RH) to the square root error term holding (with the expected variance) for all $\varepsilon > 0$ and *almost* all intervals $[x, x+y]$ which are *short* in the sense that $y = x^\theta$ for some small (fixed) $\theta > 0$. (This is a rough statement; a more precise formulation can be found in [**GoMo1987**].) It seems to me that reformulation of the full GUE hypothesis in terms of primes should be similar, but would assert that the error term in the prime number theorem (as well as variants of this theorem for almost primes) in short intervals enjoys the expected normal distribution; I don't know of a precise formulation of this assertion, but calculations in this direction lie in [**BoKe1996**].)

Notes. This article first appeared at `terrytao.wordpress.com/2009/07/12`. Thanks to anonymous commenters for corrections.

Emmanuel Kowalski discusses the relationship between Selberg's limit theorem and the Erdös-Kac theorem further at

`http://blogs.ethz.ch/kowalski/2009/02/28/a-beautiful-analogy-2/`

3.9. $P = NP$, relativisation, and multiple choice exams

The most fundamental unsolved problem in complexity theory is undoubtedly the *P=NP problem*, which asks (roughly speaking) whether a problem which can be solved by a *non-deterministic polynomial-time (NP)* algorithm, can also be solved by a *deterministic polynomial-time (P)* algorithm. The general belief is that $P \neq NP$, i.e. there exist problems which can be solved by non-deterministic polynomial-time algorithms but not by deterministic polynomial-time algorithms.

One reason why the $P \neq NP$ question is so difficult to resolve is that a certain generalisation of this question has an affirmative answer in some cases, and a negative answer in other cases. More precisely, if we give all the algorithms access to an *oracle*, then for one choice A of this oracle, all the problems that are solvable by non-deterministic polynomial-time algorithms that calls A (NP^A), can also be solved by a deterministic polynomial-time algorithm algorithm that calls A (P^A), thus $P^A = NP^A$; but for another choice B of this oracle, there exist problems solvable by non-deterministic polynomial-time algorithms that call B, which *cannot* be solved by a deterministic polynomial-time algorithm that calls B, thus $P^B \neq NP^B$. One particular consequence of this result (which is due to Baker, Gill, and Solovay [**BaGiSo1975**]) is that there cannot be any *relativisable* proof of either $P = NP$ or $P \neq NP$, where "relativisable" means that the proof would also work without any changes in the presence of an oracle.

The Baker-Gill-Solovay result was quite surprising, but the idea of the proof turns out to be rather simple. To get an oracle A such that $P^A = NP^A$, one basically sets A to be a powerful simulator that can simulate non-deterministic machines (and, furthermore, can also simulate *itself*); it turns out that any *PSPACE-complete* oracle would suffice for this task. To get an oracle B for which $P^B \neq NP^B$, one has to be a bit sneakier, setting B to be a query device for a sparse set of random (or high-complexity) strings, which are too complex to be guessed at by any deterministic polynomial-time algorithm.

Unfortunately, the simple idea of the proof can be obscured by various technical details (e.g. using *Turing machines* to define P and NP precisely), which require a certain amount of time to properly absorb. To help myself try to understand this result better, I have decided to give a sort of "allegory" of the proof, based around a (rather contrived) story about various students trying to pass a multiple choice test, which avoids all the technical details but still conveys the basic ideas of the argument.

3.9.1. P **and** NP **students.** In this story, two students, named P and NP (and which for sake of grammar, I will arbitrarily assume to be male), are preparing for their final exam in a maths course, which

will consist of a long, tedious sequence of multiple-choice questions, or more precisely true-false questions. The exam has a reasonable but fixed time limit (e.g. three hours), and unlimited scratch paper is available during the exam. Students are allowed to bring one small index card into the exam. Other than scratch paper, an index card, and a pencil, no other materials are allowed. Students cannot leave questions blank; they must answer each question true or false. The professor for this course is dull and predictable; everyone knows in advance the type of questions that will be on the final, the only issue being the precise numerical values that will be used in the actual questions.

For each student response to a question, there are three possible outcomes:

- **Correct answer.** The student answers the question correctly.
- **False negative.** The student answers "false", but the actual answer is "true".
- **False positive.** The student answers "true", but the actual answer is "false".

We will assume a certain asymmetry in the grading: a few points are deducted for false negatives, but a large number of points are deducted for false positives. (There are many real-life situations in which one type of error is considered less desirable than another; for instance, when deciding on guilt in a capital crime, a false positive is generally considered a much worse mistake than a false negative.) So, while students would naturally like to ace the exam by answering all questions correctly, they would tend to err on the side of caution and put down "false" when in doubt.

Student P is hard working and careful, but unimaginative and with a poor memory. His exam strategy is to put all the techniques needed to solve the exam problems on the index card, so that they can be applied by rote during the exam. If the nature of the exam is such that P can be guaranteed to ace it by this method, we say that the exam *is in class* P. For instance, if the exam will consist of verifying various multiplication problems (e.g. "Is $231 * 136 =$

31516?"), then this exam is in class P, since P can put the algorithm for long multiplication, together with a multiplication table, on the index card, and perform these computations during the exam. A more non-trivial example of an exam in class P would be an exam consisting solely of determining whether various large numbers are prime; here P could be guaranteed to ace the test by writing down on his index card the details of the *AKS primality test*.

Student NP is similar to P, but is substantially less scrupulous; he has bribed the proctor of the exam to supply him with a full solution key, containing not only the answers, but also the worked computations that lead to that answer (when the answer is "true"). The reason he has asked (and paid) for the latter is that he does not fully trust the proctor to give reliable answers, and is terrified of the impact to his grades if he makes a false positive. Thus, if the answer key asserts that the answer to a question is "true", he plans to check the computations given to the proctor himself before putting down "true"; if he cannot follow these computations, and cannot work out the problem himself, he will play it safe and put down "false" instead.

We will say that the exam *is in class* NP if

- NP is guaranteed to ace the exam if the information given to him by the proctor is reliable;
- NP is guaranteed not to make a false positive, even if the proctor has given him unreliable information.

For instance, imagine an exam consisting of questions such as "Is Fermat's last theorem provable in ten pages or less?". Such an exam is in the class NP, as the student can bribe the proctor to ask for a ten-page proof of FLT, if such exists, and then would check that proof carefully before putting down "True". This way, the student is guaranteed not to make a false positive (which, in this context, would be a severe embarrassment to any reputable mathematician), and will ace the exam if the proctor actually does happen to have all the relevant proofs available.

It is clear that NP is always going to do at least as well as P, since NP always has the option of ignoring whatever the proctor gives him, and copying P's strategy instead. But how much of an

advantage does NP have over P? In particular, if we give P a little bit more time (and a somewhat larger index card), could every exam that is in class NP, also be in class P? This, roughly speaking, is the $P = NP$ problem. It is believed that $P \neq NP$, thus there are exams which NP will ace (with reliable information) and will at least not make a false positive (even with unreliable information), but for which P is not guaranteed to ace, even with a little extra time and space.

3.9.2. Oracles. Now let's modify the exams a bit by allowing a limited amount of computer equipment in the exam. In addition to the scratch paper, pencil, and index card, every student in the exam is now also given access to a computer A which can perform a carefully limited set of tasks that are intended to assist the student. Examples of tasks permitted by A could include a scientific calculator, a mathematics package such as Matlab or SAGE, or access to Wikipedia or Google. We say that an exam is *in class* P^A if it can be guaranteed to be aced by P if he has access to A, and similarly the exam *is in class* NP^A if it can be guaranteed to be aced by NP if he has access to A and the information obtained from the proctor was reliable, and if he is at least guaranteed not to make a false positive with access to A if the information from the proctor turned out to be unreliable. Again, it is clear that NP will have the advantage over P, in the sense that every exam in class P^A will also be in class NP^A. (In other words, the proof that $P \subset NP$ *relativises*.) But what about the converse - is every exam in class NP^A, also in class P^A (if we give P a little more time and space, and perhaps also a slightly larger and faster version of A)?

We now give an example of a computer A with the property that $P^A = NP^A$, i.e. that every exam in class NP^A, is also in class P^A. Here, A is an extremely fast computer with reasonable amount of memory and a compiler for a general-purpose programming language, but with no additional capabilities. (More precisely, A should be a *PSPACE-complete* language, but let me gloss over the precise definition of this term here.)

Suppose that an exam is in class NP^A, thus NP will ace the exam if he can access A and has reliable information, and will not give any

false positive if he can access A and has unreliable information. We now claim that P can also ace this exam, if given a little bit more time and a slightly larger version of A. The way he does it is to program his version of A to simulate NP's strategy, by looping through all possible values of the solution key that NP might be given, and also simulating NP's copy of A as well. (The latter task is possible as long as P's version of A is slightly larger and faster than NP's version.) There are of course an extremely large number of combinations of solution key to loop over (for instance, consider how many possible proofs of Fermat's last theorem under ten pages there could be), but we assume that the computer is so fast that it can handle all these combinations without difficulty. If at least one of the possible choices for a solution key causes the simulation of NP to answer "true", then P will answer "true" also; if instead none of the solution keys cause NP to answer "true", then P will answer "false" instead. If the exam is in class NP^A, it is then clear that P will ace the exam.

Now we give an example of a computer B with the property that $P^B \neq NP^B$, i.e. there exists an exam which is in class NP^B, but for which P is not guaranteed to ace even with the assistance of B. The only software loaded on B is a web browser, which can fetch any web page desired after typing in the correct URL. However, rather than being connected to the internet, the browser can only access a local file system of pages. Furthermore, there is no directory or search feature in this file system; the only way to find a page is to type in its URL, and if you can't guess the URL correctly, there is no way to access that page. (In particular, there are no links between pages.)

Furthermore, to make matters worse, the URLs are not designed according to any simple scheme, but have in fact been generated randomly, by the following procedure. For each positive integer n, flip a coin. If the coin is heads, then create a URL of n random characters and place a web page at that URL. Otherwise, if the coin is tails, do nothing. Thus, for each n, there will either be one web page with a URL of length n, or there will be no web pages of this length; but in the former case, the web page will have an address consisting of complete gibberish, and there will be no means to obtain this address other than by guessing.

The exam will consist of a long series of questions such as "Is there a web page on B with a URL of 1254 characters in length?".

It is clear that this exam is in class NP^B. Indeed, for NP to ace this exam, he just needs to bribe the proctor for the URLs of all the relevant web pages (if they exist). He can then confirm their existence by typing them into B, and then answer "true" if he finds the page, and "false" otherwise. It is clear that NP will ace the exam if the proctor information is reliable, and will avoid false positives otherwise.

On the other hand, poor P will have no chance to ace this exam if the length of the URLs are long enough, for two reasons. Firstly, the browser B is useless to him: any URL he can guess will have almost no chance of being the correct one, and so the only thing he can generate on the browser is an endless stream of "404 Not Found" messages. (Indeed, these URLs are very likely to have a high *Kolmogorov complexity*, and thus cannot be guessed by P. Admittedly, P does have B available, but one can show by induction on the number of queries that B is useless to P. We also make the idealised assumption that *side-channel attacks* are not available.) As B is useless, the only hope P has is to guess the sequence of coin flips that were used to determine the set of n for which URLs exist of that length. But the random sequence of coin flips is also likely to have high Kolmogorov complexity, and thus cannot be guaranteed to be guessed by P either. Thus $P^B \neq NP^B$.

Remark 3.9.1. Note how the existence of long random strings could be used to make an oracle that separates P from NP. In the absence of oracles, it appears that separation of P from NP is closely connected to the existence of long *pseudorandom* strings - strings of numbers which can be deterministically generated (perhaps from a given seed) in a reasonable amount of time, but are difficult to distinguish from genuinely random strings by any quick tests.

Notes. This article first appeared at `terrytao.wordpress.com/2009/08/01`. Thanks to Tom for corrections.

There was some discussion on the relationship between $P = NP$ and $P = BPP$. Greg Kuperberg gave some further examples of oracles that shed some light on this:

- Consider as an oracle an extremely large book of randomly generated numbers. This oracle could be used to simulate any probabilistic algorithm, so $P = BPP$ relative to this oracle. On the other hand, if one assigns the task to determine whether a given string of numbers exists in some range in the book, this question is in NP but not in P.
- Another example of an oracle would be an extremely large book, in which most of the pages contained the answer to the problem at hand, but for which the n^{th} page was blank for every natural number n that could be quickly created by any short deterministic algorithm. This type of oracle could be used to create a scenario in which $P \neq BPP$ and $P \neq NP$.
- A third example, this time of an *advice function* rather than an oracle, would be if the proctor wrote a long random string on the board before starting the exam (with the length of the string depending on the length of the exam). This can be used to show the inclusion $BPP \subset P/poly$.

By using written oracles instead of computer oracles, it also became more obvious that the oracles were non-interactive (i.e. subsequent responses by the oracle did not depend on earlier queries).

3.10. Moser's entropy compression argument

There are many situations in combinatorics in which one is running some sort of iteration algorithm to continually "improve" some object A; each loop of the algorithm replaces A with some better version A' of itself, until some desired property of A is attained and the algorithm halts. In order for such arguments to yield a useful conclusion, it is often necessary that the algorithm halts in a finite amount of time, or (even better), in a bounded amount of time[10].

[10] In general, one cannot use infinitary iteration tools, such as *transfinite induction* or *Zorn's lemma* (Section 2.4), in combinatorial settings, because the iteration processes used to improve some target object A often degrade some other finitary quantity B in the process, and an infinite iteration would then have the undesirable effect of making B infinite.

A basic strategy to ensure termination of an algorithm is to exploit a *monotonicity property*, or more precisely to show that some key quantity keeps increasing (or keeps decreasing) with each loop of the algorithm, while simultaneously staying bounded. (Or, as the economist Herbert Stein was fond of saying, "If something cannot go on forever, it must stop.")

Here are four common flavours of this monotonicity strategy:

- The *mass increment argument*. This is perhaps the most familiar way to ensure termination: make each improved object A' "heavier" than the previous one A by some non-trivial amount (e.g. by ensuring that the cardinality of A' is strictly greater than that of A, thus $|A'| \geq |A| + 1$). Dually, one can try to force the amount of "mass" remaining "outside" of A in some sense to decrease at every stage of the iteration. If there is a good upper bound on the "mass" of A that stays essentially fixed throughout the iteration process, and a lower bound on the mass increment at each stage, then the argument terminates. Many "greedy algorithm" arguments are of this type. The proof of the Hahn decomposition theorem (Theorem 1.2.2) also falls into this category. The general strategy here is to keep looking for useful pieces of mass outside of A, and add them to A to form A', thus exploiting the additivity properties of mass. Eventually no further usable mass remains to be added (i.e. A is *maximal* in some L^1 sense), and this should force some desirable property on A.
- The *density increment argument*. This is a variant of the mass increment argument, in which one increments the "density" of A rather than the "mass". For instance, A might be contained in some ambient space P, and one seeks to improve A to A' (and P to P') in such a way that the density of the new object in the new ambient space is better than that of the previous object (e.g. $|A'|/|P'| \geq |A|/|P| + c$ for some $c > 0$). On the other hand, the density of A is clearly bounded above by 1. As long as one has a sufficiently good lower bound on the density increment at each stage, one

can conclude an upper bound on the number of iterations in the algorithm. The prototypical example of this is Roth's proof of his theorem[**Ro1953**] that every set of integers of positive upper density contains an arithmetic progression of length three. The general strategy here is to keep looking for useful density fluctuations inside A, and then "zoom in" to a region of increased density by reducing A and P appropriately. Eventually no further usable density fluctuation remains (i.e. A is *uniformly distributed*), and this should force some desirable property on A.

- The *energy increment argument*. This is an "L^2" analogue of the "L^1"-based mass increment argument (or the "L^∞"-based density increment argument), in which one seeks to increments the amount of "energy" that A captures from some reference object X, or (equivalently) to decrement the amount of energy of X which is still "orthogonal" to A. Here A and X are related somehow to a Hilbert space, and the energy involves the norm on that space. A classic example of this type of argument is the existence of orthogonal projections onto closed subspaces of a Hilbert space; this leads among other things to the construction of *conditional expectation* in measure theory, which then underlies a number of arguments in ergodic theory, as discussed for instance in Section 2.8 of *Poincaré's Legacies, Vol. I.* Another basic example is the standard proof of the *Szemerédi regularity lemma* (where the "energy" is often referred to as the "index"). These examples are related; see Section 4.2 for further discussion. The general strategy here is to keep looking for useful pieces of energy orthogonal to A, and add them to A to form A', thus exploiting square-additivity properties of energy, such as Pythagoras' theorem. Eventually, no further usable energy outside of A remains to be added (i.e. A is *maximal* in some L^2 sense), and this should force some desirable property on A.

- The *rank reduction argument*. Here, one seeks to make each new object A' to have a lower "rank", "dimension", or "order" than the previous one. A classic example here is the proof of the linear algebra fact that given any finite set of vectors, there exists a linearly independent subset which spans the same subspace; the proof of the more general *Steinitz exchange lemma* is in the same spirit. The general strategy here is to keep looking for "collisions" or "dependencies" within A, and use them to collapse A to an object A' of lower rank. Eventually, no further usable collisions within A remain, and this should force some desirable property on A.

Much of my own work in additive combinatorics relies heavily on at least one of these types of arguments (and, in some cases, on a nested combination of two or more of them). Many arguments in nonlinear partial differential equations also have a similar flavour, relying on various *monotonicity formulae* for solutions to such equations, though the objective in PDE is usually slightly different, in that one wants to keep control of a solution as one approaches a singularity (or as some time or space coordinate goes off to infinity), rather than to ensure termination of an algorithm. (On the other hand, many arguments in the theory of *concentration compactness*, which is used heavily in PDE, does have the same algorithm-terminating flavour as the combinatorial arguments; see Section 2.1 of *Structure and Randomness* for more discussion.)

Recently, a new species of monotonicity argument was introduced by Moser[**Mo2009**], as the primary tool in his elegant new proof of the *Lovász local lemma*. This argument could be dubbed an *entropy compression argument*, and only applies to *probabilistic algorithms* which require a certain collection R of random "bits" or other random choices as part of the input, thus each loop of the algorithm takes an object A (which may also have been generated randomly) and some portion of the random string R to (deterministically) create a better object A' (and a shorter random string R', formed by throwing away those bits of R that were used in the loop). The key point is to design the algorithm to be partially *reversible*, in the sense that given A' and

R' and some additional data H' that logs the cumulative *history* of the algorithm up to this point, one can reconstruct A together with the remaining portion R not already contained in R'. Thus, each stage of the argument *compresses* the information-theoretic content of the string $A + R$ into the string $A' + R' + H'$ in a lossless fashion. However, a random variable such as $A + R$ cannot be compressed losslessly into a string of expected size smaller than the *Shannon entropy* of that variable. Thus, if one has a good lower bound on the entropy of $A + R$, and if the length of $A' + R' + H'$ is significantly less than that of $A + R$ (i.e. we need the marginal growth in the length of the history file H' per iteration to be less than the marginal amount of randomness used per iteration), then there is a limit as to how many times the algorithm can be run, much as there is a limit as to how many times a random data file can be compressed before no further length reduction occurs.

It is interesting to compare this method with the ones discussed earlier. In the previous methods, the failure of the algorithm to halt led to a new iteration of the object A which was "heavier", "denser", captured more "energy", or "lower rank" than the previous instance of A. Here, the failure of the algorithm to halt leads to new information that can be used to "compress" A (or more precisely, the full state $A + R$) into a smaller amount of space. I don't know yet of any application of this new type of termination strategy to the fields I work in, but one could imagine that it could eventually be of use (perhaps to show that solutions to PDE with sufficiently "random" initial data can avoid singularity formation?), so I thought I would discuss (a special case of) it here.

Rather than deal with the Lovász local lemma in full generality, I will work with a special case of this lemma involving the *k-satisfiability problem* (in *conjunctive normal form*). Here, one is given a set of *boolean variables* $x_1, \ldots, x_n$ together with their negations $\neg x_1, \ldots, \neg x_n$; we refer to the $2n$ variables and their negations collectively as *literals*. We fix an integer $k \geq 2$, and define a (length k) *clause* to be a *disjunction* of k literals, for instance

$$x_3 \vee \neg x_5 \vee x_9$$

is a clause of length three, which is true unless x_3 is false, x_5 is true, and x_9 is false. We define the *support* of a clause to be the set of variables that are involved in the clause, thus for instance $x_3 \vee \neg x_5 \vee x_9$ has support $\{x_3, x_5, x_9\}$. To avoid degeneracy we assume that no clause uses a variable more than once (or equivalently, all supports have cardinality exactly k), thus for instance we do not consider $x_3 \vee x_3 \vee x_9$ or $x_3 \vee \neg x_3 \vee x_9$ to be clauses.

Note that the failure of a clause reveals complete information about all k of the boolean variables in the support; this will be an important fact later on.

The *k-satisfiability problem* is the following: given a set S of clauses of length k involving n boolean variables $x_1, \ldots, x_n$, is there a way to assign truth values to each of the $x_1, \ldots, x_n$, so that all of the clauses are simultaneously satisfied?

For general S, this problem is easy for $k = 2$ (essentially equivalent to the problem of 2-colouring a graph), but NP-complete for $k \geq 3$ (this is the famous *Cook-Levin theorem*). But the problem becomes simpler if one makes some more assumptions on the set S of clauses. For instance, if the clauses in S have disjoint supports, then they can be satisfied independently of each other, and so one easily has a positive answer to the satisfiability problem in this case. (Indeed, one only needs each clause in S to have *one* variable in its support that is disjoint from all the other supports in order to make this argument work.)

Now suppose that the clauses S are not completely disjoint, but have a limited amount of overlap; thus *most* clauses in S have disjoint supports, but not all. With too much overlap, of course, one expects satisfability to fail (e.g. if S is the set of *all* length k clauses). But with a sufficiently small amount of overlap, one still has satisfiability:

Theorem 3.10.1 (Lovász local lemma, special case)**.** *Suppose that S is a set of length k clauses, such that the support of each clause s in S intersects at most 2^{k-C} supports of clauses in S (including s itself), where C is a sufficiently large absolute constant. Then the clauses in S are simultaneously satisfiable.*

One of the reasons that this result is powerful is that the bounds here are uniform in the number n of variables. Apart from the loss of C, this result is sharp; consider for instance the set S of all 2^k clauses with support $\{x_1, \ldots, x_k\}$, which is clearly unsatisfiable.

The standard proof of this theorem proceeds by assigning each of the n boolean variables $x_1, \ldots, x_n$ a truth value $a_1, \ldots, a_n \in \{\text{true}, \text{false}\}$ independently at random (with each truth value occurring with an equal probability of $1/2$); then each of the clauses in S has a positive zero probability of holding (in fact, the probability is $1 - 2^{-k}$). Furthermore, if E_s denotes the event that a clause $s \in S$ is satisfied, then the E_s are mostly independent of each other; indeed, each event E_s is independent of all but most 2^{k-C} other events $E_{s'}$. Applying the *Lovász local lemma*, one concludes that the E_s simultaneously hold with positive probability (if C is a little bit larger than $\log_2 e$), and the claim follows.

The textbook proof of the Lovász local lemma is short but non-constructive; in particular, it does not easily offer any quick way to compute an actual satisfying assignment for $x_1, \ldots, x_n$, only saying that such an assignment exists. Moser's argument, by contrast, gives a simple and natural algorithm to locate such an assignment (and thus prove Theorem 3.10.1). (The constant C becomes 3 rather than $\log_2 e$, although the $\log_2 e$ bound has since been recovered in a paper of Moser and Tardos.)

As with the usual proof, one begins by randomly assigning truth values $a_1, \ldots, a_n \in \{\text{true}, \text{false}\}$ to $x_1, \ldots, x_n$; call this random assignment $A = (a_1, \ldots, a_n)$. If A satisfied all the clauses in S, we would be done. However, it is likely that there will be some non-empty subset T of clauses in S which are not satisfied by A.

We would now like to modify A in such a manner to reduce the number $|T|$ of violated clauses. If, for instance, we could always find a modification A' of A whose set T' of violated clauses was strictly smaller than T (assuming of course that T is non-empty), then we could iterate and be done (this is basically a mass decrement argument). One obvious way to try to achieve this is to pick a clause s in T that is violated by A, and modify the values of A on the support

of s to create a modified set A' that satisfies s, which is easily accomplished; in fact, any non-trivial modification of A on the support will work here. In order to maximize the amount of entropy in the system (which is what one wants to do for an entropy compression argument), we will choose this modification of A' *randomly*; in particular, we will use k fresh random bits to replace the k bits of A in the support of s. (By doing so, there is a small probability (2^{-k}) that we in fact do not change A at all, but the argument is (very) slightly simpler if we do not bother to try to eliminate this case.)

If all the clauses had disjoint supports, then this strategy would work without difficulty. But when the supports are not disjoint, one has a problem: every time one modifies A to "fix" a clause s by modifying the variables on the support of s, one may cause other clauses s' whose supports overlap those of s to fail, thus potentially increasing the size of T by as much as $2^{k-C} - 1$. One could then try fixing all the clauses which were broken by the first fix, but it appears that the number of clauses needed to repair could grow indefinitely with this procedure, and one might never terminate in a state in which all clauses are simultaneously satisfied.

The key observation of Moser, as alluded earlier, is that each failure of a clause s for an assignment A reveals k bits of information about A, namely that the exact values that A assigns to the support of s. The plan is then to use each failure of a clause as a part of a *compression protocol* that compresses A (plus some other data) losslessly into a smaller amount of space. A crucial point is that at each stage of the process, the clause one is trying to fix is almost always going to be one that overlapped the clause that one had just previously fixed. Thus the total number of possibilities for each clause, given the previous clauses, is basically 2^{k-C}, which requires only $k - C$ bits of storage, compared with the k bits of entropy that have been eliminated. This is what is going to force the algorithm to terminate in finite time (with positive probability).

Let's make the details more precise. We will need the following objects:

- A truth assignment A of n truth values $a_1, \ldots, a_n$, which is initially assigned randomly, but which will be modified as the algorithm progresses;
- A long random string R of bits, from which we will make future random choices, with each random bit being removed from R as it is read.

We also need a recursive algorithm Fix(s), which modifies the string A to satisfy a clause s in S (and, as a bonus, may also make A obey some other clauses in S that it did not previously satisfy). It is defined recursively:

- Step 1. If A already satisfies s, do nothing (i.e. leave A unchanged).
- Step 2. Otherwise, read off k random bits from R (thus shortening R by k bits), and use these to replace the k bits of A on the support of s in the obvious manner (ordering the support of s by some fixed ordering, and assigning the j^{th} bit from R to the j^{th} variable in the support for $1 \leq j \leq k$).
- Step 3. Next, find all the clauses s' in S whose supports intersect s, and which A now violates; this is a collection of at most 2^{k-C} clauses, possibly including s itself. Order these clauses s' in some arbitrary fashion, and then apply Fix(s') to each such clause in turn. (Thus the original algorithm Fix(s) is put "on hold" on some CPU stack while all the child processes Fix(s') are executed; once all of the child processes are complete, Fix(s) then terminates also.)

An easy induction shows that if Fix(s) terminates, then the resulting modification of A will satisfy s; and furthermore, any other clause s' in S which was already satisfied by A before Fix(s) was called, will continue to be satisfied by A after Fix(s) is called. Thus, Fix(s) can only serve to decrease the number of unsatisfied clauses T in S, and so one can fix all the clauses by calling Fix(s) once for each clause in T - provided that these algorithms all terminate.

Each time Step 2 of the Fix algorithm is called, the assignment A changes to a new assignment A', and the random string R changes to a shorter string R'. Is this process reversible? Yes - provided that one

knows what clause s was being fixed by this instance of the algorithm. Indeed, if s, A', R' are known, then A can be recovered by changing the assignment of A' on the support of s to the only set of choices that violates s, while R can be recovered from R' by appending to R' the bits of A on the support of s.

This type of reversibility does not seem very useful for an entropy compression argument, because while R' is shorter than R by k bits, it requires about $\log |S|$ bits to store the clause s. So the map $A+R \mapsto A' + R' + s$ is only a compression if $\log |S| < k$, which is not what is being assumed here (and in any case the satisfiability of S in the case $\log |S| < k$ is trivial from the union bound).

The key trick is that while it does indeed take $\log |S|$ bits to store any given clause s, there is an economy of scale: after many recursive applications of the fix algorithm, the *marginal* amount of bits needed to store s drops to merely $k - C + O(1)$, which is less than k if C is large enough, and which will therefore make the entropy compression argument work.

Let's see why this is the case. Observe that the clauses s for which the above algorithm $\mathrm{Fix}(s)$ is called come in two categories. Firstly, there are those s which came from the original list T of failed clauses. Each of these will require $O(\log |S|)$ bits to store - but there are only $|T|$ of them. Since $|T| \leq |S|$, the net amount of storage space required for these clauses is $O(|S| \log |S|)$ at most. Actually, one can just store the subset T of S using $|S|$ bits (one for each element of S, to record whether it lies in T or not).

Of more interest is the other category of clauses s, in which $\mathrm{Fix}(s)$ is called recursively from some previously invoked call $\mathrm{Fix}(s')$ to the fix algorithm. But then s is one of the at most 2^{k-C} clauses in S whose support intersects that of s'. Thus one can encode s using s' and a number between 1 and 2^{k-C}, representing the position of s (with respect to some arbitrarily chosen fixed ordering of S) in the list of all clauses in S whose supports intersect that of s'. Let us call this number the *index* of the call $\mathrm{Fix}(s)$.

Now imagine that while the Fix routine is called, a running *log file* (or history) H of the routine is kept, which records s each time one of the original $|T|$ calls $\mathrm{Fix}(s)$ with $s \in T$ is invoked, and also

records the index of any other call Fix(s) made during the recursive procedure. Finally, we assume that this log file records a termination symbol whenever a Fix routine terminates. By performing a *stack trace*, one sees that whenever a Fix routine is called, the clause s that is being repaired by that routine can be deduced from an inspection of the log file H up to that point.

As a consequence, at any intermediate stage in the process of all these fix calls, the original state $A + R$ of the assignment and the random string of bits can be deduced from the current state $A' + R'$ of these objects, plus the history H' up to that point.

Now suppose for contradiction that S is not satisfiable; thus the stack of fix calls can never completely terminate. We trace through this stack for M steps, where M is some large number to be chosen later. After these steps, the random string R has shortened by an amount of Mk; if we set R to initially have length Mk, then the string is now completely empty, $R' = \emptyset$. On the other hand, the history H' has size at most $O(|S|) + M(k - C + O(1))$, since it takes $|S|$ bits to store the initial clauses in T, $O(|S|) + O(M)$ bits to record all the instances when Step 1 occurs, and every subsequent call to Fix generates a $k - C$-bit number, plus possibly a termination symbol of size $O(1)$. Thus we have a lossless compression algorithm $A + R \mapsto A' + H'$ from $n + Mk$ completely random bits to $n + O(|S|) + M(k-C+O(1))$ bits (recall that A and R were chosen randomly, and independently of each other). But since $n + Mk$ random bits cannot be compressed losslessly into any smaller space, we have the entropy bound

$$n + O(|S|) + M(k - C + O(1)) \geq n + Mk \tag{3.49}$$

which leads to a contradiction if M is large enough (and if C is larger than an absolute constant). This proves Theorem 3.10.1.

Remark 3.10.2. Observe that the above argument in fact gives an explicit bound on M, and with a small bit of additional effort, it can be converted into a probabilistic algorithm that (with high probability) computes a satisfying assignment for S in time polynomial in $|S|$ and n.

Remark 3.10.3. One can replace the usage of randomness and Shannon entropy in the above argument with *Kolmogorov complexity* instead; thus, one sets $A+R$ to be a string of $n+Mk$ bits which cannot be computed by any algorithm of length $n + O(|S| \log |S|) + M(k - C + O(1))$, the existence of which is guaranteed as soon as (3.49) is violated; the proof now becomes deterministic, except of course for the problem of building the high-complexity string, which by their definition can only be constructed quickly by probabilistic methods.

Notes. This article first appeared at `terrytao.wordpress.com/2009/08/05`, but is based on an earlier blog post by Lance Fortnow at

`blog.computationalcomplexity.org/2009/06`.

Thanks to harrison, Heinrich, nh, and anonymous commenters for corrections.

There was some discussion online about the tightness of bounds in the argument.

3.11. The AKS primality test

The *Agrawal-Kayal-Saxena (AKS) primality test*, discovered in 2002, is the first provably deterministic algorithm to determine the primality of a given number with a run time which is guaranteed to be polynomial in the number of digits, thus, given a large number n, the algorithm will correctly determine whether that number is prime or not in time $O(\log^{O(1)} n)$. (Many previous primality testing algorithms existed, but they were either probabilistic in nature, had a running time slower than polynomial, or the correctness could not be guaranteed without additional hypotheses such as GRH.)

In this article I sketch the details of the test (and the proof that it works) here. (Of course, full details can be found in the original paper[**AgKaSa2004**], which is nine pages in length and almost entirely elementary in nature.) It relies on polynomial identities that are true modulo n when n is prime, but cannot hold for n non-prime as they would generate a large number of additional polynomial identities, eventually violating the *factor theorem* (which asserts that a polynomial identity of degree at most d can be obeyed by at most d

values of the unknown). To remove some clutter in the notation, I have relied (somewhat loosely) on *asymptotic notation* in this article.

Our starting point is *Fermat's little theorem*, which asserts that

$$a^p = a \bmod p \tag{3.50}$$

for every prime p and every a. This theorem suggests an obvious primality test: to test whether a number n is prime, pick a few values of a and see whether $a^n = a \bmod n$. (Note that a^n can be computed in time $O(\log^{O(1)} n)$ for any fixed a by expressing n in binary, and repeatedly squaring a.) If the statement $a^n = a \bmod n$ fails for some a, then n would be composite. Unfortunately, the converse is not true: there exist non-prime numbers n, known as *Carmichael numbers*, for which $a^n = a \bmod n$ for all a coprime to n (561 is the first example). So Fermat's little theorem cannot be used, by itself, to establish primality for general n, because it is too weak to eliminate all non-prime numbers. (The situation improves though for more special types of n, such as Mersenne numbers; see Section 1.7 of *Poincaré's Legacies, Vol. I* for more discussion.)

However, there is a stronger version of Fermat's little theorem which does eliminate all non-prime numbers. Specifically, if p is prime and a is arbitrary, then one has the polynomial identity

$$(X + a)^p = X^p + a \bmod p \tag{3.51}$$

where X is an indeterminate variable. (More formally, we have the identity $(X + a)^p = X^p + a$ in the ring $F_p[X]$ of polynomials of one variable X over the finite field F_p of p elements.) This identity (a manifestation of the *Frobenius endomorphism*) clearly implies (3.50) by setting $X = 0$; conversely, one can easily deduce (3.51) from (3.50) by expanding out $(X + a)^p$ using the *binomial theorem* and the observation that the binomial coefficients $\binom{p}{i} = \frac{p \cdot \ldots \cdot (p-i+1)}{i!}$ are divisible by p for all $1 \leq i < p$. Conversely, if

$$(X + a)^n = X^n + a \bmod n \tag{3.52}$$

(i.e. $(X+a)^n = X^n + a$ in $(\mathbf{Z}/n\mathbf{Z})[X]$) for some a coprime to n, then by comparing coefficients using the binomial theorem we see that $\binom{n}{i}$ is divisible by n for all $1 \leq i < n$. But if n is divisible by some smaller prime p, then by setting i equal to the largest power of p that

divides n, one sees that $\binom{n}{i}$ is not divisible by enough powers of p to be divisible by n, a contradiction. Thus one can use (3.52) (for a single value of a coprime to n) to decide whether n is prime or not.

Unfortunately, this algorithm, while deterministic, is not polynomial-time, because the polynomial $(X+a)^n$ has $n+1$ coefficients and will therefore take at least $O(n)$ time to compute. However, one can speed up the process by descending to a quotient ring of $(\mathbf{Z}/n\mathbf{Z})[X]$, such as $F_p[X]/(X^r-1)$ for some r. Clearly, if the identity $(X+a)^n = X^n+a$ holds in $(\mathbf{Z}/n\mathbf{Z})[X]$, then it will also hold in $(\mathbf{Z}/n\mathbf{Z})[X]/(X^r-1)$, thus

$$(X+a)^n = X^n + a \bmod n, X^r - 1. \tag{3.53}$$

The point of doing this is that (if r is not too large) the left-hand side of (3.53) can now be computed quickly (again by expanding n in binary and performing repeated squaring), because all polynomials can be reduced to be of degree less than r, rather than being as large as n. Indeed, if $r = O(\log^{O(1)} n)$, then one can test (3.53) in time $O(\log^{O(1)} n)$.

We are not done yet, because it could happen that (3.53) holds but (3.52) fails. But we have the following key theorem:

Theorem 3.11.1 (AKS theorem)**.** *Suppose that for all* $1 \leq a, r \leq O(\log^{O(1)} n)$, (3.53) *holds, and* a *is coprime to* n*. Then* n *is either a prime, or a power of a prime.*

Of course, coprimality of a and n can be quickly tested using the *Euclidean algorithm*, and if coprimality fails then n is of course composite. Also, it is easy to quickly test for the property that n is a power of an integer (just compute the roots $n^{1/k}$ for $1 \leq k \leq \log_2 n$), and such powers are clearly composite. From all this (and (3.51), one soon sees that theorem gives rise to a deterministic polynomial-time test for primality. One can optimise the powers of $\log n$ in the bounds for a, r (as is done in [**AgKaSa2004**]), but we will not do so here to keep the exposition uncluttered.

Actually, we don't need (3.53) satisfied for all that many exponents r to make the theorem work; just one well-chosen r will do. More precisely, we have

Theorem 3.11.2 (AKS theorem, key step)**.** *Let r be coprime to n, and such that n has order greater than $\log_2^2 n$ in the multiplicative group $(\mathbf{Z}/r\mathbf{Z})^\times$ (i.e. the residues n^i mod r for $1 \leq i \leq \log^2 n$ are distinct). Suppose that for all $1 \leq a \leq O(r \log^{O(1)} n)$, (3.53) holds, and a is coprime to n. Then n is either a prime, or a power of a prime.*

To find an r with the above properties we have

Lemma 3.11.3 (Existence of good r)**.** *There exists $r = O(\log^{O(1)} n)$ coprime to n, such that n has order greater than $\log_2^2 n$ in $(\mathbf{Z}/r\mathbf{Z})^\times$.*

Proof. For each $1 \leq i \leq \log_2^2 n$, the number $n^i - 1$ has at most $O(\log^{O(1)} n)$ prime divisors (by the fundamental theorem of arithmetic). If one picks r to be the first prime not equal to any of these prime divisors, one obtains the claim. (One can use a crude version of the prime number theorem to get the upper bound on r.) □

It is clear that Theorem 3.11.1 follows from Theorem 3.11.2 and Lemma 3.11.3, so it suffices now to prove Theorem 3.11.2.

Suppose for contradiction that Theorem 3.11.2 fails. Then n is divisible by some smaller prime p, but is not a power of p. Since n is coprime to all numbers of size $O(\log^{O(1)} n)$ we know that p is not of polylogarithmic size, thus we may assume $p \geq \log^C n$ for any fixed C. As r is coprime to n, we see that r is not a multiple of p (indeed, one should view p as being much larger than r).

Let F be a field extension of F_p by a primitive r^{th} root of unity X, thus $F = F_p[X]/h(X)$ for some factor $h(X)$ (in $F_p[X]$) of the r^{th} *cyclotomic polynomial* $\Phi_r(X)$. From the hypothesis (3.53), we see that

$$(X + a)^n = X^n + a$$

in F for all $1 \leq a \leq A$, where $A = O(r \log^{O(1)} n)$. Note that n is coprime to every integer less than A, and thus $A < p$.

Meanwhile, from (3.51) one has

$$(X + a)^p = X^p + a$$

in F for all such a. The two equations give

$$(X^p + a)^{n/p} = (X^p)^{n/p} + a.$$

Note that the p^{th} power X^p of a primitive r^{th} root of unity X is again a primitive r^{th} root of unity (and conversely, every primitive r^{th} root arises in this fashion) and hence we also have

$$(X+a)^{n/p} = X^{n/p} + a$$

in F for all $1 \leq a \leq A$.

Inspired by this, we define a key concept: a positive integer m is said to be *introspective* if one has

$$(X+a)^m = X^m + a$$

in F for all $1 \leq a \leq A$, or equivalently if $(X+a)^m = \phi_m(X+a)$, where $\phi_m : F \to F$ is the ring homomorphism that sends X to X^m.

We have just shown that $p, n, n/p$ are all introspective; 1 is also trivially introspective. Furthermore, if m and m' are introspective, it is not hard to see that mm' is also introspective. Thus we in fact have a lot of introspective integers: any number of the form $p^i(n/p)^j$ for $i, j \geq 0$ is introspective.

It turns out in fact that it is not possible to create so many different introspective numbers, basically the presence of so many polynomial identities in the field would eventually violate the *factor theorem*. To see this, let $\mathcal{G} \subset F^\times$ be the multiplicative group generated by the quantities $X+a$ for $1 \leq a \leq A$. Observe that $z^m = \phi_m(z)$ for all $z \in \mathcal{G}$. We now show that this places incompatible lower and upper bounds on $\mathcal{G}$. We begin with the lower bound:

Proposition 3.11.4 (Lower bound on $\mathcal{G}$)**.** $|\mathcal{G}| \geq 2^t$.

Proof. Let $P(X)$ be a product of less than t of the quantities $X+1, \ldots, X+A$ (allowing repetitions), then $P(X)$ lies in $\mathcal{G}$. Since $A \geq 2r \geq 2t$, there are certainly at least 2^t ways to pick such a product. So to establish the proposition it suffices to show that all these products are distinct.

Suppose for contradiction that $P(X) = Q(X)$, where P, Q are different products of less than t of the $X+1, \ldots, X+A$. Then, for every introspective m, $P(X^m) = Q(X^m)$ as well (note that $P(X^m) = \phi_m(P(X))$). In particular, this shows that $X^{m_1}, \ldots, X^{m_t}$ are all roots of the polynomial $P - Q$. But this polynomial has degree less than t,

and the $X^{m_1}, \ldots, X^{m_t}$ are distinct by hypothesis, and we obtain the desired contradiction by the factor theorem. □

Proposition 3.11.5 (Upper bound on $\mathcal{G}$)**.** *Suppose that there are exactly t residue classes modulo r of the form $p^i(n/p)^j \bmod r$ for $i, j \geq 0$. Then $|\mathcal{G}| \leq n^{\sqrt{t}}$.*

Proof. By the pigeonhole principle, we must have a collision

$$p^i(n/p)^j = p^{i'}(n/p)^{j'} \bmod r$$

for some $0 \leq i, j, i', j' \leq \sqrt{t}$ with $(i, j) \neq (i', j')$. Setting $m := p^i(n/p)^j$ and $m' := p^{i'}(n/p)^{j'}$, we thus see that there are two distinct introspective numbers m, m' of size most $n^{\sqrt{t}}$ which are equal modulo r. (To ensure that m, m' are distinct, we use the hypothesis that n is not a power of p.) This implies that $\phi_m = \phi_{m'}$, and thus $z^m = z^{m'}$ for all $z \in \mathcal{G}$. But the polynomial $z^m - z^{m'}$ has degree at most $n^{\sqrt{t}}$, and the claim now follows from the factor theorem. □

Since n has order greater than $\log^2 n$ in $(\mathbf{Z}/r\mathbf{Z})^\times$, we see that the number t of residue classes r of the form $p^i(n/p)^j$ is at least $\log^2 n$. But then $2^t > n^{\sqrt{t}}$, and so Propositions 3.11.4, 3.11.5 are incompatible.

Notes. This article first appeared at `terrytao.wordpress.com/2009/08/11`. Thanks to Leandro, theoreticalminimum and windfarmmusic for corrections.

A thorough discussion of the AKS algorithm can be found at [**Gr2005**].

3.12. The prime number theorem in arithmetic progressions, and dueling conspiracies

A fundamental problem in analytic number theory is to understand the distribution of the prime numbers $\{2, 3, 5, \ldots\}$. For technical reasons, it is convenient not to study the primes directly, but a proxy for the primes known as the *von Mangoldt function* $\Lambda : \mathbf{N} \to \mathbf{R}$, defined by setting $\Lambda(n)$ to equal $\log p$ when n is a prime p (or a power of that prime) and zero otherwise. The basic reason why the von Mangoldt

function is useful is that it encodes the *fundamental theorem of arithmetic* (which in turn can be viewed as the defining property of the primes) very neatly via the identity

$$\log n = \sum_{d|n} \Lambda(d) \tag{3.54}$$

for every natural number n.

The most important result in this subject is the *prime number theorem*, which asserts that the number of prime numbers less than a large number x is equal to $(1+o(1))\frac{x}{\log x}$:

$$\sum_{p \leq x} 1 = (1+o(1))\frac{x}{\log x}.$$

Here, of course, $o(1)$ denotes a quantity that goes to zero as $x \to \infty$.

It is not hard to see (e.g. by *summation by parts*) that this is equivalent to the asymptotic

$$\sum_{n \leq x} \Lambda(n) = (1+o(1))x \tag{3.55}$$

for the von Mangoldt function (the key point being that the squares, cubes, etc. of primes give a negligible contribution, so $\sum_{n \leq x} \Lambda(n)$ is essentially the same quantity as $\sum_{p \leq x} \log p$). Understanding the nature of the $o(1)$ term is a very important problem, with the conjectured optimal decay rate of $O(\sqrt{x}\log x)$ being equivalent to the *Riemann hypothesis*, but this will not be our concern here.

The prime number theorem has several important generalisations (for instance, there are analogues for other number fields such as the *Chebotarev density theorem*). One of the more elementary such generalisations is the *prime number theorem in arithmetic progressions*, which asserts that for fixed a and q with a coprime to q (thus $(a,q)=1$), the number of primes less than x equal to a mod q is equal to $(1+o_q(1))\frac{1}{\phi(q)}\frac{x}{\log x}$, where $\phi(q) := \#\{1 \leq a \leq q : (a,q)=1\}$ is the *Euler totient function*:

$$\sum_{p \leq x : p = a \bmod q} 1 = (1+o_q(1))\frac{1}{\phi(q)}\frac{x}{\log x}.$$

(Of course, if a is not coprime to q, the number of primes less than x equal to a mod q is $O(1)$. The subscript q in the $o()$ and $O()$ notation denotes that the implied constants in that notation is allowed to depend on q.) This is a more quantitative version of *Dirichlet's theorem*, which asserts the weaker statement that the number of primes equal to a mod q is infinite. This theorem is important in many applications in analytic number theory, for instance in *Vinogradov's theorem* that every sufficiently large odd number is the sum of three odd primes. (Imagine for instance if almost all of the primes were clustered in the residue class 2 mod 3, rather than 1 mod 3. Then almost all sums of three odd primes would be divisible by 3, leaving dangerously few sums left to cover the remaining two residue classes. Similarly for other moduli than 3. This does not fully rule out the possibility that Vinogradov's theorem could still be true, but it does indicate why the prime number theorem in arithmetic progressions is a relevant tool in the proof of that theorem.)

As before, one can rewrite the prime number theorem in arithmetic progressions in terms of the von Mangoldt function as the equivalent form

$$\sum_{n \leq x: n = a \bmod q} \Lambda(n) = (1 + o_q(1)) \frac{1}{\phi(q)} x.$$

Philosophically, one of the main reasons why it is so hard to control the distribution of the primes is that we do not currently have too many tools with which one can rule out "conspiracies" between the primes, in which the primes (or the von Mangoldt function) decide to correlate with some structured object (and in particular, with a totally multiplicative function) which then visibly distorts the distribution of the primes. For instance, one could imagine a scenario in which the probability that a randomly chosen large integer n is prime is not asymptotic to $\frac{1}{\log n}$ (as is given by the prime number theorem), but instead to fluctuate depending on the phase of the complex number n^{it} for some fixed real number t, thus for instance the probability might be significantly less than $1/\log n$ when $t \log n$ is close to an integer, and significantly more than $1/\log n$ when $t \log n$ is close to a half-integer. This would contradict the prime number theorem, and so this scenario would have to be somehow eradicated in the course

of proving that theorem. In the language of *Dirichlet series*, this conspiracy is more commonly known as a zero of the Riemann zeta function at $1 + it$.

In the above scenario, the primality of a large integer n was somehow sensitive to asymptotic or "Archimedean" information about n, namely the approximate value of its logarithm. In modern terminology, this information reflects the local behaviour of n at the infinite *place* ∞. There are also potential consipracies in which the primality of n is sensitive to the local behaviour of n at finite places, and in particular to the residue class of n mod q for some fixed modulus q. For instance, given a *Dirichlet character* $\chi : \mathbf{Z} \to \mathbf{C}$ of modulus q, i.e. a *completely multiplicative* function on the integers which is periodic of period q (and vanishes on those integers not coprime to q), one could imagine a scenario in which the probability that a randomly chosen large integer n is prime is large when $\chi(n)$ is close to $+1$, and small when $\chi(n)$ is close to -1, which would contradict the prime number theorem in arithmetic progressions. (Note the similarity between this scenario at q and the previous scenario at ∞; in particular, observe that the functions $n \to \chi(n)$ and $n \to n^{it}$ are both totally multiplicative.) In the language of Dirichlet series, this conspiracy is more commonly known as a zero of the *L-function* of χ at 1.

An especially difficult scenario to eliminate is that of *real characters*, such as the *Kronecker symbol* $\chi(n) = \left(\frac{n}{q}\right)$, in which numbers n which are quadratic nonresidues mod q are very likely to be prime, and quadratic residues mod q are unlikely to be prime. Indeed, there is a scenario of this form - the *Siegel zero* scenario - which we are still not able to eradicate (without assuming powerful conjectures such as the *Generalised Riemann Hypothesis (GRH)*), though fortunately Siegel zeroes are not quite strong enough to destroy the prime number theorem in arithmetic progressions.

It is difficult to prove that no conspiracy between the primes exist. However, it is not entirely impossible, because we have been able to exploit two important phenomena. The first is that there is often a "all or nothing dichotomy" (somewhat resembling the *zero-one laws* in probability) regarding conspiracies: in the asymptotic limit, the primes can either conspire totally (or more precisely, anti-conspire

totally) with a multiplicative function, or fail to conspire at all, but there is no middle ground. (In the language of Dirichlet series, this is reflected in the fact that zeroes of a meromorphic function can have order 1, or order 0 (i.e. are not zeroes after all), but cannot have an intermediate order between 0 and 1.) As a corollary of this fact, the prime numbers cannot conspire with two distinct multiplicative functions at once (by having a partial correlation with one and another partial correlation with another); thus one can use the existence of one conspiracy to exclude all the others. In other words, there is at most one conspiracy that can significantly distort the distribution of the primes. Unfortunately, this argument is *ineffective*, because it doesn't give any control at all on what that conspiracy is, or even if it exists in the first place!

But now one can use the second important phenomenon, which is that because of symmetries, one type of conspiracy can lead to another. For instance, because the von Mangoldt function is real-valued rather than complex-valued, we have conjugation symmetry; if the primes correlate with, say, n^{it}, then they must also correlate with n^{-it}. (In the language of Dirichlet series, this reflects the fact that the zeta function and L-functions enjoy symmetries with respect to reflection across the real axis (i.e. complex conjugation).) Combining this observation with the all-or-nothing dichotomy, we conclude that the primes cannot correlate with n^{it} for any non-zero t, which in fact leads directly to the prime number theorem (3.55), as we shall discuss below. Similarly, if the primes correlated with a Dirichlet character $\chi(n)$, then they would also correlate with the conjugate $\overline{\chi}(n)$, which also is inconsistent with the all-or-nothing dichotomy, except in the exceptional case when χ is real - which essentially means that χ is a quadratic character. In this one case (which is the only scenario which comes close to threatening the truth of the prime number theorem in arithmetic progressions), the above tricks fail and one has to instead exploit the algebraic number theory properties of these characters instead, which has so far led to weaker results than in the non-real case.

As mentioned previously in passing, these phenomena are usually presented using the language of Dirichlet series and complex analysis.

This is a very slick and powerful way to do things, but I would like here to present the elementary approach to the same topics, which is slightly weaker but which I find to also be very instructive. (However, I will not be *too* dogmatic about keeping things elementary, if this comes at the expense of obscuring the key ideas; in particular, I will rely on multiplicative Fourier analysis (both at ∞ and at finite places) as a substitute for complex analysis in order to expedite various parts of the argument. Also, the emphasis here will be more on heuristics and intuition than on rigour.)

The material here is closely related to the theory of *pretentious characters* developed in [**GrSo2007**], as well as the earlier paper [**Gr1992**].

3.12.1. A heuristic elementary proof of the prime number theorem. To motivate some of the later discussion, let us first give a highly non-rigorous *heuristic* elementary "proof" of the prime number theorem (3.55). Since we clearly have

$$\sum_{n \leq x} 1 = x + O(1)$$

one can view the prime number theorem as an assertion that the von Mangoldt function Λ "behaves like 1 on the average",

$$\Lambda(n) \approx 1, \tag{3.56}$$

where we will be deliberately vague as to what the "$\approx$" symbol means. (One can think of this symbol as denoting some sort of proximity in the *weak topology* or *vague topology*, after suitable normalisation.)

To see why one would expect (3.56) to be true, we take divisor sums of (3.56) to heuristically obtain

$$\sum_{d|n} \Lambda(d) \approx \sum_{d|n} 1. \tag{3.57}$$

By (3.54), the left-hand side is $\log n$; meanwhile, the right-hand side is the *divisor function* $\tau(n)$ of n, by definition. So we have a heuristic relationship between (3.56) and the informal approximation

$$\tau(n) \approx \log n. \tag{3.58}$$

In particular, we expect

$$\sum_{n\leq x} \tau(n) \approx \sum_{n\leq x} \log n. \tag{3.59}$$

The right-hand side of (3.59) can be approximated using the *integral test* as

$$\sum_{n\leq x} \log n = \int_1^x \log t \; dt + O(\log x) = x \log x - x + O(\log x) \tag{3.60}$$

(one can also use *Stirling's formula* to obtain a similar asymptotic). As for the left-hand side, we write $\tau(n) = \sum_{d|n} 1$ and then make the substitution $n = dm$ to obtain

$$\sum_{n\leq x} \tau(n) = \sum_{d,m:dm\leq x} 1.$$

The right-hand side is the number of lattice points underneath the hyperbola $dm = x$, and can be counted using the *Dirichlet hyperbola method*:

$$\sum_{d,m:dm\leq x} 1 = \sum_{d\leq\sqrt{x}} \sum_{m\leq x/d} 1 + \sum_{m\leq\sqrt{x}} \sum_{d\leq x/m} 1 - \sum_{d\leq\sqrt{x}} \sum_{m\leq\sqrt{x}} 1.$$

The third sum is equal to $(\sqrt{x} + O(1))^2 = x + O(\sqrt{x})$. The second sum is equal to the first. The first sum can be computed as

$$\sum_{d\leq\sqrt{x}} \sum_{m\leq x/d} 1 = \sum_{d\leq\sqrt{x}} (\frac{x}{d} + O(1)) = x \sum_{d\leq\sqrt{x}} \frac{1}{d} + O(1);$$

meanwhile, from the *integral test* and the definition of *Euler's constant* $\gamma = 0.577\ldots$ one has

$$\sum_{d\leq y} \frac{1}{d} = \log y + \gamma + O(1/y) \tag{3.61}$$

for any $y \geq 1$; combining all these estimates one obtains

$$\sum_{n\leq x} \tau(n) = x \log x + (2\gamma - 1)x + O(\sqrt{x}). \tag{3.62}$$

Comparing this with (3.60) we do see that $\tau(n)$ and $\log n$ are roughly equal "to top order" on average, thus giving some form of (3.58) and hence (3.57); if one could somehow invert the divisor sum operation, one could hope to get (3.56) and thus the prime number theorem.

(Looking at the next highest order terms in (3.60), (3.62), we see that we expect $\tau(n)$ to in fact be slightly larger than $\log n$ on the average, and so $\Lambda(n)$ should be slightly less than 1 on the average. There is indeed a slight effect of this form; for instance, it is possible (using the prime number theorem) to prove

$$\sum_{d \leq y} \frac{\Lambda(d)}{d} = \log y - \gamma + o(1),$$

which should be compared with (3.61).)

One can partially translate the above discussion into the language of Dirichlet series, by transforming various arithmetical functions $f(n)$ to their associated Dirichlet series

$$F(s) := \sum_{n=1}^{\infty} \frac{f(n)}{n^s},$$

ignoring for now the issue of convergence of this series. By definition, the constant function 1 transforms to the Riemann zeta function $\zeta(s)$. Taking derivatives in s, we see (formally, at least) that if $f(n)$ has Dirichlet series $F(s)$, then $f(n) \log n$ has Dirichlet series $-F'(s)$; thus, for instance, $\log n$ has Dirichlet series $-\zeta'(s)$.

Most importantly, though, if $f(n), g(n)$ have Dirichlet series $F(s), G(s)$ respectively, then their *Dirichlet convolution* $f*g(n) := \sum_{d|n} f(d) g(\frac{n}{d})$ has Dirichlet series $F(s)G(s)$; this is closely related to the well-known ability of the Fourier transform to convert convolutions to pointwise multiplication. Thus, for instance, $\tau(n)$ has Dirichlet series $\zeta(s)^2$. Also, from (3.54) and the preceding discussion, we see that $\Lambda(n)$ has Dirichlet series $-\zeta'(s)/\zeta(s)$ (formally, at least). This already suggests that the von Mangoldt function will be sensitive to the zeroes of the zeta function.

An integral test computation closely related to (3.61) gives the asymptotic

$$\zeta(s) = \frac{1}{s-1} + \gamma + O(s-1)$$

for s close to one (and $\mathrm{Re}(s) > 1$, to avoid issues of convergence). This implies that the Dirichlet series $-\zeta'(s)/\zeta(s)$ for $\Lambda(n)$ has asymptotic

$$\frac{-\zeta'(s)}{\zeta(s)} = \frac{1}{s-1} - \gamma + O(s-1)$$

thus giving support to (3.56); similarly, the Dirichlet series for $\log n$ and $\tau(n)$ have asymptotic

$$-\zeta'(s) = \frac{1}{(s-1)^2} + O(1)$$

and

$$\zeta(s)^2 = \frac{1}{(s-1)^2} + \frac{2\gamma}{s-1} + O(1)$$

which gives support to (3.58) (and is also consistent with (3.60), (3.62)).

Remark 3.12.1. One can connect the properties of Dirichlet series $F(s)$ more rigorously to asymptotics of partial sums $\sum_{n\leq x} f(n)$ by means of various transforms in Fourier analysis and complex analysis, in particular contour integration or the Hilbert transform, but this becomes somewhat technical and we will not do so here. I will remark, though, that asymptotics of $F(s)$ for s close to 1 are not enough, by themselves, to get really precise asymptotics for the sharply truncated partial sums $\sum_{n\leq x} f(n)$, for reasons related to the uncertainty principle; in order to control such sums one also needs to understand the behaviour of F far away from $s = 1$, and in particular for $s = 1 + it$ for large real t. On the other hand, the asymptotics for $F(s)$ for s near 1 are just about all one needs to control *smoothly* truncated partial sums such as $\sum_n f(n)\eta(n/x)$ for suitable cutoff functions η. Also, while Dirichlet series are very powerful tools, particularly with regards to understanding Dirichlet convolution identities, and controlling everything in terms of the zeroes and poles of such series, they do have the drawback that they do not easily encode such fundamental "physical space" facts as the pointwise inequalities $|\mu(n)| \leq 1$ and $\Lambda(n) \geq 0$, which are also an important aspect of the theory.

3.12.2. Almost primes. One can hope to make the above heuristics precise by applying the *Möbius inversion formula*

$$1_{n=1} = \sum_{d|n} \mu(d)$$

where $\mu(d)$ is the *Möbius function*, defined as $(-1)^k$ when d is the product of k distinct primes for some $k \geq 0$, and zero otherwise. In terms of Dirichlet series, we thus see that μ has the Dirichlet series of

$1/\zeta(s)$, and so can invert the divisor sum operation $f(n) \mapsto \sum_{d|n} f(d)$ (which corresponds to multiplication by $\zeta(s)$):

$$f(n) = \sum_{m|n} \mu(m)(\sum_{d|n/m} f(d)).$$

From (3.54) we then conclude

$$\Lambda(n) = \sum_{d|n} \mu(d) \log \frac{n}{d} \tag{3.63}$$

while from $\tau(n) = \sum_{d|n} 1$ we have

$$1 = \sum_{d|n} \mu(d)\tau(\frac{n}{d}). \tag{3.64}$$

One can now hope to derive the prime number theorem (3.55) from the formulae (3.60), (3.62). Unfortunately, this doesn't quite work: the prime number theorem is equivalent to the assertion

$$\sum_{n\leq x} (\Lambda(n) - 1) = o(x), \tag{3.65}$$

but if one inserts (3.63), (3.64) into the left-hand side of (3.65), one obtains

$$\sum_{d\leq x} \mu(d) \sum_{m\leq x/d} (\log m - \tau(m)),$$

which if one then inserts (3.60), (3.62) and the trivial bound $\mu(d) = O(1)$, leads to

$$2Cx \sum_{d\leq x} \frac{\mu(d)}{d} + O(x).$$

Using the elementary inequality

$$|\sum_{d\leq x} \frac{\mu(d)}{d}| \leq 1, \tag{3.66}$$

(see [**Ta2010b**]), we only obtain a bound of $O(x)$ for (3.65) instead of $o(x)$. (A refinement of this argument, though, shows that the prime number theorem would follow if one had the asymptotic $\sum_{n\leq x} \mu(n) = o(x)$, which is in fact equivalent to the prime number theorem.)

We remark that if one computed $\sum_{n\leq x} \tau(n)$ or $\sum_{n\leq x} \Lambda(n)$ by the above methods, one would eventually be led to a variant of (3.66), namely

$$\sum_{d\leq x} \frac{\mu(d)}{d} \log \frac{x}{d} = O(1), \tag{3.67}$$

which is an estimate which will be useful later.

So we see that when trying to sum the von Mangoldt function Λ by elementary means, the error term $O(x)$ overwhelms the main term x. But there is a slight tweaking of the von Mangoldt function, the *second von Mangoldt function* Λ_2, that increases the size of the main term to $2x \log x$ while keeping the error term at $O(x)$, thus leading to a useful estimate; the price one pays for this is that this function is now a proxy for the *almost primes* rather than the primes. This function is defined by a variant of (3.63), namely

$$\Lambda_2(n) = \sum_{d|n} \mu(d) \log^2 \frac{n}{d}. \tag{3.68}$$

It is not hard to see that $\Lambda_2(n)$ vanishes once n has at least three distinct prime factors (basically because the quadratic function $x \mapsto x^2$ vanishes after being differentiated three or more times). Indeed, one can easily verify the identity

$$\Lambda_2(n) = \Lambda(n) \log n + \Lambda * \Lambda(n) \tag{3.69}$$

(which corresponds to the Dirichlet series identity $\zeta''(s)/\zeta(s) = -(-\zeta'(s)/\zeta(s))' + (-\zeta'(s)/\zeta(s))^2$); the first term $\Lambda(n) \log n$ is mostly concentrated on primes, while the second term $\Lambda * \Lambda(n)$ is mostly concentrated on *semiprimes* (products of two distinct primes).

Now let us sum $\Lambda_2(n)$. In analogy with the previous discussion, we will do so by comparing the function $\log^2 n$ with something involving the divisor function. In view of (3.58), it is reasonable to try the approximation

$$\log^2 n \approx \tau(n) \log n;$$

from the identity

$$2 \log n = \sum_{d|n} \mu(d) \tau(\frac{n}{d}) \log \frac{n}{d} \tag{3.70}$$

(which corresponds to the Dirichlet series identity $-2\zeta'(s) = \frac{1}{\zeta(s)} - (\zeta^2(s))'$) we thus expect

$$\Lambda_2(n) \approx 2\log n. \tag{3.71}$$

Now we make these heuristics more precise. From the integral test we have

$$\sum_{n\leq x} \log^2 n = x\log^2 x + C_1 x\log x + C_2 x + O(\log^2 x)$$

while from (3.62) and summation by parts one has

$$\sum_{n\leq x} \tau(n)\log n = x\log^2 x + C_3 x\log x + C_4 x + O(\sqrt{x}\log x)$$

where C_1, C_2, C_3, C_4 are explicit absolute constants whose exact value is not important here. Thus

$$\sum_{n\leq x} (\log^2 n - \tau(n)\log n) = C_5 x\log x + C_6 x + O(\sqrt{x}\log x) \tag{3.72}$$

for some other constants C_5, C_6.

Meanwhile, from (3.68), (3.70) one has

$$\sum_{n\leq x} (\Lambda_2(n) - 2\log(n)) = \sum_{d\leq x} \mu(d) \sum_{m\leq x/d} \log^2 n - \tau(n)\log n;$$

applying (3.72), (3.66), (3.67) we see that the right-hand side is $O(x)$. Computing $\sum_{n\leq x} \log n$ by the integral test, we deduce the *Selberg symmetry formula*

$$\sum_{n\leq x} \Lambda_2(n) = 2x\log x + O(x). \tag{3.73}$$

One can view (3.73) as the "almost prime number theorem" - the analogue of the prime number theorem for almost primes.

The fact that the almost primes have a relatively easy asymptotic, while the genuine primes do not, is a reflection of the *parity problem* in sieve theory; see Section 3.10 of *Structure and Randomness* for further discussion. The symmetry formula is however enough to get "within a factor of two" of the prime number theorem: if we discard

the semiprimes $\Lambda * \Lambda$ from (3.69), we see that $\Lambda(n) \log n \leq \Lambda_2(n)$, and thus

$$\sum_{n \leq x} \Lambda(n) \log n \leq 2x \log x + O(x)$$

which by a summation by parts argument leads to

$$0 \leq \sum_{n \leq x} \Lambda(n) \leq 2x + O(\frac{x}{\log x}),$$

which is within a factor of 2 of (3.55) in some sense.

One can "twist" all of the above arguments by a Dirichlet character χ. For instance, (3.68) twists to

$$\Lambda_2(n)\chi(n) = \sum_{d|n} \mu(d)\chi(d) \log^2 \frac{n}{d} \chi(\frac{n}{d}).$$

On the other hand, if χ is a non-principal character of modulus q, then it has mean zero on any interval with length q, and it is then not hard to establish the asymptotic

$$\sum_{n \leq y} \log^2 n \chi(n) = O_q(\log^2 y).$$

This soon leads to the twisted version of (3.73):

$$\sum_{n \leq x} \Lambda_2(n)\chi(n) = O_q(x), \tag{3.74}$$

thus almost primes are asymptotically unbiased with respect to non-principal characters.

From the multiplicative Fourier analysis of Dirichlet characters modulo q (and the observation that Λ_2 is quite small on residue classes not coprime to q) one then has an "almost prime number theorem in arithmetic progressions":

$$\sum_{n \leq x: n = a \bmod q} \Lambda_2(n) = \frac{2}{\phi(q)} x \log x + O_q(x).$$

As before, this lets us come within a factor of two of the actual prime number theorem in arithmetic progressions:

$$\sum_{n \leq x: n = a \bmod q} \Lambda(n) \leq \frac{2}{\phi(q)} x + O_q(\frac{x}{\log x}).$$

One can also twist things by the completely multiplicative function $n \mapsto n^{it}$, but with the caveat that the approximation $2\log n$ to $\Lambda_2(n)$ can locally correlate with n^{it}. Thus for instance one has

$$\sum_{n\leq x}(\Lambda_2(n) - 2\log n)\chi(n)n^{it} = O_q(x)$$

for any fixed t and χ; in particular, if χ is non-principal, one has

$$\sum_{n\leq x}\Lambda_2(n)\chi(n)n^{it} = O_q(x).$$

3.12.3. The all-or-nothing dichotomy. To summarise so far, the almost primes (as represented by Λ_2) are quite uniformly distributed. These almost primes can be split up into the primes (as represented by $\Lambda(n)\log n$) and the semiprimes (as represented by $\Lambda*\Lambda(n)$), thanks to (3.69).

One can rewrite (3.69) as a recursive formula for Λ:

$$\Lambda(n) = \frac{1}{\log n}\Lambda_2(n) - \frac{1}{\log n}\Lambda * \Lambda(n). \tag{3.75}$$

One can also twist this formula by a character χ and/or a completely multiplicative function $n \mapsto n^{it}$, thus for instance

$$\Lambda\chi(n) = \frac{1}{\log n}\Lambda_2\chi(n) - \frac{1}{\log n}\Lambda\chi * \Lambda\chi(n). \tag{3.76}$$

This recursion, combined with the uniform distribution properties on Λ_2, lead to various *all-or-nothing* dichotomies for Λ. Suppose, for instance, that $\Lambda\chi$ behaves like a constant c on the average for some non-principal character χ:

$$\Lambda\chi(n) \approx c.$$

Then (from (3.58)) we expect $\Lambda\chi * \Lambda\chi$ to behave like $c^2\log n$, thus

$$\frac{1}{\log n}\Lambda\chi * \Lambda\chi(n) \approx c^2.$$

On the other hand, from (3.74), $\frac{1}{\log n}\Lambda_2(n)$ is asymptotically uncorrelated with χ:

$$\frac{1}{\log n}\Lambda_2\chi \approx 0.$$

Putting all this together, one obtains

$$c \approx -c^2$$

which suggests that c must be either close to 0, or close to -1.

Basically, the point is that there are only two equilibria for the recursion (3.76). One equilibrium occurs when Λ is asymptotically uncorrelated with χ; the other is when it is completely anti-correlated with χ, so that $\Lambda(n)$ is supported primarily on those n for which $\chi(n)$ is close to -1. Note in the latter case $\chi(n) \approx -1$ for most primes n, and thus $\chi(n) \approx +1$ for most semiprimes n, thus leading to an equidistribution of $\chi(n)$ for almost primes (weighted by Λ_2). Any intermediate distribution of $\Lambda\chi$ would be inconsistent with the distribution of $\Lambda_2\chi$. (In terms of Dirichlet series, this assertion corresponds to the fact that the L-function of χ either has a zero of order 1, or a zero of order 0 (i.e. not a zero at all) at $s = 1$.)

A similar phenomenon occurs when twisting Λ by n^{it}; basically, the average value of $(\Lambda(n)-1)n^{it}$ must asymptotically either be close to 0, or close to -1; no other asymptotic ends up being compatible with the distribution of $(\Lambda_2(n)-2\log n)n^{it}$. (Again, this corresponds to the fact that the Riemann zeta function has a zero of order 1 or 0 at $1+it$.) More generally, the average value of $(\Lambda(n)-1)\chi(n)n^{it}$ must asymptotically approach either 0 or -1.

Remark 3.12.2. One can make the above heuristics precise either by using Dirichlet series (and analytic continuation, and the theory of zeroes of meromorphic functions), or by smoothing out arithmetic functions such as $\Lambda\chi$ by a suitable multiplicative convolution with a mollifier (as is basically done in elementary proofs of the prime number theorem); see also [**GrSo2007**] for a closely related theory. We will not pursue these details here, however.

3.12.4. Dueling conspiracies. In the previous section we have seen (heuristically, at least), that the von Mangoldt function $\Lambda(n)$ (or more precisely, $\Lambda(n)-1$) will either have no correlation, or a maximal amount of anti-correlation, with a completely multiplicative function such as $\chi(n)$, n^{it}, or $\chi(n)n^{it}$. On the other hand, it is not possible for this function to maximally anti-correlate (or to *conspire*) with two such functions; thus the presence of one conspiracy excludes the presence of all others.

Suppose for instance that we had two distinct non-principal characters χ, χ' for which one had maximal anti-correlation:

$$\Lambda(n)\chi(n), \Lambda(n)\chi'(n) \approx -1.$$

One could then combine the two statements to obtain

$$\Lambda(n)(\chi(n) + \chi'(n)) \approx -2.$$

Meanwhile, $\frac{1}{\log n}\Lambda_2(n)$ doesn't correlate with either χ or χ'. It will be convenient to exploit this to normalise Λ, obtaining

$$(\Lambda(n) - \frac{1}{2\log n}\Lambda_2(n))(\chi(n) + \chi'(n)) \approx -2.$$

(Note from (3.56), (3.71) that we expect $\Lambda(n) - \frac{1}{2\log n}\Lambda_2(n)$ to have mean zero.)

On the other hand, since $0 \leq \Lambda(n)\log n \leq \Lambda_2(n)$, one has

$$|\Lambda(n) - \frac{1}{2\log n}\Lambda_2(n)| \leq \frac{1}{2\log n}\Lambda_2(n)$$

and hence by the triangle inequality

$$\Lambda_2(n)|\chi(n) + \chi'(n)| \gtrapprox 4\log n$$

in the sense that averages of the left-hand side should be at least as large as averages of the right-hand side. From this, (3.71), and Cauchy-Schwarz, one thus expects

$$\Lambda_2(n)|\chi(n) + \chi'(n)|^2 \gtrapprox 8\log n.$$

But if one expands out the left-hand side using (3.71), (3.74), one only ends up with $4\log n + O_q(1)$ on the average, a contradiction for n sufficiently large.

Remark 3.12.3. The above argument belongs to a family of L^2-based arguments which go by various names (*almost orthogonality*, TT^*, *large sieve*, etc.). The L^2 argument can more generally be used to establish square-summability estimates on averages such as $\frac{1}{x}\sum_{n\leq x}\Lambda(n)\chi(n)$ as χ varies, but we will not make this precise here.

As one consequence of the above arguments, one can show that $\Lambda(n)$ cannot maximally anti-correlate with any non-real character χ, since (by the reality of Λ) it would then also maximally anti-correlate with the complex conjugate $\overline{\chi}$, which is distinct from χ. A similar

argument shows that $\Lambda(n)$ cannot maximally anti-correlate with n^{it} for any non-zero t, a fact which can soon lead to the prime number theorem, either by Dirichlet series methods, by Fourier-analytic means, or by elementary means. (Sketch of Fourier-analytic proof: L^2 methods provide L^2-type bounds on the averages of $\Lambda(n)n^{it}$ in t, while the above arguments show that these averages are also small in L^∞. Applying (3.75) a few times to take advantage of the smoothing effects of convolution, one eventually concludes that these averages can be made arbitrarily small in L^1 asymptotically, at which point the prime number theorem follows from Fourier inversion.)

Remark 3.12.4. There is a slightly different argument of an L^1 nature rather than an L^2 nature (i.e. using tools such as the triangle inequality, union bound, etc.) that can also achieve similar results. For instance, suppose that $\Lambda(n)$ maximally anti-correlates with χ and χ'. Then $\chi(n), \chi'(n) \approx -1$ for most primes n, which implies that $\chi\chi'(n) \approx +1$ for most primes n, which is incompatible with the all-or-nothing dichotomy unless $\chi\chi'$ is principal. This is an alternate way to exclude correlation with non-real characters. Similarly, if $\Lambda(n)n^{it} \approx -1$, then $\Lambda(n)n^{2it} \approx +1$, which is also incompatible with the zero-one law; this is essentially the method underlying the standard proof of the prime number theorem (which relates $\zeta(1+it)$ with $\zeta(1+2it)$).

3.12.5. Quadratic characters. The one difficult scenario to eliminate is that of maximal anti-correlation with a real non-principal (i.e. quadratic) character χ, thus

$$\Lambda(n)\chi(n) \approx -1.$$

This scenario implies that the quantity

$$L(1,\chi) := \sum_{n=1}^{\infty} \frac{\chi(n)}{n}$$

vanishes. Indeed, if one starts with the identity

$$\log n\chi(n) = \sum_{d|n} \Lambda\chi(d)\chi(\frac{n}{d})$$

and sums in n, one sees that

$$\sum_{n\leq x} \log n\chi(n) = \sum_{d,m:dm\leq x} \Lambda\chi(d)\chi(m).$$

The left-hand side is $O_q(\log x)$ by the mean zero and periodicity properties of χ. To estimate the right-hand side, we use the hyperbola method and rewrite it as

$$\sum_{m \leq M} \chi(m) \sum_{d \leq x/m} \Lambda\chi(d) + \sum_{d \leq x/M} \Lambda\chi(d) \sum_{M < m \leq x/d} \chi(m)$$

for some parameter M (sufficiently slowly growing in x) to be optimised later. Writing $\sum_{d \leq x/m} \Lambda\chi(d) = (-1+o_q(1))x/m$ and $\sum_{M<m\leq x/d} \chi(m) = O_q(1)$, we can express this as

$$x(\sum_{m \leq M} \frac{\chi(m)}{m} + o_q(1)) + O_q(x/M);$$

sending $x \to \infty$ (and $M \to \infty$ at a slower rate) we conclude $L(1,\chi) = 0$ as required.

It is remarkably difficult to show that $L(1,\chi)$ does not, in fact, vanish. One way to do this is to use the *class number formula*, that relates this quantity to the class number of the quadratic number field $\mathbf{Q}(\sqrt{-d})$ associated to the conductor d of χ, together with some related number-theoretic quantities. A more elementary (but significantly weaker) method proceeds by using the easily verified fact that the convolution $1 * \chi$ is non-negative, and is at least 1 on the squares; this should be interpreted as a fact from algebraic number theory, and basically corresponds to the fact that the number of representations of an integer n as the norm $x^2 + dy^2$ of an integer in $\mathbf{Z}(\sqrt{d})$ (or more generally, as the norm of an ideal in that ring) is non-negative, and is at least 1 on the squares. In particular we have

$$\sum_{n \leq x} \frac{1 * \chi(n)}{\sqrt{n}} \geq \frac{1}{2} \log x + O(1).$$

On the other hand, from the hyperbola method we can express the left-hand side as

$$\sum_{d \leq \sqrt{x}} \frac{\chi(d)}{\sqrt{d}} \sum_{m \leq x/d} \frac{1}{\sqrt{m}} + \sum_{m < \sqrt{x}} \frac{1}{\sqrt{m}} \sum_{\sqrt{x} < d \leq x/m} \frac{\chi(d)}{\sqrt{d}}. \tag{3.77}$$

From the mean zero and periodicity properties of χ we have $\sum_{\sqrt{x}<d\leq x/m} \frac{\chi(d)}{\sqrt{d}} = O_q(x^{-1/4})$, so the second term in (3.77) is $O_q(1)$. Meanwhile, from

the *midpoint rule*, $\sum_{m\leq y}\frac{1}{\sqrt{m}} = 2\sqrt{y} - \frac{3}{2} + O(1/\sqrt{y})$, and so the first term in (3.77) is

$$2\sqrt{x}\sum_{d\leq\sqrt{x}}\frac{\chi(d)}{d} + O(|\sum_{d\leq\sqrt{x}}\frac{\chi(d)}{\sqrt{d}}|) + O(1) = 2\sqrt{x}L(1,\chi) + O(1).$$

Putting all this together we have

$$\frac{1}{2}\log x + O(1) \leq 2\sqrt{x}L(1,\chi) + O_q(1),$$

which leads to a contradiction as $x\to\infty$ if $L(1,\chi)$ vanishes.

Note in fact that the above argument shows that $L(1,\chi)$ is positive. If one carefully computes the dependence of the above argument on the modulus q, one obtains a lower bound of the form $L(1,\chi) \geq \exp(-q^{1/2+o(1)})$, which is quite poor. Using a non-trivial improvement on the error term in counting lattice points under the hyperbola (or better still, by smoothing the sum $\sum_{n\leq x}$), one can improve this a bit, to $L(1,\chi) \geq q^{-O(1)}$. In contrast, the class number method gives a bound $L(1,\chi) \geq q^{-1/2+o(1)}$.

We can improve this even further for all but at most one real primitive character χ:

Theorem 3.12.5 (Siegel's theorem)**.** *For every $\varepsilon > 0$, one has $L(1,\chi) \gg_\varepsilon q^{-\varepsilon}$ for all but at most one real primitive character χ, where the implied constant is effective, and q is the modulus of χ.*

Throwing in this (hypothetical) one exceptional character, we conclude that $L(1,\chi) \gg_\varepsilon q^{-\varepsilon}$ for *all* real primitive characters χ, but now the implied constant is ineffective, which is the usual way in which Siegel's theorem is formulated (but the above nearly effective refinement can be obtained by the same methods). It is a major open problem in the subject to eliminate this exceptional character and recover an effective estimate for some $\varepsilon < 1/2$.

Proof. Let $\varepsilon > 0$ (which we can assume to be small), and let $c > 0$ be a small number depending (effectively) on ε to be chosen later. Our task is to show that $L(1,\chi) \geq cq^{-\varepsilon}$ for all but at most one primitive real character χ. Note we may assume q is large (effectively) depending on ε, as the claim follows from the previous bounds on $L(1,\chi)$ otherwise.

Suppose then for contradiction that $L(1,\chi) < cq^{-\varepsilon}$ and $L(1,\chi') < c(q')^{-\varepsilon}$ for two distinct primitive real characters χ, χ' of (large) modulus q, q' respectively.

We begin by modifying the proof that $L(1,\chi)$ was positive, which relied (among other things) on the observation that $1 * \chi$, and equals 1 at 1. In particular, one has

$$\sum_{n \leq x} \frac{1 * \chi(n)}{n^s} \geq 1 \tag{3.78}$$

for any $x \geq 1$ and any real s. (One can get slightly better bounds by exploiting that $1 * \chi$ is also at least 1 on square numbers, as before, but this is really only useful for $s \leq 1/2$, and we are now going to take s much closer to 1.)

On the other hand, one has the asymptotics

$$\sum_{n \leq x} \frac{1}{n^s} = \zeta(s) + \frac{x^{1-s}}{1-s} + O(x^{-s})$$

for any real s close (but not equal) to 1, and similarly

$$\sum_{n \leq x} \frac{\chi(n)}{n^s} = L(s,\chi) + O(q^{O(1)} x^{-s})$$

for any real s close to 1; similarly for $\chi', \chi\chi'$. From the hyperbola method, we can then conclude

$$\sum_{n \leq x} \frac{1 * \chi(n)}{n^s} = \zeta(s) L(s,\chi) + \frac{x^{1-s}}{1-s} L(1,\chi) + O(q^{O(1)} x^{0.5-s}) \tag{3.79}$$

for all real s sufficiently close to 1. Indeed, one can expand the left-hand side of (3.79) as

$$\sum_{d \leq \sqrt{x}} \frac{\chi(d)}{d^s} \sum_{m \leq x/d} \frac{1}{m^s} + \sum_{m < \sqrt{x}} \frac{1}{m^s} \sum_{\sqrt{x} < d \leq x/m} \frac{\chi(d)}{d^s}$$

and the claim then follows from the previous asymptotics. (One can improve the error term by smoothing the summation, but we will not need to do so here.)

Now set $x = Cq^C$ for a large absolute constant C. If $0.99 \leq s < 1$, then the error term in $O(q^{O(1)} x^{0.5-s})$ is then at most $1/2$ (say) if C

is large enough. We conclude from (3.78) that

$$\zeta(s)L(s,\chi) \geq \frac{1}{2} - O(\frac{q^{O(1-s)}}{1-s}L(1,\chi))$$

for $0.99 \leq s < 1$. Since $L(1,\chi) \leq cq^{-\varepsilon}$ and c is assumed small (depending on ε), this implies that $\zeta(s)L(s,\chi)$ is positive in the range

$$L(1,\chi) \ll 1-s \ll \varepsilon$$

(this can be seen by checking the cases $1-s \leq 1/\log q$ and $1-s > 1/\log q$ separately). On the other hand, $\zeta(s)L(s,\chi)$ has a simple pole at $s=1$ with positive residue, and is thus negative for $s<1$ extremely close to 1. By the intermediate value theorem, we conclude that $L(s,\chi)$ has a zero for some $s = 1 - O(L(1,\chi))$. Conversely, it is not difficult (using summation by parts) to show that $L'(s,\chi) = O(\log^2 q)$ for $s = 1 - O(1/\log q)$, and so by the mean value theorem we see that the zero of $L(s,\chi)$ must also obey $1-s \gg L(1,\chi)/\log^2 q$. Thus $L(s,\chi)$ has a zero for some $s<1$ with

$$L(1,\chi)/\log^2 q \ll 1-s \ll L(1,\chi). \tag{3.80}$$

Similarly, $L(s',\chi')$ has a zero for some $s'<1$ with

$$L(1,\chi')/\log^2 q' \ll 1-s' \ll L(1,\chi'). \tag{3.81}$$

Now, we consider the function

$$f := 1 * \chi * \chi' * \chi\chi'.$$

One can also show that f is non-negative and equals 1 at 1, thus

$$\sum_{n\leq x} \frac{f(n)}{n^s} \geq 1.$$

(The algebraic number theory interpretation of this positivity is that $f(n)$ is the number of representations of n as the norm of an ideal in the *biquadratic field* generated by $\sqrt{q}$ and $\sqrt{q'}$.)

Also, by (a more complicated version of) the derivation of (3.79), one has

$$\sum_{n\leq x} \frac{f(n)}{n^s} = \zeta(s)L(s,\chi)L(s,\chi')L(s,\chi\chi') + \frac{x^{1-s}}{1-s}L(1,\chi)L(1,\chi')L(1,\chi\chi') + O((qq')^{O(1)}x^{0.9-s})$$

(say). Arguing as before, we conclude that

$$\zeta(s)L(s,\chi)L(s,\chi')L(s,\chi\chi') \geq \frac{1}{2} - O(\frac{(qq')^{O(1-s)}}{1-s} L(1,\chi)L(1,\chi')L(1,\chi\chi'))$$

for $0.99 \leq s < 1$. Using the bound $L(1,\chi\chi') \ll \log(qq')$ (which can be established by summation by parts), we conclude that $\zeta(s)L(s,\chi)L(s,\chi')L(s,\chi\chi')$ is positive in the range

$$L(1,\chi)L(1,\chi')\log(qq') \ll 1-s \ll \varepsilon.$$

Since we already know $L(s,\chi)$ and $L(s',\chi')$ have zeroes for some s, s' obeying (3.80), (3.81)

$$\frac{L(1,\chi)}{\log^2 q}, \frac{L(1,\chi')}{\log^2 q'} \ll L(1,\chi)L(1,\chi')\log(qq');$$

taking geometric means and rearranging we obtain

$$L(1,\chi)L(1,\chi') \gg \log(qq')^{-O(1)}.$$

But this contradicts the hypotheses $L(1,\chi) \leq cq^{-\varepsilon}$, $L(1,\chi') \leq c(q')^{-\varepsilon}$ if c is small enough. □

Remark 3.12.6. Siegel's theorem leads to a version of the prime number theorem in arithmetic progressions known as the *Siegel-Walfisz theorem*. As with Siegel's theorem, the bounds are ineffective unless one is allowed to exclude a single exceptional modulus q (and its multiples), in which case one has a modified prime number theorem which favours the quadratic nonresidues mod q; see [**Gr1992**].

Remark 3.12.7. One can improve the effective bounds in Siegel's theorem if one is allowed to exclude a larger set of bad moduli. For instance, the arguments in Section 3.12.4 allow one to establish a bound of the form $L(1,\chi) \gg \log^{-O(1)} q$ after excluding at most one q in each hyper-dyadic range $2^{100^k} \leq q \leq 2^{100^{k+1}}$ for each k; one can of course replace 100 by other exponents here, but at the cost of worsening the $O(1)$ term. (This is essentially an observation of Landau.)

Notes. This article first appeared at `terrytao.wordpress.com/2009/09/24`. Thanks to anonymous commenters for corrections.

David Speyer noted the connection between Siegel's theorem and the classification of imaginary quadratic fields with unique factorisation.

3.13. Mazur's swindle

Let d be a natural number. A basic operation in the topology of oriented, connected, compact, d-dimensional manifolds (hereby referred to simply as *manifolds* for short) is that of *connected sum*: given two manifolds M, N, the connected sum $M\#N$ is formed by removing a small ball from each manifold and then gluing the boundary together (in the orientation-preserving manner). This gives another oriented, connected, compact manifold, and the exact nature of the balls removed and their gluing is not relevant for topological purposes (any two such procedures give homeomorphic manifolds). It is easy to see that this operation is associative and commutative up to homeomorphism, thus $M\#N \cong N\#M$ and $(M\#N)\#O \cong M\#(N\#O)$, where we use $M \cong N$ to denote the assertion that M is homeomorphic to N.

(It is important that the orientation is preserved; if, for instance, $d = 3$, and M is a chiral 3-manifold which is *chiral* (thus $M \ncong -M$, where $-M$ is the orientation reversal of M), then the connect sum $M\#M$ of M with itself is also chiral (by the *prime decomposition*; in fact one does not even need the irreducibility hypothesis for this claim), but $M\# - M$ is not. A typical example of an irreducible chiral manifold is the complement of a *trefoil knot*. Thanks to Danny Calegari for this example.)

The d-dimensional sphere S^d is an identity (up to homeomorphism) of connect sum: $M\#S^d \cong M$ for any M. A basic result in the subject is that the sphere is itself irreducible:

Theorem 3.13.1 (Irreducibility of the sphere)**.** *If $S^d \cong M\#N$, then $M, N \cong S^d$.*

For $d = 1$ (curves), this theorem is trivial because the only connected 1-manifolds are homeomorphic to circles. For $d = 2$ (surfaces), the theorem is also easy by considering the *genus* of $M, N, M\#N$. For $d = 3$ the result follows from the *prime decomposition*. But for higher

d, these *ad hoc* methods no longer work. Nevertheless, there is an elegant proof of Theorem 3.13.1, due to Mazur[**Ma1959**], and known as *Mazur's swindle*. The reason for this name should become clear when one sees the proof, which I reproduce below.

Suppose $M\#N \cong S^d$. Now consider the infinite connected sum

$$(M\#N)\#(M\#N)\#(M\#N)\#\ldots.$$

This is an infinite connected sum of spheres, and can thus be viewed as a half-open cylinder, which is topologically equivalent to a sphere with a small ball removed; alternatively, one can contract the boundary at infinity to a point to recover the sphere S^d. On the other hand, by using the associativity of connected sum (which will still work for the infinite connected sum, if one thinks about it carefully), the above manifold is also homeomorphic to

$$M\#(N\#M)\#(N\#M)\#\ldots$$

which is the connected sum of M with an infinite sequence of spheres, or equivalently M with a small ball removed. Contracting the small balls to a point, we conclude that $M \cong S^d$, and a similar argument gives $N \cong S^d$.

A typical corollary of Theorem 3.13.1 is a generalisation of the *Jordan curve theorem*: any *locally flat* embedded copy of S^{d-1} in S^d divides the sphere S^d into two regions homeomorphic to balls B^d. (Some sort of regularity hypothesis, such as local flatness, is essential, thanks to the counterexample of the *Alexander horned sphere*. If one assumes smoothness instead of local flatness, the problem is known as the *Schönflies problem*, and is apparently quite subtle, especially in the four-dimensional case $d = 4$.)

One can ask whether there is a way to prove Theorem 3.13.1 for general d without recourse to the infinite sum swindle. I do not know the complete answer to this, but some evidence against this hope can be seen by noting that if one works in the smooth category instead of the topological category (i.e. working with smooth manifolds, and only equating manifolds that are diffeomorphic, and not merely homeomorphic), then the *exotic spheres* in five and higher dimensions provide a counterexample to the smooth version of Theorem 3.13.1: it is

possible to find two exotic spheres whose connected sum is diffeomorphic to the standard sphere. (Indeed, in five and higher dimensions, the exotic sphere structures on S^d form a finite abelian group under connect sum, with the standard sphere being the identity element. The situation in four dimensions is much less well understood.) The problem with the swindle here is that the homeomorphism generated by the infinite number of applications of the associativity law is not smooth when one identifies the boundary with a point.

The basic idea of the swindle - grouping an alternating infinite sum in two different ways - also appears in a few other contexts. Most classically, it is used to show that the sum $1-1+1-1+\ldots$ does not converge in any sense which is consistent with the infinite associative law, since this would then imply that $1 = 0$; indeed, one can view the swindle as a dichotomy between the infinite associative law and the presence of non-trivial cancellation. (In the topological manifold category, one has the former but not the latter, whereas in the case of $1 - 1 + 1 - 1 + \ldots$, one has the latter but not the former.) The *alternating series test* can also be viewed as a variant of the swindle.

Another variant of the swindle arises in the proof of the *Cantor-BernsteinSchröder theorem*. Suppose one has two sets A, B, together with injections from A to B and from B to A. The first injection leads to an identification $B \cong C \uplus A$ for some set C, while the second injection leads to an identification $A \cong D \uplus B$. Iterating this leads to identifications

$$A \cong (D \uplus C \uplus D \uplus \ldots) \uplus X$$

and

$$B \cong (C \uplus D \uplus C \uplus \ldots) \uplus X$$

for some additional set X. Using the identification $D \uplus C \cong C \uplus D$ then yields an explicit bijection between A and B.

Notes. This article first appeared at `terrytao.wordpress.com/2009/10/05`. Thanks to Jan, Peter, and an anonymous commenter for corrections.

Thanks to Danny Calegari for telling me about the swindle, while we were both waiting to catch an airplane.

Several commenters provided further examples of swindle-type arguments. Scott Morrison noted that Mazur's argument also shows

that non-trivial knots do not have inverses: one cannot untie a knot by tying another one. Qiaochu Yuan provided a swindle argument that showed that $GL(H)$ is contractible for any infinite-dimensional Hilbert space H. In a similar spirit, Pace Nielsen recalled the Eilenberg swindle that shows that for every projective module P, there exists a free module F with $P \oplus F \equiv F$. Tim Gowers also mentioned Pelczynski's decomposition method in the theory of Banach spaces as a similar argument.

3.14. Grothendieck's definition of a group

In his wonderful article [**Th1994**],, Bill Thurston describes (among many other topics) how one's understanding of given concept in mathematics (such as that of the derivative) can be vastly enriched by viewing it simultaneously from many subtly different perspectives; in the case of the derivative, he gives seven standard such perspectives (infinitesimal, symbolic, logical, geometric, rate, approximation, microscopic) and then mentions a much later perspective in the sequence (as describing a flat connection for a graph).

One can of course do something similar for many other fundamental notions in mathematics. For instance, the notion of a *group* G can be thought of in a number of (closely related) ways, such as the following:

(0) **Motivating examples**: A group is an abstraction of the operations of addition/subtraction or multiplication/division in arithmetic or linear algebra, or of composition/inversion of transformations.

(1) **Universal algebraic**: A group is a set G with an identity element e, a unary inverse operation $\cdot^{-1} : G \to G$, and a binary multiplication operation $\cdot : G \times G \to G$ obeying the relations (or axioms) $e \cdot x = x \cdot e = x$, $x \cdot x^{-1} = x^{-1} \cdot x = e$, $(x \cdot y) \cdot z = x \cdot (y \cdot z)$ for all $x, y, z \in G$.

(2) **Symmetric**: A group is all the ways in which one can transform a space V to itself while preserving some object or structure O on this space.

(3) **Representation theoretic**: A group is identifiable with a collection of transformations on a space V which is closed under composition and inverse, and contains the identity transformation.

(4) **Presentation theoretic**: A group can be generated by a collection of generators subject to some number of relations.

(5) **Topological**: A group is the fundamental group $\pi_1(X)$ of a connected topological space X.

(6) **Dynamic**: A group represents the passage of time (or of some other variable(s) of motion or action) on a (reversible) dynamical system.

(7) **Category theoretic**: A group is a category with one object, in which all morphisms have inverses.

(8) **Quantum**: A group is the classical limit $q \to 0$ of a quantum group.

- etc.

One can view a large part of group theory (and related subjects, such as representation theory) as exploring the interconnections between various of these perspectives. As one's understanding of the subject matures, many of these formerly distinct perspectives slowly merge into a single unified perspective.

From a recent talk by Ezra Getzler, I learned a more sophisticated perspective on a group, somewhat analogous to Thurston's example of a sophisticated perspective on a derivative (and coincidentally, flat connections play a central role in both):

(37) **Sheaf theoretic**: A group is identifiable with a (set-valued) sheaf on the category of simplicial complexes such that the morphisms associated to collapses of d-simplices are bijective for $d > 1$ (and merely surjective for $d \leq 1$).

This interpretation of the group concept is apparently due to Grothendieck, though it is motivated also by homotopy theory. One of the key advantages of this interpretation is that it generalises easily to the notion of an *n-group* (simply by replacing 1 with n in (37)), whereas the other interpretations listed earlier require a certain

amount of subtlety in order to generalise correctly (in particular, they usually themselves require higher-order notions, such as *n-categories*).

The connection of (37) with any of the other perspectives of a group is elementary, but not immediately obvious; I enjoyed working out exactly what the connection was, and thought it might be of interest to some readers here, so I reproduce it below the fold.

3.14.1. Flat connections. To see the relationship between (37) and more traditional concepts of a group, such as (1), we will begin by recalling the machinery of flat connections.

Let G be a group, X be a topological space. A *principal G-connection* ω on X can be thought of as an assignment of a group element $\omega(\gamma) \in G$ to every path γ in X which obey the following four properties:

- Invariance under reparameterisation: if γ' is a reparameterisation of γ, then $\omega(\gamma) = \omega(\gamma')$.
- Identity: If γ is a constant path, then $\omega(\gamma)$ is the identity element.
- Inverse: If $-\gamma$ is the reversal of a path γ, then $\omega(-\gamma)$ is the inverse of $\omega(\gamma)$.
- Groupoid homomorphism: If γ_2 starts where γ_1 ends (so that one can define the concatenation $\gamma_1 + \gamma_2$), then $\omega(\gamma_1 + \gamma_2) = \omega(\gamma_2)\omega(\gamma_1)$. (Depending on one's conventions, one may wish to reverse the order of the group multiplication on the right-hand side.)

Intuitively, $\omega(\gamma)$ represents a way to use the group G to connect (or "parallel transport") the fibre at the initial point of γ to the fibre at the final point; see Section 1.4 of *Poincaré's Legacies, Vol. II* for more discussion. Note that the identity property is redundant, being implied by the other three properties.

We say that a connection ω is *flat* if $\omega(\gamma)$ is the identity element for every "short" closed loop γ, thus strengthening the identity property. One could define "short" rigorously (e.g. one could use "*contractible*" as a substitute), but we will prefer here to leave the concept intentionally vague.

Typically, one studies connections when the structure group G and the base space X are continuous rather than discrete. However, there is a combinatorial model for connections which is suitable for discrete groups, in which the base space X is now an *(abstract) simplicial complex* Δ - a vertex set V, together with a number of *simplices* in V, by which we mean ordered $d+1$-tuples $(x_0, \ldots, x_d)$ of distinct vertices in V for various integers d (with d being the *dimension* of the simplex $(x_0, \ldots, x_d)$). In our definition of a simplicial complex, we add the requirement that if a simplex lies in the complex, then all faces of that simplex (formed by removing one of the vertices, but leaving the order of the remaining vertices unchanged) also lie in the complex. We also assume a well defined *orientation*, in the sense that every $d+1$-tuple $\{x_0, \ldots, x_d\}$ is represented by at most one simplex (thus, for instance, a complex cannot contain both an edge $(0,1)$ and its reversal $(1,0)$). Though it will not matter too much here, one can think of the vertex set V here as being restricted to be finite.

A *path* γ in a simplicial complex Δ is then a sequence of 1-simplices (x_i, x_{i+1}) or their formal reverses $-(x_i, x_{i+1})$, with the final point of each 1-simplex being the initial point of the next. If G is a (discrete) group, a *principal G-connection* ω on Δ is then an assignment of a group element $\omega(\gamma) \in G$ to each such path γ, obeying the groupoid homomorphism property and the inverse property (and hence the identity property). Note that the reparameterisation property is no longer needed in this abstract combinatorial model. Note that a connection can be determined by the group elements $\omega(b \leftarrow a)$ it assigns to each 1-simplex (a,b). (I have written the simplex $b \leftarrow a$ from right to left, as this makes the composition law cleaner.)

So far, only the 1-skeleton (i.e. the simplices of dimension at most 1) of the complex have been used. But one can use the 2-skeleton to define the notion of a *flat* connection: we say that a principal G-connection ω on Δ is flat if the boundary of every 2-simplex (a,b,c), oriented appropriately, is assigned the identity element, or more precisely that $\omega(c \leftarrow a)^{-1}\omega(c \leftarrow b)\omega(b \leftarrow a) = e$, or in other words that $\omega(c \leftarrow a) = \omega(c \leftarrow b)\omega(b \leftarrow a)$; thus, in this context, a "short loop"

means a loop that is the boundary of a 2-simplex. Note that this corresponds closely to the topological concept of a flat connection when applied to, say, a triangulated manifold.

Fix a group G. Given any simplicial complex Δ, let $\mathcal{O}(\Delta)$ be the set of flat connections on Δ. One can get some feeling for this set by considering some basic examples:

- If Δ is a single 0-dimensional simplex (i.e. a point), then there is only the trivial path, which must be assigned the identity element e of the group. Thus, in this case, $\mathcal{O}(\Delta)$ can be identified with $\{e\}$.
- If Δ is a 1-dimensional simplex, say $(0, 1)$, then the path from 0 to 1 can be assigned an arbitrary group element $\omega(1 \leftarrow 0) \in G$, and this is the only degree of freedom in the connection. So in this case, $\mathcal{O}(\Delta)$ can be identified with G.
- Now suppose Δ is a 2-dimensional simplex, say $(0, 1, 2)$. Then the group elements $\omega(1 \leftarrow 0)$ and $\omega(2 \leftarrow 1)$ are arbitrary elements of G, but $\omega(2 \leftarrow 0)$ is constrained to equal $\omega(2 \leftarrow 1)\omega(1 \leftarrow 0)$. This determines the entire flat connection, so $\mathcal{O}(\Delta)$ can be identified with G^2.
- Generalising this example, if Δ is a k-dimensional simplex, then $\mathcal{O}(\Delta)$ can be identified with G^k.

An important operation one can do on flat connections is that of *pullback*. Let $\phi : \Delta \to \Delta'$ be a *morphism* from one simplicial complex Δ to another Δ'; by this, we mean a map from the vertex set of Δ to the vertex set of Δ' such that every simplex in Δ maps to a simplex in Δ' in an order preserving manner (though note that ϕ is allowed to be non-injective, so that the dimension of the simplex can decrease by mapping adjacent vertices to the same vertex). Given such a morphism, and given a flat connection ω' on Δ', one can then pull back that connection to yield a flat connection $\phi^*\omega'$ on Δ, defined by the formula

$$\phi^*\omega'(w \leftarrow v) := \omega'(\phi(w) \leftarrow \phi(v))$$

for any 1-simplex (v, w) in Δ, with the convention that $\omega'(u \leftarrow u)$ is the identity for any u. It is easy to see that this is still a flat connection. Also, if $\phi : \Delta \to \Delta'$ and $\psi : \Delta' \to \Delta''$ are morphisms, then the operations of pullback by ψ and then by ϕ compose to equal the operation of pullback by $\psi \circ \phi$: $\phi^* \psi^* = (\psi \circ \phi)^*$. In the language of category theory, pullback is a contravariant functor from the category of simplicial complexes to the category of sets (with each simplicial complex being mapped to its set of flat connections).

A special case of a morphism is an *inclusion morphism* $\iota : \Delta \to \Delta'$ to a simplicial complex Δ' from a subcomplex Δ. The associated pullback operation is the *restriction* operation, that maps a flat connection ω' on Δ' to its restriction $\omega' \restriction_\Delta$ to Δ.

3.14.2. Sheaves. We currently have a set $\mathcal{O}(\Delta)$ of flat connections assigned to each simplicial complex Δ, together with pullback maps (and in particular, restriction maps) connecting these sets to each other. One can easily observe that this system of structures obeys the following axioms:

- (Identity) There is only one flat connection on a point.
- (Locality) If $\Delta = \Delta_1 \cup \Delta_2$ is the union of two simplicial complexes, then a flat connection on Δ is determined by its restrictions to Δ_1 and Δ_2. In other words, the map $\omega \mapsto (\omega \restriction_{\Delta_1}, \omega \restriction_{\Delta_2})$ is an injection from $\mathcal{O}(\Delta)$ to $\mathcal{O}(\Delta_1) \times \mathcal{O}(\Delta_2)$.
- (Gluing) If $\Delta = \Delta_1 \cup \Delta_2$, and ω_1, ω_2 are flat connections on Δ_1, Δ_2 which agree when restricted to $\Delta_1 \cap \Delta_2$, (and if the orientations of Δ_1, Δ_2 on the intersection $\Delta_1 \cap \Delta_2$ agree) then there exists a flat connection ω on Δ which agrees with ω_1, ω_2 on Δ_1, Δ_2. (Note that this gluing of ω_1 and ω_2 is unique, by the previous axiom. It is important that the orientations match; we cannot glue $(0, 1)$ to $(1, 0)$, for instance.)

One can consider more abstract assignments of sets to simplicial complexes, together with pullback maps, which obey these three axioms. A system which obeys the first two axioms is known as a *presheaf*, while a system that obeys all three is known as a *sheaf*. (One can also consider pre-sheaves and sheaves on more general topological

spaces than simplicial complexes, for instance the spaces of smooth (or continuous, or holomorphic, etc.) functions (or forms, sections, etc.) on open subsets of a manifold form a sheaf.)

Thus, flat connections associated to a group G form a sheaf. But flat connections form a special type of sheaf that obeys an additional property (listed above as (37)). To explain this property, we first consider a key example when $\Delta = (0,1,2)$ is the standard 2-simplex (together with subsimplices), and Δ' is the subcomplex formed by removing the 2-face $(0,1,2)$ and the 1-face $(0,2)$, leaving only the 1-faces $(0,1), (1,2)$ and the 0-faces $0, 1, 2$. Then of course every flat connection on Δ restricts to a flat connection on Δ'. But the flatness property ensures that this restriction is invertible: given a flat connection on Δ', there exists a unique extension of this connection back to Δ. This is nothing more than the property, remarked earlier, that to specify a flat connection on the 2-simplex $(0,1,2)$, it suffices to know what the connection is doing on $(0,1)$ and $(1,2)$, as the behaviour on the remaining edge can then be deduced from the group law; conversely, any specification of the connection on those two 1-simplices determines a connection on the remainder of the 2-simplex.

This observation can be generalised. Given any simplicial complex Δ, define a k-dimensional *collapse* Δ' of Δ to be a simplicial complex obtained from Δ by removing the interior of a k-simplex, together with one of its faces; thus for instance the complex consisting of $(0,1), (1,2)$ (and subsimplices) is a 2-dimensional collapse of the 2-simplex $(0,1,2)$ (and subsimplices). We then see that the sheaf of flat connections obeys an additional axiom:

- (Grothendieck's axiom) If Δ' is a k-dimensional collapse of Δ, then the restriction map from $\mathcal{O}(\Delta)$ to $\mathcal{O}(\Delta')$ is surjective for all k, and bijective for $k \geq 2$.

This axiom is trivial for $k = 0$. For $k = 1$, it is true because if an edge (and one of its vertices) can be removed from a complex, then it is not the boundary of any 2-simplex, and the value of a flat connection on that edge is thus completely unconstrained. (In any event, the $k = 1$ case of this axiom can be deduced from the sheaf axioms.) For $k = 2$, it follows because if one can remove a 2-simplex and one of its edges from a complex, then the edge is not

the boundary of any other 2-simplex and thus the connection on that edge only constrained to precisely be the product of the connection on the other two edges of the 2-simplex. For $k = 3$, it follows because if oen removes a 3-simplex and one of its 2-simplex faces, the constraint associated to that 2-simplex is implied by the constraints coming from the other three faces of the 3-simplex (I recommend drawing a tetrahedron and chasing some loops around to see this), and so one retains bijectivity. For $k \geq 4$, the axiom becomes trivial again because the k-simplices and $k-1$-simplices have no impact on the definition of a flat connection.

Grothendieck's beautiful observation is that the converse holds: if a (concrete) sheaf $\Delta \mapsto \mathcal{O}(\Delta)$ obeys Grothendieck's axiom, then it is equivalent to the sheaf of flat connections of some group G defined canonically from the sheaf. Let's see how this works. Suppose we have a sheaf $\Delta \mapsto \mathcal{O}(\Delta)$, which is concrete in the sense that $\mathcal{O}(\Delta)$ is a set, and the morphisms between these sets are given by functions. In analogy with the preceding discussion, we'll refer to elements of $\mathcal{O}(\Delta)$ as (abstract) flat connections, though *a priori* we do not assume there is a group structure behind these connections.

By the sheaf axioms, there is only one flat connection on a point, which we will call the trivial connection. Now consider the space $\mathcal{O}(0,1)$ of flat connections on the standard 1-simplex $(0,1)$. If the sheaf was indeed the sheaf of flat connections on a group G, then $\mathcal{O}(0,1)$ is canonically identifiable with G. Inspired by this, we will *define* G to equal the space $\mathcal{O}(0,1)$ of flat connections on $(0,1)$. The flat connections on any other 1-simplex (u,v) can then be placed in one-to-one correspondence with elements of G by the morphism $u \mapsto 0, v \mapsto 1$, so flat connections on (u,v) can be viewed as being *equivalent* to an element of G.

At present, G is merely a set, not a group. To make it into a group, we need to introduce an identity element, an inverse operation, and a multiplication operation, and verify the group axioms.

To obtain an identity element, we look at the morphism from $(0,1)$ to a point, and pull back the trivial connection on that point to obtain a flat connection e on $(0,1)$, which we will declare to be the

identity element. (Note from the functorial nature of pullback that it does not matter which point we choose for this.)

Now we define the multiplication operation. Let $g, h \in G$, then g and h are flat connections on $(0,1)$. By using the morphism $i \mapsto i-1$ from $(1,2)$ to $(0,1)$, we can pull back h to $(1,2)$ to create a flat connection $\tilde{h}$ on $(1,2)$ that is equivalent to h. The restriction of g and $\tilde{h}$ to the point 1 is trivial, so by the gluing axiom we can glue g and $\tilde{h}$ to a flat connection on the complex $(0,1), (1,2)$. By Grothendieck's axiom, one can then uniquely extend this connection to the 2-simplex $(0,1,2)$, which can then be restricted to the edge $(0,2)$. Mapping this edge back to $(0,1)$, we obtain an element of G, which we will define to be hg.

This operation is closed. To verify the identity property, observe that if $g \in G$, then by starting with the simplex $(0,1,2)$ and pulling back g under the morphism that sends 2 to 1 but is the identity on $0, 1$, we obtain a flat connection on $(0,1,2)$ which is equal to g on $(0,1)$, equivalent to the identity on $(1,2)$, and is equivalent to g on $(0,2)$ (after identifying $(0,2)$ with $(0,1)$). From the definition of group multiplication, this shows that $eg = g$; a similar argument (using a slightly different morphism from $(0,1,2)$ to $(0,1)$) gives $ge = g$.

Now we establish associativity. Let $f, g, h \in G$. Using the definition of multiplication, we can create a flat connection on the 2-simplex $(0,1,2)$ which equals h on $(0,1)$, is equivalent to g on $(1,2)$, and is equivalent to gh on $(0,2)$. We then glue on the edge $(2,3)$ and extend the flat connection to be equivalent to f on $(2,3)$. Using Grothendieck's axiom and the definition of multiplication, we can then extend the flat connection to the 2-simplex $(0,2,3)$ to be equivalent to $f(gh)$ on $(0,3)$. By another use of that axiom, we can also extend the flat connection to the 2-simplex $(1,2,3)$, to be equivalent to fg on $(1,3)$. Now that we have three of the four faces of the 3-simplex $(0,1,2,3)$, we can now apply the $k=3$ case of Grothendieck's axiom and extend the flat connection to the entire 3-simplex, and in particular to the 2-simplex $(0,1,3)$. Using the definition of multiplication again, we thus see that $f(gh) = (fg)h$, giving associativity.

Next, we establish the inverse property. It will suffice to establish the existence of a left-inverse and a right-inverse for each group element, since the associativity property will then guarantee that these two inverses equal each other. We shall just establish the left-inverse property, as the right-inverse is analogous. Let $g \in G$ be arbitrary. By the gluing axiom, one can form a flat connection on the complex $(0,1), (0,2)$ which equals g on $(0,1)$ and is equivalent to the identity on $(0,2)$. By Grothendieck's axiom, this can be extended to a flat connection on $(0,1,2)$; the restriction of this connection to $(1,2)$ is equivalent to some element of G, which we define to be g^{-1}. By construction, $g^{-1}g = e$ as required.

We have just shown that G is a group. The last thing to do is to show that this abstract sheaf $\mathcal{O}$ can be indeed identified with the sheaf of G-flat connections. This is fairly straightforward: given an abstract flat connection on a complex, the restriction of that connection to any edge is equivalent to an element of G. To verify that this genuinely determines a G-connection on that complex, we need to verify that if (u,v) and (v,u) are both in the complex, that the group elements g, h assigned to these edges invert each other. But we can pullback $(u,v), (v,u)$ to the 2-simplex $(0,1,2)$ by mapping $0, 2$ to u and 1 to v, creating a flat connection that is equal to g on $(0,1)$, equivalent to h on $(1,2)$, and equivalent to the identity on $(0,2)$; by definition of multiplication or inverse we conclude that g, h invert each other as desired.

Thus the abstract connection defines a G-connection. From the definition of multiplication it is also clear that every 2-simplex in the complex imposes the right relation on the three elements of G associated to the edges of that simplex that makes the G-connection flat. Thus we have a canonical way to associate a G-flat connection to each abstract flat connection; the only remaining things to do are verify that this association is bijective.

We induct on the size of the complex. If the complex is not a single simplex, the claim follows from the induction hypothesis by viewing the complex as the union of two (possibly overlapping) smaller complexes, and using the gluing and locality axioms. So we may assume that the complex consists of a single simplex. If the simplex

is 0 or 1-dimensional the claim is easy; for $k \geq 2$ the claim follows from Grothendieck's axiom (which applies both for the abstract flat connections (by hypothesis) and for G-flat connections (as verified earlier)) and the induction hypothesis.

Notes. This article first appeared at `terrytao.wordpress.com/2009/10/19`. Thanks to Lior, Raj, and anonymous commenters for corrections.

Raj and Ben Wieland noted the close connection to the Kan extension property.

3.15. The "no self-defeating object" argument

A fundamental tool in any mathematician's toolkit is that of *reductio ad absurdum*: showing that a statement X is false by assuming first that X is true, and showing that this leads to a logical contradiction. A particulary pure example of *reductio ad absurdum* occurs when establishing the non-existence of a hypothetically overpowered object or structure X, by showing that X's powers are "self-defeating": the very existence of X and its powers can be used (by some clever trick) to construct a counterexample to that power. Perhaps the most well-known example of a self-defeating object comes from the *omnipotence paradox* in philosophy ("Can an omnipotent being create a rock so heavy that He cannot lift it?"); more generally, a large number of other paradoxes in logic or philosophy can be reinterpreted as a proof that a certain overpowered object or structure does not exist.

In mathematics, perhaps the first example of a self-defeating object one encounters is that of a largest natural number:

Proposition 3.15.1 (No largest natural number)**.** *There does not exist a natural number N which is larger than all other natural numbers.*

Proof. Suppose for contradiction that there was such a largest natural number N. Then $N+1$ is also a natural number which is strictly larger than N, contradicting the hypothesis that N is the largest natural number. □

Note the argument does not apply to the *extended natural number system* in which one adjoins an additional object ∞ beyond the natural numbers, because $\infty + 1$ is defined equal to ∞. However, the above argument does show that the existence of a largest number is not compatible with the *Peano axioms*.

This argument, by the way, is perhaps the only mathematical argument I know of which is routinely taught to primary school children *by other primary school children*, thanks to the schoolyard game of naming the largest number. It is arguably one's first exposure to a mathematical *non-existence result*, which seems innocuous at first but can be surprisingly deep, as such results preclude in advance all future attempts to establish existence of that object, no matter how much effort or ingenuity is poured into this task. One sees this in a typical instance of the above schoolyard game; one player tries furiously to cleverly construct some impressively huge number N, but no matter how much effort is expended in doing so, the player is defeated by the simple response "... plus one!" (unless, of course, N is infinite, ill-defined, or otherwise not a natural number).

It is not only individual objects (such as natural numbers) which can be self-defeating; structures (such as orderings or enumerations) can also be self-defeating. (In modern set theory, one considers structures to themselves be a kind of object, and so the distinction between the two concepts is often blurred.) Here is one example (related to, but subtly different from, the previous one):

Proposition 3.15.2 (The natural numbers cannot be finitely enumerated)**.** *The natural numbers* $\mathbf{N} = \{0, 1, 2, 3, \ldots\}$ *cannot be written as* $\{a_1, \ldots, a_n\}$ *for any finite collection* $a_1, \ldots, a_n$ *of natural numbers.*

Proof. Suppose for contradiction that such an enumeration $\mathbf{N} = \{a_1, \ldots, a_n\}$ existed. Then consider the number $a_1 + \ldots + a_n + 1$; this is a natural number, but is larger than (and hence not equal to) any of the natural numbers $a_1, \ldots, a_n$, contradicting the hypothesis that $\mathbf{N}$ is enumerated by $a_1, \ldots, a_n$. □

Here it is the *enumeration* which is self-defeating, rather than any individual natural number such as a_1 or a_n. (For this post, we allow enumerations to contain repetitions.)

The above argument may seem trivial, but a slight modification of it already gives an important result, namely *Euclid's theorem*:

Proposition 3.15.3 (The primes cannot be finitely enumerated)**.** *The prime numbers* $\mathcal{P} = \{2, 3, 5, 7, \ldots\}$ *cannot be written as* $\{p_1, \ldots, p_n\}$ *for any finite collection of prime numbers.*

Proof. Suppose for contradiction that such an enumeration $\mathcal{P} = \{p_1, \ldots, p_n\}$ existed. Then consider the natural number $p_1 \times \ldots \times p_n + 1$; this is a natural number larger than 1 which is not divisible by any of the primes $p_1, \ldots, p_n$. But, by the *fundamental theorem of arithmetic* (or by the method of *Infinite descent*, which is another classic application of *reductio ad absurdum*), every natural number larger than 1 must be divisible by some prime, contradicting the hypothesis that $\mathcal{P}$ is enumerated by $p_1, \ldots, p_n$. □

Remark 3.15.4. Continuing the number-theoretic theme, the "dueling conspiracies" arguments in Section 3.12.4 can also be viewed as an instance of this type of "no-self-defeating-object"; for instance, a zero of the Riemann zeta function at $1 + it$ implies that the primes anti-correlate almost completely with the multiplicative function n^{it}, which is self-defeating because it also implies complete anti-correlation with n^{-it}, and the two are incompatible. Thus we see that the *prime number theorem* (a much stronger version of Proposition 3.15.3) also emerges from this type of argument.

In this post I would like to collect several other well-known examples of this type of "no self-defeating object" argument. Each of these is well studied, and probably quite familiar to many of you, but I feel that by collecting them all in one place, the commonality of theme between these arguments becomes more apparent. (For instance, as we shall see, many well-known "paradoxes" in logic and philosophy can be interpreted mathematically as a rigorous "no self-defeating object" argument.)

3.15.1. Set theory. Many famous foundational results in set theory come from a "no self-defeating object" argument. (Here, we shall

be implicitly be using a standard axiomatic framework for set theory, such as *Zermelo-Frankel set theory*; the situation becomes different for other set theories, much as results such as Proposition 3.15.1 changes if one uses other number systems such as the extended natural numbers.) The basic idea here is that any sufficiently overpowered set-theoretic object is capable of encoding a version of the *liar paradox* ("this sentence is false", or more generally a statement which can be shown to be logically equivalent to its negation) and thus lead to a contradiction. Consider for instance this variant of *Russell's paradox*:

Proposition 3.15.5 (No universal set)**.** *There does not exist a set which contains all sets (including itself).*

Proof. Suppose for contradiction that there existed a universal set X which contained all sets. Using the *axiom schema of specification*, one can then construct the set

$$Y := \{A \in X : A \notin A\}$$

of all sets in the universe which did not contain themselves. As X is universal, Y is contained in X. But then, by definition of Y, one sees that $Y \in Y$ if and only if $Y \notin Y$, a contradiction. □

Remark 3.15.6. As a corollary, there also does not exist any set Z which contains all *other* sets (not including itself), because the set $X := Z \cup \{Z\}$ would then be universal.

One can "localise" the above argument to a smaller domain than the entire universe, leading to the important

Proposition 3.15.7 (Cantor's theorem)**.** *Let X be a set. Then the power set $2^X := \{A : A \subset X\}$ of X cannot be enumerated by X, i.e. one cannot write $2^X := \{A_x : x \in X\}$ for some collection $(A_x)_{x \in X}$ of subsets of X.*

Proof. Suppose for contradiction that there existed a set X whose power set 2^X could be enumerated as $\{A_x : x \in X\}$ for some $(A_x)_{x \in X}$. Using the axiom schema of specification, one can then construct the set

$$Y := \{x \in X : x \notin A_x\}.$$

The set Y is an element of the power set 2^X. As 2^X is enumerated by $\{A_x : x \in X\}$, we have $Y = A_y$ for some $y \in X$. But then by the definition of Y, one sees that $y \in A_y$ if and only if $y \notin A_y$, a contradiction. □

As is well-known, one can adapt Cantor's argument to the real line, showing that the reals are uncountable:

Proposition 3.15.8 (The real numbers cannot be countably enumerated)**.** *The real numbers* $\mathbf{R}$ *cannot be written as* $\{x_n : n \in \mathbf{N}\}$ *for any countable collection* $x_1, x_2, \ldots$ *of real numbers.*

Proof. Suppose for contradiction that there existed a countable enumeration of $\mathbf{R}$ by a sequence $x_1, x_2, \ldots$ of real numbers. Consider the decimal expansion of each of these numbers. Note that, due to the well-known "$0.999\ldots = 1.000\ldots$" issue, the decimal expansion may be non-unique, but each real number x_n has at most two decimal expansions. For each n, let $a_n \in \{0, 1, \ldots, 9\}$ be a digit which is not equal to the n^{th} digit of any of the decimal expansions of x_n; this is always possible because there are ten digits to choose from and at most two decimal expansions of x_n. (One can avoid any implicit invocation of the *axiom of choice* here by setting a_n to be (say) the *least* digit which is not equal to the n^{th} digit of any of the decimal expansions of x_n.) Then the real number given by the decimal expansion $0.a_1a_2a_3\ldots$ differs in the n^{th} digit from any of the decimal expansions of x_n for each n, and so is distinct from every x_n, a contradiction. □

Remark 3.15.9. One can of course deduce Proposition 3.15.8 directly from Proposition 3.15.7, by using the decimal representation to embed $2^{\mathbf{N}}$ into $\mathbf{R}$. But notice how the two arguments have a slightly different (though closely related) basis; the former argument proceeds by encoding the liar paradox, while the second proceeds by exploiting Cantor's diagonal argument. The two perspectives are indeed a little different: for instance, Cantor's diagonal argument can also be modified to establish the *Arzela-Ascolí theorem*, whereas I do not see any obvious way to prove that theorem by encoding the liar paradox.

Remark 3.15.10. It is an interesting psychological phenomenon that Proposition 3.15.8 is often considered far more unintuitive than any

of the other propositions here, despite being in the same family of arguments; most people have no objection to the fact that every effort to finitely enumerate the natural numbers, for instance, is doomed to failure, but for some reason it is much harder to let go of the belief that, at some point, some sufficiently ingenious person will work out a way to countably enumerate the real numbers. I am not exactly sure why this disparity exists, but I suspect it is at least partly due to the fact that the rigorous construction of the real numbers is quite sophisticated and often not presented properly until the advanced undergraduate level. (Or perhaps it is because we do not play the game "enumerate the real numbers" often enough in schoolyards.)

Remark 3.15.11. One can also use the diagonal argument to show that any reasonable notion of a "constructible real number" must itself be non-constructive (this is related to the *interesting number paradox*). Part of the problem is that the question of determining whether a proposed construction of a real number is actually well-defined is a variant of the *halting problem*, which we will discuss below.

While a genuinely universal set is not possible in standard set theory, one at least has the notion of an *ordinal*, which contains all the ordinals less than it. (In the discussion below, we assume familiarity with the theory of ordinals.) One can modify the above arguments concerning sets to give analogous results about the ordinals. For instance:

Proposition 3.15.12 (Ordinals do not form a set)**.** *There does not exist a set that contains all the ordinals.*

Proof. Suppose for contradiction that such a set existed. By the axiom schema of specification, one can then find a set Ω which consists precisely of all the ordinals and nothing else. But then $\Omega \cup \{\Omega\}$ is an ordinal which is not contained in Ω (by the *axiom of foundation*, also known as the *axiom of regularity*), a contradiction. □

Remark 3.15.13. This proposition(together with the theory of ordinals) can be used to give a quick proof of *Zorn's lemma*: see Section 2.4 for further discussion. Notice the similarity between this argument and the proof of Proposition 3.15.1.

Remark 3.15.14. Once one has Zorn's lemma, one can show that various other classes of mathematical objects do not form sets. Consider for instance the class of all vector spaces. Observe that any chain of (real) vector spaces (ordered by inclusion) has an upper bound (namely the union or limit of these spaces); thus, if the class of all vector spaces was a set, then Zorn's lemma would imply the existence of a maximal vector space V. But one can simply adjoin an additional element not already in V (e.g. $\{V\}$) to V, and contradict this maximality. (An alternate proof: every object is an element of some vector space, and in particular every set is an element of some vector space. If the class of all vector spaces formed a set, then by the *axiom of union*, we see that union of all vector spaces is a set also, contradicting Proposition 3.15.5.)

One can localise the above argument to smaller cardinalities, for instance:

Proposition 3.15.15 (Countable ordinals are uncountable)**.** *There does not exist a countable enumeration $\omega_1, \omega_2, \ldots$ of the countable ordinals. (Here we consider finite sets and countably infinite sets to both be countable.)*

Proof. Suppose for contradiction that there exists a countable enumeration $\omega_1, \omega_2, \ldots$ of the countable ordinals. Then the set $\Omega := \bigcup_n \omega_n$ is also a countable ordinal, as is the set $\Omega \cup \{\Omega\}$. But $\Omega \cup \{\Omega\}$ is not equal to any of the ω_n (by the axiom of foundation), a contradiction. □

Remark 3.15.16. One can show the existence of uncountable ordinals (e.g. by considering all the well-orderings of subsets of the natural numbers, up to isomorphism), and then there exists a least uncountable ordinal Ω. By construction, this ordinal consists precisely of all the countable ordinals, but is itself uncountable, much as $\mathbf{N}$ consists precisely of all the finite natural numbers, but is itself infinite (Proposition 3.15.2). The least uncountable ordinal is notorious, among other things, for providing a host of counterexamples to various intuitively plausible assertions in point set topology, and in particular in showing that the topology of sufficiently uncountable

spaces cannot always be adequately explored by countable objects such as sequences.

Remark 3.15.17. The existence of the least uncountable ordinal can explain why one cannot contradict Cantor's theorem on the uncountability of the reals simply by iterating the diagonal argument (or any other algorithm) in an attempt to "exhaust" the reals. From *transfinite induction* we see that the diagonal argument allows one to assign a different real number to each countable ordinal, but this does not establish countability of the reals, because the set of all countable ordinals is itself uncountable. (This is similar to how one cannot contradict Proposition 3.15.5 by iterating the $N \to N+1$ map, as the set of all finite natural numbers is itself infinite.) In any event, even once one reaches the first uncountable ordinal, one may not yet have completely exhausted the reals; for instance, using the diagonal argument given in the proof of Proposition 3.15.8, only the real numbers in the interval $[0,1]$ will ever be enumerated by this procedure. (Also, the question of whether *all* real numbers in $[0,1]$ can be enumerated by the iterated diagonal algorithm requires the *continuum hypothesis*, and even with this hypothesis I am not sure whether the statement is decidable.)

3.15.2. Logic. The "no self-defeating object" argument leads to a number of important non-existence results in logic. Again, the basic idea is to show that a sufficiently overpowered logical structure will eventually lead to the existence of a self-contradictory statement, such as the liar paradox. To state examples of this properly, one unfortunately has to invest a fair amount of time in first carefully setting up the language and theory of logic. I will not do so here, and instead use informal English sentences as a proxy for precise logical statements to convey a taste (but not a completely rigorous description) of these logical results here.

The liar paradox itself - the inability to assign a consistent truth value to "this sentence is false" - can be viewed as an argument demonstrating that there is no consistent way to interpret (i.e. assign a truth value to) sentences, when the sentences are (a) allowed to be self-referential, and (b) allowed to invoke the very notion of truth given by this interpretation. One's first impulse is to say that the

difficulty here lies more with (a) than with (b), but there is a clever trick, known as *Quining* (or *indirect self-reference*), which allows one to modify the liar paradox to produce a non-self-referential statement to which one still cannot assign a consistent truth value. The idea is to work not with fully formed sentences S, which have a single truth value, but instead with *predicates* S, whose truth value depends on a variable x in some range. For instance, S may be "x is thirty-two characters long.", and the range of x may be the set of strings (i.e. finite sequences of characters); then for every string T, the statement $S(T)$ (formed by replacing every appearance of x in S with T) is either true or false. For instance, $S(\text{“}a\text{''})$ is true, but $S(\text{“}ab\text{''})$ is false. Crucially, predicates are themselves strings, and can thus be fed into themselves as input; for instance, $S(S)$ is false. If however U is the predicate "x is sixty-five characters long.", observe that $U(U)$ is true.

Now consider the *Quine predicate* Q given by

"x is a predicate whose range is the set of strings, and $x(x)$ is false."

whose range is the set of strings. Thus, for any string T, $Q(T)$ is the sentence

"T is a predicate whose range is the set of strings, and $T(T)$ is false."

This predicate is defined non-recursively, but the sentence $Q(Q)$ captures the essence of the liar paradox: it is true if and only if it is false. This shows that there is no consistent way to interpret sentences in which the sentences are allowed to come from predicates, are allowed to use the concept of a string, and also allowed to use the concept of truth as given by that interpretation.

Note that the proof of Proposition 3.15.5 is basically the set-theoretic analogue of the above argument, with the connection being that one can identify a predicate $T(x)$ with the set $\{x : T(x)\,\text{true}\}$.

By making one small modification to the above argument - replacing the notion of truth with the related notion of provability - one obtains the celebrated *Gödel's (second) incompleteness theorem*:

Theorem 3.15.18 (Gödel's incompleteness theorem)**.** *(Informal statement) No consistent logical system which has the notion of a string, can provide a proof of its own logical consistency. (Note that a proof can be viewed as a certain type of string.)*

Remark 3.15.19. Because one can encode strings in numerical form (e.g. using the *ASCII code*), it is also true (informally speaking) that no consistent logical system which has the notion of a natural number, can provide a proof of its own logical consistency.

Proof. (Informal sketch only) Suppose for contradiction that one had a consistent logical system inside of which its consistency could be proven. Now let Q be the predicate given by

"x is a predicate whose range is the set of strings, and $x(x)$ is not provable"

and whose range is the set of strings. Define the *Gödel sentence* G to be the string $G := Q(Q)$. Then G is logically equivalent to the assertion "G is not provable". Thus, if G were false, then G would be provable, which (by the consistency of the system) implies that G is true, a contradiction; thus, G must be true. Because the system is provably consistent, the above argument can be placed inside the system itself, to *prove* inside that system that G must be true; thus G is provable and G is then false, a contradiction. (It becomes quite necessary to carefully distinguish the notions of truth and provability (both inside a system and externally to that system) in order to get this argument straight!) □

Remark 3.15.20. It is not hard to show that a consistent logical system which can model the standard natural numbers cannot *disprove* its own consistency either (i.e. it cannot establish the statement that one can deduce a contradiction from the axioms in the systems in n steps for some natural number n); thus the consistency of such a system is undecidable within that system. Thus this theorem strengthens the (more well known) first Gödel incompleteness theory, which asserts the existence of undecidable statements inside a consistent logical system which contains the concept of a string (or a natural number). On the other hand, the incompleteness theorem does not preclude the possibility that the consistency of a weak theory could be proven in a strictly stronger theory (e.g. the consistency of Peano arithmetic is provable in Zermelo-Frankel set theory).

Remark 3.15.21. One can use the incompleteness theorem to establish the undecidability of other overpowered problems. For instance,

Matiyasevich's theorem demonstrates that the problem of determining the solvability of a system of Diophantine equations is, in general, undecidable, because one can encode statements such as the consistency of set theory inside such a system.

3.15.3. Computability. One can adapt these arguments in logic to analogous arguments in the theory of computation; the basic idea here is to show that a sufficiently overpowered computer program cannot exist, by feeding the source code for that program into the program itself (or some slight modification thereof) to create a contradiction. As with logic, a properly rigorous formalisation of the theory of computation would require a fair amount of preliminary machinery, for instance to define the concept of Turing machine (or some other universal computer), and so I will once again use informal English sentences as an informal substitute for a precise programming language.

A fundamental "no self-defeating object" argument in the subject, analogous to the other liar paradox type arguments encountered previously, is the *Turing halting theorem*:

Theorem 3.15.22 (Turing halting theorem)**.** *There does not exist a program P which takes a string S as input, and determines in finite time whether S is a program (with no input) that halts in finite time.*

Proof. Suppose for contradiction that such a program P existed. Then one could easily modify P to create a variant program Q, which takes a string S as input, and halts if and only if S is a program (with S itself as input) that does not halts in finite time. Indeed, all Q has to do is call P with the string $S(S)$, defined as the program (with no input) formed by declaring S to be the input string for the program S. If P determines that $S(S)$ does not halt, then Q halts; otherwise, if P determines that $S(S)$ does halt, then Q performs an infinite loop and does not halt. Then observe that $Q(Q)$ will halt if and only if it does not halt, a contradiction. □

Remark 3.15.23. As one can imagine from the proofs, this result is closely related to, but not quite identical with, the Gödel incompleteness theorem. That latter theorem implies that the question of

whether a given program halts or not in general is undecidable (consider a program designed to search for proofs of the inconsistency of set theory). By contrast, the halting theorem (roughly speaking) shows that this question is *uncomputable* (i.e. there is no algorithm to decide halting in general) rather than *undecidable* (i.e. there are programs whose halting can neither be proven nor disproven).

On the other hand, the halting theorem can be used to establish the incompleteness theorem. Indeed, suppose that all statements in a suitably strong and consistent logical system were either provable or disprovable. Then one could build a program that determined whether an input string S, when run as a program, halts in finite time, simply by searching for all proofs or disproofs of the statement "S halts in finite time"; this program is guaranteed to terminate with a correct answer by hypothesis.

Remark 3.15.24. While it is not possible for the halting problem for a given computing language to be computable in that language, it is certainly possible that it is computable in a strictly stronger language. When that is the case, one can then invoke *Newcomb's paradox* to argue that the weaker language does not have unlimited "free will" in some sense.

Remark 3.15.25. In the language of *recursion theory*, the halting theorem asserts that the set of programs that do not halt is not a *decidable set* (or a *recursive set*). In fact, one can make the slightly stronger assertion that the set of programs that do not halt is not even a *semi-decidable set* (or a *recursively enumerable set*), i.e. there is no algorithm which takes a program as input and halts in finite time if and only if the input program does not halt. This is because the complementary set of programs that do halt is clearly semi-decidable (one simply runs the program until it halts, running forever if it does not), and so if the set of programs that do not halt is also semi-decidable, then it is decidable (by running both algorithms in parallel; this observation is a special case of *Post's theorem*).

Remark 3.15.26. One can use the halting theorem to exclude overly general theories for certain types of mathematical objects. For instance, one cannot hope to find an algorithm to determine the existence of smooth solutions to arbitrary nonlinear partial differential

equations, because it is possible to simulate a Turing machine using the laws of classical physics, which in turn can be modeled using (a moderately complicated system of) nonlinear PDE. Instead, progress in nonlinear PDE has instead proceeded by focusing on much more specific classes of such PDE (e.g. elliptic PDE, parabolic PDE, hyperbolic PDE, gauge theories, etc.).

One can place the halting theorem in a more "quantitative" form. Call a function $f : \mathbf{N} \to \mathbf{N}$ *computable* if there exists a computer program which, when given a natural number n as input, returns $f(n)$ as output in finite time. Define the *Busy Beaver function* $BB : \mathbf{N} \to \mathbf{N}$ by setting $BB(n)$ to equal the largest output of any program of at most n characters in length (and taking no input), which halts in finite time. Note that there are only finitely many such programs for any given n, so $BB(n)$ is well-defined. On the other hand, it is uncomputable, even to upper bound:

Proposition 3.15.27. *There does not exist a computable function f such that one has $BB(n) \leq f(n)$ for all n.*

Proof. Suppose for contradiction that there existed a computable function $f(n)$ such that $BB(n) \leq f(n)$ for all n. We can use this to contradict the halting theorem, as follows. First observe that once the Busy Beaver function can be upper bounded by a computable function, then for any n, the run time of any halting program of length at most n can also be upper bounded by a computable function. This is because if a program of length n halts in finite time, then a trivial modification of that program (of length larger than n, but by a computable factor) is capable of outputting the run time of that program (by keeping track of a suitable "clock" variable, for instance). Applying the upper bound for Busy Beaver to that increased length, one obtains the bound on run time.

Now, to determine whether a given program S halts in finite time or not, one simply simulates (runs) that program for time up to the computable upper bound of the possible running time of S, given by the length of S. If the program has not halted by then, then it never will. This provides a program P obeying the hypotheses of the halting theorem, a contradiction. □

Remark 3.15.28. A variant of the argument shows that $BB(n)$ grows faster than any computable function: thus if f is computable, then $BB(n) > f(n)$ for all sufficiently large n. We leave the proof of this result as an exercise to the reader.

Remark 3.15.29. Sadly, the most important unsolved problem in complexity theory, namely the $P \neq NP$ *problem*, does not seem to be susceptible to the no-self-defeating-object argument, basically because such arguments tend to be *relativisable* whereas the $P \neq NP$ problem is not; see Section 3.9 for more discussion. On the other hand, one has the curious feature that many proposed *proofs* that $P \neq NP$ appear to be self-defeating; this is most strikingly captured in the celebrated work of Razborov and Rudich[**RaRu1997**], who showed (very roughly speaking) that any sufficiently "natural" proof that $P \neq NP$ could be used to disprove the existence of an object closely related to the belief that $P \neq NP$, namely the existence of pseudorandom number generators. (I am told, though, that diagonalisation arguments can be used to prove some other inclusions or non-inclusions in complexity theory that are not subject to the relativisation barrier, though I do not know the details.)

3.15.4. Game theory. Another basic example of the no-self-defeating-objects argument arises from game theory, namely the *strategy stealing argument*. Consider for instance a generalised version of naughts and crosses (tic-tac-toe), in which two players take turns placing naughts and crosses on some game board (not necessarily square, and not necessarily two-dimensional), with the naughts player going first, until a certain pattern of all naughts or all crosses is obtained as a subpattern, with the naughts player winning if the pattern is all naughts, and the crosses player winning if the pattern is all crosses. (If all positions are filled without either pattern occurring, the game is a draw.) We assume that the winning patterns for the cross player are exactly the same as the winning patterns for the naughts player (but with naughts replaced by crosses, of course).

Proposition 3.15.30. *In any generalised version of naughts and crosses, there is no strategy for the second player (i.e. the crosses player) which is guaranteed to ensure victory.*

Proof. Suppose for contradiction that the second player had such a winning strategy W. The first player can then *steal* that strategy by placing a naught arbitrarily on the board, and then pretending to be the second player and using W accordingly. Note that occasionally, the W strategy will compel the naughts player to place a naught on the square that he or she has already occupied, but in such cases the naughts player may simply place the naught somewhere else instead. (It is not possible that the naughts player would run out of places, thus forcing a draw, because this would imply that W could lead to a draw as well, a contradiction.) If we denote this stolen strategy by W', then W' guarantees a win for the naughts player; playing the W' strategy for the naughts player against the W strategy for the crosses player, we obtain a contradiction. □

Remark 3.15.31. The key point here is that in naughts and crosses games, it is possible to play a *harmless move* - a move which gives up the turn of play, but does not actually decrease one's chance of winning. In games such as chess, there does not appear to be any analogue of the harmless move, and so it is not known whether black actually has a strategy guaranteed to win or not in chess, though it is suspected that this is not the case.

Remark 3.15.32. The *Hales-Jewett theorem* shows that for any fixed board length, an n-dimensional game of naughts and crosses is unable to end in a draw if n is sufficiently large. An induction argument shows that for any two-player game that terminates in bounded time in which draws are impossible, one player must have a guaranteed winning strategy; by the above proposition, this strategy must be a win for the naughts player. Note, however, that Proposition 3.15.30 provides no information as to *how* to locate this winning strategy, other than that this strategy belongs to the naughts player. Nevertheless, this gives a second example in which the no-self-defeating-object argument can be used to ensure the *existence* of some object, rather than the *non-existence* of an object. (The first example was the prime number theorem, discussed earlier.)

The strategy-stealing argument can be applied to real-world economics and finance, though as with any other application of mathematics to the real world, one has to be careful as to the implicit assumptions one is making about reality and how it conforms to one's mathematical model when doing so. For instance, one can argue that in any market or other economic system in which the net amount of money is approximately constant, it is not possible to locate a universal trading strategy which is guaranteed to make money for the user of that strategy, since if everyone applied that strategy then the net amount of money in the system would increase, a contradiction. Note however that there are many loopholes here; it may be that the strategy is difficult to copy, or relies on exploiting some other group of participants who are unaware or unable to use the strategy, and would then lose money (though in such a case, the strategy is not truly universal as it would stop working once enough people used it). Unfortunately, there can be strong psychological factors that can cause people to override the conclusions of such strategy-stealing arguments with their own rationalisations, as can be seen, for instance, in the perennial popularity of pyramid schemes, or to a lesser extent, market bubbles (though one has to be careful about applying the strategy-stealing argument in the latter case, since it is possible to have net wealth creation through external factors such as advances in technology).

Note also that the strategy-stealing argument also limits the universal predictive power of *technical analysis* to provide predictions other than that the prices obey a *martingale*, though again there are loopholes in the case of markets that are illiquid or highly volatile.

3.15.5. Physics. In a similar vein, one can try to adapt the no-self-defeating-object argument from mathematics to physics, but again one has to be much more careful with various physical and metaphysical assumptions that may be implicit in one's argument. For instance, one can argue that under the laws of special relativity, it is not possible to construct a totally immovable object. The argument would be that if one could construct an immovable object O in one inertial reference frame, then by the *principle of relativity* it should be possible to construct an object O' which is immovable in another

inertial reference frame which is moving with respect to the first; setting the two on a collision course, we obtain the classic contradiction between an irresistible force and an immovable object. Note however that there are several loopholes here which allow one to avoid contradiction; for instance, the two objects O, O' could simply pass through each other without interacting.

In a somewhat similar vein, using the laws of special relativity one can argue that it is not possible to systematically generate and detect *tachyon particles* - particles traveling faster than the speed of light - because these could be used to transmit localised information faster than the speed of light, and then (by the principle of relativity) to send localised information back into the past, from one location to a distant one. Setting up a second tachyon beam to reflect this information back to the original location, one could then send localised information back to one's own past (rather than to the past of an observer at a distant location), allowing one to set up a classic *grandfather paradox*. However, as before, there are a large number of loopholes in this argument which could let one avoid contradiction; for instance, if the apparatus needed to set up the tachyon beam may be larger than the distance the beam travels (as is for instance the case in *Mexican wave*-type tachyon beams) then there is no causality paradox; another loophole is if the tachyon beam is not fully localised, but propagates in spacetime in a manner to interfere with the second tachyon beam. A third loophole occurs if the universe exhibits quantum behaviour (in particular, the ability to exist in entangled states) instead of non-quantum behaviour, which allows for such superluminal mechanisms as wave function collapse to occur without any threat to causality or the principle of relativity. A fourth loophole occurs if the effects of relativistic gravity (i.e. general relativity) become significant. Nevertheless, the paradoxical effect of time travel is so strong that this physical argument is a fairly convincing way to rule out many commonly imagined types of faster-than-light travel or communication (and we have a number of other arguments too that exclude more modes of faster-than-light behaviour, though this is an entire blog post topic in its own right).

Notes. This article first appeared at `terrytao.wordpress.com/2009/10/27`. Thanks to Seva Lev and an anonymous commenter for corrections.

3.16. From Bose-Einstein condensates to the nonlinear Schrödinger equation

The *Schrödinger equation*

$$i\hbar\partial_t|\psi\rangle = H|\psi\rangle$$

is the fundamental equation of motion for (non-relativistic) quantum mechanics, modeling both one-particle systems and N-particle systems for $N > 1$. Remarkably, despite being a *linear* equation, solutions $|\psi\rangle$ to this equation can be governed by a *non-linear* equation in the large particle limit $N \to \infty$. In particular, when modeling a *Bose-Einstein condensate* with a suitably scaled interaction potential V in the large particle limit, the solution can be governed by the *cubic nonlinear Schrödinger equation*

$$i\partial_t\phi = \Delta\phi + \lambda|\phi|^2\phi. \tag{3.82}$$

I recently attended a talk by Natasa Pavlovic on the rigorous derivation of this type of limiting behaviour, which was initiated by the pioneering work of Hepp and Spohn, and has now attracted a vast recent literature. The rigorous details here are rather sophisticated; but the heuristic explanation of the phenomenon is fairly simple, and actually rather pretty in my opinion, involving the foundational quantum mechanics of N-particle systems. I am recording this heuristic derivation here, partly for my own benefit, but perhaps it will be of interest to some readers.

This discussion will be purely formal, in the sense that (important) analytic issues such as differentiability, existence and uniqueness, etc. will be largely ignored.

3.16.1. A quick review of classical mechanics. The phenomena discussed here are purely quantum mechanical in nature, but to motivate the quantum mechanical discussion, it is helpful to first quickly review the more familiar (and more conceptually intuitive) classical situation.

Classical mechanics can be formulated in a number of essentially equivalent ways: *Newtonian*, *Hamiltonian*, and *Lagrangian*. The formalism of Hamiltonian mechanics for a given physical system can be summarised briefly as follows:

- The physical system has a *phase space* Ω of states $\vec{x}$ (which is often parameterised by position variables q and momentum variables p). Mathematically, it has the structure of a *symplectic manifold*, with some *symplectic form* ω (which would be $\omega = dp \wedge dq$ if one had position and momentum coordinates available).
- The complete state of the system at any given time t is given (in the case of *pure states*) by a point $\vec{x}(t)$ in the phase space Ω.
- Every physical observable (e.g., energy, momentum, position, etc.) A is associated to a function (also called A) mapping the phase space Ω to the range of the observable (e.g. for real observables, A would be a function from Ω to $\mathbf{R}$). If one measures the observable A at time t, one will obtain the measurement $A(x(t))$.
- There is a special observable, the *Hamiltonian* $H : \Omega \to \mathbf{R}$, which governs the evolution of the state $\vec{x}(t)$ through time, via *Hamilton's equations of motion*. If one has position and momentum coordinates $\vec{x}(t) = (q_i(t), p_i(t))_{i=1}^n$, these equations are given by the formulae

$$\partial_t p_i = -\frac{\partial H}{\partial q_i}; \partial_t q_i = \frac{\partial H}{\partial p_i};$$

more abstractly, just from the symplectic form ω on the phase space, the equations of motion can be written as

$$\partial_t \vec{x}(t) = -\nabla_\omega H(\vec{x}(t)), \tag{3.83}$$

where $\nabla_\omega H$ is the symplectic gradient of H.

Hamilton's equation of motion can also be expressed in a dual form in terms of observables A, as *Poisson's equation of motion*

$$\partial_t A(\vec{x}(t)) = -\{H, A\}(\vec{x}(t))$$

for any observable A, where $\{H, A\} := \nabla_\omega H \cdot \nabla A$ is the *Poisson bracket*. One can express Poisson's equation more abstractly as

$$\partial_t A = -\{H, A\}. \tag{3.84}$$

In the above formalism, we are assuming that the system is in a *pure state* at each time t, which means that it only occupies a single point $\vec{x}(t)$ in phase space. One can also consider *mixed states* in which the state of the system at a time t is not fully known, but is instead given by a *probability distribution* $\rho(t, \vec{x})\ dx$ on phase space. The act of measuring an observable A at a time t will thus no longer be deterministic, but will itself be a random variable, whose expectation $\langle A \rangle$ is given by

$$\langle A \rangle(t) = \int_\Omega A(\vec{x}) \rho(t, \vec{x})\ d\vec{x}. \tag{3.85}$$

The equation of motion of a mixed state ρ is given by the *advection equation*

$$\partial_t \rho = \mathrm{div}(\rho \nabla_\omega H)$$

using the same vector field $-\nabla_\omega H$ that appears in (3.83); this equation can also be derived from (3.84), (3.85), and a duality argument.

Pure states can be viewed as the special case of mixed states in which the probability distribution $\rho(t, \vec{x})\ d\vec{x}$ is a *Dirac mass*[11] $\delta_{\vec{x}(t)}(\vec{x})$. One can thus think of mixed states as continuous averages of pure states, or equivalently the space of mixed states is the convex hull of the space of pure states.

Suppose one had a 2-particle system, in which the joint phase space $\Omega = \Omega_1 \times \Omega_2$ is the product of the two one-particle phase spaces. A pure joint state is then a point $x = (\vec{x}_1, \vec{x}_2)$ in Ω, where $\vec{x}_1$ represents the state of the first particle, and $\vec{x}_2$ is the state of the second particle. If the joint Hamiltonian $H : \Omega \to \mathbf{R}$ split as

$$H(\vec{x}_1, \vec{x}_2) = H_1(\vec{x}_1) + H_2(\vec{x}_2)$$

[11] We ignore for now the formal issues of how to perform operations such as derivatives on Dirac masses; this can be accomplished using the theory of distributions in Section 1.13 (or, equivalently, by working in the dual setting of observables) but this is not our concern here.

then the equations of motion for the first and second particles would be completely *decoupled*, with no interactions between the two particles. However, in practice, the joint Hamiltonian contains coupling terms between $\vec{x}_1, \vec{x}_2$ that prevents one from totally decoupling the system; for instance, one may have

$$H(\vec{x}_1, \vec{x}_2) = \frac{|p_1|^2}{2m_1} + \frac{|p_2|^2}{2m_2} + V(q_1 - q_2),$$

where $\vec{x}_1 = (q_1, p_1)$, $\vec{x}_2 = (q_2, p_2)$ are written using position coordinates q_i and momentum coordinates p_i, $m_1, m_2 > 0$ are constants (representing mass), and $V(q_1 - q_2)$ is some *interaction potential* that depends on the spatial separation $q_1 - q_2$ between the two particles.

In a similar spirit, a mixed joint state is a joint probability distribution $\rho(\vec{x}_1, \vec{x}_2)\, d\vec{x}_1 d\vec{x}_2$ on the product state space. To recover the (mixed) state of an individual particle, one must consider a *marginal distribution* such as

$$\rho_1(\vec{x}_1) := \int_{\Omega_2} \rho(\vec{x}_1, \vec{x}_2)\, d\vec{x}_2$$

(for the first particle) or

$$\rho_2(\vec{x}_2) := \int_{\Omega_1} \rho(\vec{x}_1, \vec{x}_2)\, d\vec{x}_1$$

(for the second particle). Similarly for N-particle systems: if the joint distribution of N distinct particles is given by $\rho(\vec{x}_1, \ldots, \vec{x}_N)\, \vec{dx}_1 \ldots \vec{dx}_N$, then the distribution of the first particle (say) is given by

$$\rho_1(\vec{x}_1) = \int_{\Omega_2 \times \ldots \times \Omega_N} \rho(\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_N)\, d\vec{x}_2 \ldots d\vec{x}_N,$$

the distribution of the first two particles is given by

$$\rho_{12}(\vec{x}_1, \vec{x}_2) = \int_{\Omega_3 \times \ldots \times \Omega_N} \rho(\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_N)\, d\vec{x}_3 \ldots d\vec{x}_N,$$

and so forth.

A typical Hamiltonian in this case may take the form

$$H(\vec{x}_1, \ldots, \vec{x}_n) = \sum_{j=1}^{N} \frac{|p_j|^2}{2m_j} + \sum_{1 \leq j < k \leq N} V_{jk}(q_j - q_k)$$

which is a combination of single-particle Hamiltonians H_j and interaction perturbations. If the momenta p_j and masses m_j are normalised to be of size $O(1)$, and the potential V_{jk} has an average value (i.e. an L^1 norm) of $O(1)$ also, then the former sum has size $O(N)$ and the latter sum has size $O(N^2)$, so the latter will dominate. In order to balance the two components and get a more interesting limiting dynamics when $N \to \infty$, we shall therefore insert a normalising factor of $\frac{1}{N}$ on the right-hand side, giving a Hamiltonian

$$H(\vec{x}_1, \ldots, \vec{x}_n) = \sum_{j=1}^{N} \frac{|p_j|^2}{2m_j} + \frac{1}{N} \sum_{1 \leq j < k \leq N} V_{jk}(q_j - q_k).$$

Now imagine a system of N *indistinguishable* particles. By this, we mean that all the state spaces $\Omega_1 = \ldots = \Omega_N$ are identical, and all observables (including the Hamiltonian) are symmetric functions of the product space $\Omega = \Omega_1^N$ (i.e. invariant under the action of the symmetric group S_N). In such a case, one may as well average over this group (since this does not affect any physical observable), and assume that all mixed states ρ are also symmetric. (One cost of doing this, though, is one has to largely give up pure states $(\vec{x}_1, \ldots, \vec{x}_N)$, since such states will not be symmetric except in the very exceptional case $\vec{x}_1 = \ldots = \vec{x}_N$.)

A typical example of a symmetric Hamiltonian is

$$H(\vec{x}_1, \ldots, \vec{x}_n) = \sum_{j=1}^{N} \frac{|p_j|^2}{2m} + \frac{1}{N} \sum_{1 \leq j < k \leq N} V(q_j - q_k)$$

where V is even (thus all particles have the same individual Hamiltonian, and interact with the other particles using the same interaction potential). In many physical systems, it is natural to consider only *short-range* interaction potentials, in which the interaction between q_j and q_k is localised to the region $q_j - q_k = O(r)$ for some small r. We model this by considering Hamiltonians of the form

$$H(\vec{x}_1, \ldots, \vec{x}_n) = \sum_{j=1}^{N} H(\vec{x}_j) + \frac{1}{N} \sum_{1 \leq j < k \leq N} \frac{1}{r^d} V(\frac{\vec{x}_j - \vec{x}_k}{r})$$

where d is the ambient dimension of each particle (thus in physical models, d would usually be 3); the factor of $\frac{1}{r^d}$ is a normalisation

factor designed to keep the L^1 norm of the interaction potential of size $O(1)$. It turns out that an interesting limit occurs when r goes to zero as N goes to infinity by some power law $r = N^{-\beta}$; imagine for instance N particles of "radius" r bouncing around in a box, which is a basic model for classical gases.

An important example of a symmetric mixed state is a *factored* state

$$\rho(\vec{x}_1, \ldots, \vec{x}_N) = \rho_1(\vec{x}_1) \ldots \rho_1(\vec{x}_N)$$

where ρ_1 is a single-particle probability density function; thus ρ is the tensor product of N copies of ρ_1. If there are no interaction terms in the Hamiltonian, then Hamiltonian's equation of motion will preserve the property of being a factored state (with ρ_1 evolving according to the one-particle equation); but with interactions, the factored nature may be lost over time.

3.16.2. A quick review of quantum mechanics. Now we turn to quantum mechanics. This theory is fundamentally rather different in nature than classical mechanics (in the sense that the basic objects, such as states and observables, are a different type of mathematical object than in the classical case), but shares many features in common also, particularly those relating to the Hamiltonian and other observables. (This relationship is made more precise via the *correspondence principle*, and more precise still using *semi-classical analysis*.)

The formalism of quantum mechanics for a given physical system can be summarised briefly as follows:

- The physical system has a *phase space* $\mathbf{H}$ of states $|\psi\rangle$ (which is often parameterised as a complex-valued function of the position space). Mathematically, it has the structure of a complex *Hilbert space*, which is traditionally manipulated using *bra-ket notation*.
- The complete *state* of the system at any given time t is given (in the case of *pure states*) by a unit vector $|\psi(t)\rangle$ in the phase space $\mathbf{H}$.

- Every physical observable A is associated to a linear operator on $\mathbf{H}$; real-valued observables are associated to self-adjoint linear operators. If one measures the observable A at time t, one will obtain the random variable whose expectation $\langle A \rangle$ is given by $\langle \psi(t)|A|\psi(t)\rangle$. (The full distribution of A is given by the *spectral measure* of A relative to $|\psi(t)\rangle$.)
- There is a special observable, the *Hamiltonian* $H : \mathbf{H} \to \mathbf{H}$, which governs the evolution of the state $|\psi(t)\rangle$ through time, via *Schrödinger's equations of motion*

$$i\hbar \partial_t |\psi(t)\rangle = H|\psi(t)\rangle. \tag{3.86}$$

Schrödinger's equation of motion can also be expressed in a dual form in terms of observables A, as *Heisenberg's equation of motion*

$$\partial_t \langle \psi|A|\psi\rangle = \frac{i}{\hbar}\langle \psi|[H, A]|\psi\rangle$$

or more abstractly as

$$\partial_t A = \frac{i}{\hbar}[H, A] \tag{3.87}$$

where $[,]$ is the *commutator* or *Lie bracket* (compare with (3.84)).

The states $|\psi\rangle$ are pure states, analogous to the pure states x in Hamiltonian mechanics. One also has *mixed states* ρ in quantum mechanics. Whereas in classical mechanics, a mixed state ρ is a probability distribution (a non-negative function of total mass $\int_\Omega \rho = 1$), in quantum mechanics a mixed state is a non-negative (i.e. *positive semi-definite*) *operator* ρ on $\mathbf{H}$ of total *trace* $\operatorname{tr} \rho = 1$. If one measures an observable A at a mixed state ρ, one obtains a random variable with expectation $\operatorname{tr} A\rho$. From (3.87) and duality, one can infer that the correct equation of motion for mixed states must be given by

$$\partial_t \rho = \frac{i}{\hbar}[H, \rho]. \tag{3.88}$$

One can view pure states as the special case of mixed states which are rank one projections,

$$\rho = |\psi\rangle\langle\psi|.$$

Morally speaking, the space of mixed states is the convex hull of the space of pure states (just as in the classical case), though things are a

little trickier than this when the phase space $\mathbf{H}$ is infinite dimensional, due to the presence of continuous spectrum in the *spectral theorem*.

Pure states suffer from a *phase ambiguity*: a phase rotation $e^{i\theta}|\psi\rangle$ of a pure state $|\psi\rangle$ leads to the same mixed state, and the two states cannot be distinguished by any physical observable.

In a single particle system, modeling a (scalar) quantum particle in a d-dimensional position space $\mathbf{R}^d$, one can identify the Hilbert space $\mathbf{H}$ with $L^2(\mathbf{R}^d \to \mathbf{C})$, and describe the pure state $|\psi\rangle$ as a *wave function* $\psi : \mathbf{R}^d \to \mathbf{C}$, which is normalised as

$$\int_{\mathbf{R}^d} |\psi(x)|^2 \, dx = 1$$

as $|\psi\rangle$ has to be a unit vector. (If the quantum particle has additional features such as *spin*, then one needs a fancier wave function, but let's ignore this for now.) A mixed state is then a function $\rho : \mathbf{R}^d \times \mathbf{R}^d \to \mathbf{C}$ which is Hermitian (i.e. $\rho(x,x') = \overline{\rho(x',x)}$) and positive definite, with unit trace $\int_{\mathbf{R}^d} \rho(x,x)\, dx = 1$; a pure state ψ corresponds to the mixed state $\rho(x,x') = \psi(x)\overline{\psi(x')}$.

A typical Hamiltonian in this setting is given by the operator

$$H\psi(x) := \frac{|p|^2}{2m}\psi(x) + V(x)\psi(x)$$

where $m > 0$ is a constant, p is the *momentum operator* $p := -i\hbar\nabla_x$, and ∇_x is the gradient in the x variable (so $|p|^2 = -\hbar^2\Delta_x$, where Δ_x is the Laplacian; note that ∇_x is skew-adjoint and should thus be thought of as being imaginary rather than real), and $V : \mathbf{R}^d \to \mathbf{R}$ is some potential. Physically, this depicts a particle of mass m in a potential well given by the potential V.

Now suppose one has an N-particle system of scalar particles. A pure state of such a system can then be given by an N-particle wave function $\psi : (\mathbf{R}^d)^N \to \mathbf{C}$, normalised so that

$$\int_{(\mathbf{R}^d)^N} |\psi(x_1, \ldots, x_N)|^2 \, dx_1 \ldots dx_N = 1$$

and a mixed state is a Hermitian positive semi-definite function $\rho : (\mathbf{R}^d)^N \times (\mathbf{R}^d)^N \to \mathbf{C}$ with trace

$$\int_{(\mathbf{R}^d)^N} \rho(x_1, \ldots, x_N; x_1, \ldots, x_N) \, dx_1 \ldots dx_N = 1,$$

with a pure state ψ being identified with the mixed state

$$\rho(x_1, \ldots, x_N; x'_1, \ldots, x'_N) := \psi(x_1, \ldots, x_N)\overline{\psi(x'_1, \ldots, x'_N)}.$$

In classical mechanics, the state of a single particle was the marginal distribution of the joint state. In quantum mechanics, the state of a single particle is instead obtained as the *partial trace* of the joint state. For instance, the state of the first particle is given as

$$\rho_1(x_1; x'_1) := \int_{(\mathbf{R}^d)^{N-1}} \rho(x_1, x_2, \ldots, x_N; x'_1, x_2, \ldots, x_N)\ dx_2 \ldots dx_N,$$

the state of the first two particles is given as

$$\rho_{12}(x_1, x_2; x'_1, x'_2) := \int_{(\mathbf{R}^d)^{N-2}} \rho(x_1, x_2, x_3, \ldots, x_N;$$
$$x'_1, x'_2, x_3, \ldots, x_N)\ dx_3 \ldots dx_N,$$

and so forth. (These formulae can be justified by considering observables of the joint state that only affect, say, the first two position coordinates x_1, x_2 and using duality.)

A typical Hamiltonian in this setting is given by the operator

$$H\psi(x_1, \ldots, x_N) = \sum_{j=1}^{N} \frac{|p_j|^2}{2m_j}\psi(x_1, \ldots, x_N)$$
$$+\frac{1}{N}\sum_{1 \le j < k \le N} V_{jk}(x_j - x_k)\psi(x_1, \ldots, x_N)$$

where we normalise just as in the classical case, and $p_j := -i\hbar\nabla_{x_j}$.

An interesting feature of quantum mechanics - not present in the classical world - is that even if the N-particle system is in a pure state, individual particles may be in a mixed state: the partial trace of a pure state need not remain pure. Because of this, when considering a subsystem of a larger system, one cannot always assume that the subsystem is in a pure state, but must work instead with mixed states throughout, unless there is some reason (e.g. a lack of coupling) to assume that pure states are somehow preserved.

Now consider a system of N indistinguishable quantum particles. As in the classical case, this means that all observables (including the Hamiltonian) for the joint system are invariant with respect to the action of the symmetric group S_N. Because of this, one may as well

assume that the (mixed) state of the joint system is also symmetric with respect to this action. In the special case when the particles are *bosons*, one can also assume that pure states $|\psi\rangle$ are also symmetric with respect to this action (in contrast to *fermions*, where the action on pure states is anti-symmetric). A typical Hamiltonian in this setting is given by the operator

$$H\psi(x_1,\ldots,x_N) = \sum_{j=1}^N \frac{|p_j|^2}{2m}\psi(x_1,\ldots,x_N)$$

$$+\frac{1}{N}\sum_{1\leq j<k\leq N} V(x_j-x_k)\psi(x_1,\ldots,x_N)$$

for some even potential V; if one wants to model short-range interactions, one might instead pick the variant

(3.89)
$$H\psi(x_1,\ldots,x_N) = \sum_{j=1}^N \frac{|p_j|^2}{2m}\psi(x_1,\ldots,x_N)+\frac{1}{N}\sum_{1\leq j<k\leq N} r^d V(\frac{x_j-x_k}{r})\psi(x_1,\ldots,x_N)$$

for some $r > 0$. This is a typical model for an N-particle *Bose-Einstein condensate*. (Longer-range models can lead to more non-local variants of NLS for the limiting equation, such as the *Hartree equation*.)

3.16.3. NLS. Suppose we have a Bose-Einstein condensate given by a (symmetric) mixed state

$$\rho(t,x_1,\ldots,x_N;x'_1,\ldots,x'_N)$$

evolving according to the equation of motion (3.88) using the Hamiltonian (3.89). One can take a partial trace of the equation of motion (3.88) to obtain an equation for the state $\rho_1(t,x_1;x'_1)$ of the first particle (note from symmetry that all the other particles will have the same state function). If one does take this trace, one soon finds that the equation of motion becomes

$$\partial_t\rho_1(t,x_1;x'_1) = \frac{i}{\hbar}[(\frac{|p_1|^2}{2m}-\frac{|p'_1|^2}{2m})\rho_1(t,x_1;x'_1)$$

$$+\frac{1}{N}\sum_{j=2}^N \int_{\mathbf{R}^d}\frac{1}{r^d}[V(\frac{x_1-x_j}{r})-V(\frac{x'_1-x_j}{r})]\rho_{1j}(t,x_1,x_j;x'_1,x_j)\,dx_j$$

where ρ_{1j} is the partial trace to the $1, j$ particles. Using symmetry, we see that all the summands in the j summation are identical, so we can simplify this as

$$\partial_t \rho_1(t, x_1; x_1') = \frac{i}{\hbar}[(\frac{|p_1|^2}{2m} - \frac{|p_1'|^2}{2m})\rho_1(t, x_1; x_1')$$
$$+\frac{N-1}{N}\int_{\mathbf{R}^d} \frac{1}{r^d}[V(\frac{x_1 - x_2}{r}) - V(\frac{x_1' - x_2}{r})]\rho_{12}(t, x_1, x_2; x_1', x_2)\ dx_2.$$

This does not completely describe the dynamics of ρ_1, as one also needs an equation for ρ_{12}. But one can repeat the same argument to get an equation for ρ_{12} involving ρ_{123}, and so forth, leading to a system of equations known as the *BBGKY hierarchy*. But for simplicity we shall just look at the first equation in this hierarchy.

Let us now formally take two limits in the above equation, sending the number of particles N to infinity and the interaction scale r to zero. The effect of sending N to infinity should simply be to eliminate the $\frac{N-1}{N}$ factor. The effect of sending r to zero should be to send $\frac{1}{r^d}V(\frac{x}{r})$ to the Dirac mass $\lambda\delta(x)$, where $\lambda := \int_{\mathbf{R}^d} V$ is the total mass of V. *Formally* performing these two limits, one is led to the equation

$$\partial_t \rho_1(t, x_1; x_1') = \frac{i}{\hbar}[(\frac{|p_1|^2}{2m} - \frac{|p_1'|^2}{2m})\rho_1(t, x_1; x_1')$$
$$+\lambda(\rho_{12}(t, x_1, x_1; x_1', x_1) - \rho_{12}(t, x_1, x_1'; x_1', x_1'))].$$

One can perform a similar formal limiting procedure for the other equations in the BBGKY hierarchy, obtaining a system of equations known as the *Gross-Pitaevskii hierarchy*.

We next make an important simplifying assumption, which is that in the limit $N \to \infty$ any two particles in this system become *decoupled*, which means that the two-particle mixed state factors as the tensor product of two one-particle states:

$$\rho_{12}(t, x_1, x_2; x_1', x_2) \approx \rho_1(t, x_1; x_1')\rho_1(t, x_2; x_2').$$

One can view this as a *mean field approximation*, modeling the interaction of one particle x_1 with all the other particles by the mean field ρ_1.

Making this assumption, the previous equation simplifies to

$$\partial_t \rho_1(t, x_1; x_1') = \frac{i}{\hbar}[(\frac{|p_1|^2}{2m} - \frac{|p_1'|^2}{2m})$$

$$+\lambda(\rho_1(t,x_1;x_1)-\rho_1(t,x_1';x_1'))]\rho_1(t,x_1;x_1').$$

If we assume furthermore that ρ_1 is a pure state, thus

$$\rho_1(t,x_1;x_1') = \psi(t,x_1)\overline{\psi(t,x_1')}$$

then (up to the phase ambiguity mentioned earlier), $\psi(t,x)$ obeys the *Gross-Pitaevskii equation*

$$\partial_t \psi(t,x) = \frac{i}{\hbar}[(\frac{|p|^2}{2m} + \lambda|\psi(t,x)|^2]\psi(t,x)$$

which (up to some factors of $\hbar$ and m, which can be renormalised away) is essentially (3.82).

An alternate derivation of (3.82), using a slight variant of the above mean field approximation, comes from studying the Hamiltonian (3.89). Let us make the (very strong) assumption that at some fixed time t, one is in a completely factored pure state

$$\psi(x_1,\ldots,x_N) = \psi_1(x_1)\ldots\psi_1(x_N),$$

where ψ_1 is a one-particle wave function, in particular obeying the normalisation

$$\int_{\mathbf{R}^d} |\psi_1(x)|^2\ dx = 1.$$

(This is an unrealistically strong version of the mean field approximation. In practice, one only needs the two-particle partial traces to be completely factored for the discussion below.) The expected value of the Hamiltonian,

$$\langle\psi|H|\psi\rangle = \int_{(\mathbf{R}^d)^N} \psi(x_1,\ldots,x_N)\overline{H\psi(x_1,\ldots,x_N)}\ dx_1\ldots dx_N,$$

can then be simplified as

$$N\int_{\mathbf{R}^d} \psi_1(x)\overline{\frac{|p_1|^2}{2m}\psi_1(x)}\ dx$$

$$+\frac{N-1}{2}\int_{\mathbf{R}^d\times\mathbf{R}^d} r^{-d}V(\frac{x_1-x_2}{r})|\psi_1(x_1)|^2|\psi_1(x_2)|\ dx_1dx_2.$$

Again sending $r\to 0$, this formally becomes

$$N\int_{\mathbf{R}^d} \psi_1(x)\overline{\frac{|p_1|^2}{2m}\psi_1(x)}\ dx + \frac{N-1}{2}\lambda\int_{\mathbf{R}^d\times\mathbf{R}^d} |\psi_1(x_1)|^4\ dx_1$$

which in the limit $N \to \infty$ is asymptotically

$$N \int_{\mathbf{R}^d} \psi_1(x) \overline{\frac{|p_1|^2}{2m} \psi_1(x)} + \frac{\lambda}{2} |\psi_1(x_1)|^4 \, dx_1.$$

Up to some normalisations, this is the Hamiltonian for the NLS equation (3.82).

There has been much progress recently in making the above derivations precise, see e.g. [**Sc2006**], [**KlMa2008**], [**KiScSt2008**], [**ChPa2009**]. A key step is to show that the Gross-Pitaevskii hierarchy necessarily preserves the property of being a completely factored state. This requires a uniqueness theory for this hierarchy, which is surprisingly delicate, due to the fact that it is a system of infinitely many coupled equations over an unbounded number of variables.

Remark 3.16.1. Interestingly, the above heuristic derivation only works when the interaction scale r is much larger than N^{-1}. For $r \sim N^{-1}$, the coupling constant λ acquires a nonlinear correction, becoming essentially the *scattering length* of the potential rather than its mean. (Thanks to Bob Jerrard for pointing out this subtlety.)

Notes. This article first appeared at `terrytao.wordpress.com/2009/11/26`. Thanks to CJ, liuyao, Mio and M.S. for corrections.

Bob Jerrard provided a heuristic argument as to why the coupling constant becomes nonlinear in the regime $r \sim N^{-1}$.

Chapter 4

Technical articles

4.1. Polymath1 and three new proofs of the density Hales-Jewett theorem

During the first few months of 2009, I was involved in the *Polymath1 project*, a massively collaborative mathematical project whose purpose was to investigate the viability of various approaches to proving the *density Hales-Jewett theorem*. For simplicity I will focus attention here on the model case $k = 3$ of a three-letter alphabet, in which case the theorem reads as follows:

Theorem 4.1.1 ($k = 3$ density Hales-Jewett theorem)**.** *Let $0 < \delta \leq 1$. Then if n is a sufficiently large integer, any subset A of the cube $[3]^n = \{1, 2, 3\}^n$ of density $|A|/3^n$ at least δ contains at least one combinatorial line $\{\ell(1), \ell(2), \ell(3)\}$, where $\ell \in \{1, 2, 3, x\}^n \backslash [3]^n$ is a string of 1s, 2s, 3s, and x's containing at least one "wildcard" x, and $\ell(i)$ is the string formed from ℓ by replacing all x's with i's.*

The full density Hales-Jewett theorem is the same statement, but with [3] replaced by $[k]$ for some $k \geq 1$. (The case $k = 1$ is trivial, and the case $k = 2$ follows from *Sperner's theorem*.) As a result of the project, three new proofs of this theorem were established, at least one of which has extended[**Po2009**] to cover the case of general k.

This theorem was first proven by Furstenberg and Katznelson[**FuKa1989**], by first converting it to a statement in ergodic theory; the original paper of Furstenberg-Katznelson argument was for the $k = 3$ case only, and gave only part of the proof in detail, but in a subsequent paper[**FuKa1991**] a full proof in general k was provided. The remaining components of the original $k = 3$ argument were later completed in unpublished notes of McCutcheon[1]. One of the new proofs is essentially a finitary translation of this $k = 3$ argument; in principle one could also finitise the significantly more complicated argument of Furstenberg and Katznelson for general k, but this has not been properly carried out yet (the other two proofs are likely to generalise much more easily to higher k). The result is considered quite deep; for instance, the general k case of the density Hales-Jewett theorem already implies *Szemerédi's theorem*, which is a highly non-trivial theorem in its own right, as a special case.

[1] `http://www.msci.memphis.edu/ randall/preprints/HJk3.pdf`

Another of the proofs is based primarily on the density increment method that goes back to Roth, and also incorporates some ideas from a paper of Ajtai and Szemerédi[**AjSz1974**] establishing what we have called the *corners theorem* (and which is also implied by the $k = 3$ case of the density Hales-Jewett theorem). A key new idea involved studying the correlations of the original set A with special subsets of $[3]^n$, such as *ij-insensitive sets*, or intersections of ij-insensitive and ik-insensitive sets.

This correlations idea inspired a new ergodic proof of the density Hales-Jewett theorem for all values of k by Austin[**Au2009b**], which is in the spirit of the *triangle removal lemma* (or hypergraph removal lemma) proofs of *Roth's theorem* (or the *multidimensional Szemerédi theorem*). A finitary translation of this argument in the $k = 3$ case has been sketched out; I believe it also extends in a relatively straightforward manner to the higher k case (in analogy with some proofs of the hypergraph removal lemma).

4.1.1. Simpler cases of density Hales-Jewett. In order to motivate the known proofs of the density Hales-Jewett theorem, it is instructive to consider some simpler theorems which are implied by this theorem. The first is the *corners theorem* of Ajtai and Szemerédi:

Theorem 4.1.2 (Corners theorem)**.** *Let $0 < \delta \leq 1$. Then if n is a sufficiently large integer, any subset A of the square $[n]^2$ of density $|A|/n^2$ at least δ contains at least one right-angled triangle (or "corner") $\{(x, y), (x + r, y), (x, y + r)\}$ with $r \neq 0$.*

The $k = 3$ density Hales-Jewett theorem implies the corners theorem; this is proven by utilising the map $\phi : [3]^n \to [n]^2$ from the cube to the square, defined by mapping a string $x \in [3]^n$ to a pair (a, b), where a, b are the number of 1s and 2s respectively in x. The key point is that ϕ maps combinatorial lines to corners. (Strictly speaking, this mapping only establishes the corners theorem for dense subsets of $[n/3 - \sqrt{n}, n/3 + \sqrt{n}]^2$, but it is not difficult to obtain the general case from this by replacing n by n^2 and using translation-invariance.)

The corners theorem is also closely related to the problem of finding dense sets of points in a triangular grid without any equilateral triangles, a problem which we have called *Fujimura's problem*.

The corners theorem in turn implies

Theorem 4.1.3 (Roth's theorem)**.** *Let $0 < \delta \leq 1$. Then if n is a sufficiently large integer, any subset A of the interval $[n]$ of density $|A|/n$ at least δ contains at least one arithmetic progression $a, a + r, a + 2r$ of length three.*

Roth's theorem can be deduced from the corners theorem by considering the map $\psi : [n]^2 \to [3n]$ defined by $\psi(a, b) := a + 2b$; the key point is that ψ maps corners to arithmetic progressions of length three.

There are higher k analogues of these implications; the general k version of the density Hales-Jewett theorem implies a general k version of the corners theorem known as the multidimensional Szemerédi theorem, which in term implies a general version of Roth's theorem known as *Szemerédi's theorem*.

4.1.2. The density increment argument. The strategy of the density increment argument, which goes back to Roth's proof[**Ro1953**] of Theorem 4.1.3, is to perform a downward induction on the density δ. Indeed, the theorem is obvious for high enough values of δ; for instance, if $\delta > 2/3$, then partitioning the cube $[3]^n$ into lines and applying the pigeonhole principle will already give a combinatorial line. So the idea is to deduce the claim for a fixed density δ from that of a higher density δ.

A key concept here is that of an m-dimensional *combinatorial subspace* of $[3]^n$ - a set of the form $\phi([3]^m)$, where $\phi \in \{1, 2, 3, *_1, \ldots, *_m\}^n$ is a string formed using the base alphabet and m wildcards $*_1, \ldots, *_m$ (with each wildcard appearing at least opnce), and $\phi(a_1 \ldots a_m)$ is the string formed by substituting a_i for $*_i$ for each i. (Thus, for instance, a combinatorial line is a combinatorial subspace of dimension 1.) The identification ϕ between $[3]^m$ and the combinatorial space $\phi([3]^m)$ maps combinatorial lines to combinatorial lines. Thus, to prove Theorem 4.1.1, it suffices to show

Proposition 4.1.4 (Lack of lines implies density increment)**.** *Let $0 < \delta \leq 1$. Then if n is a sufficiently large integer, and $A \subset [3]^n$ has density at least δ and has no combinatorial lines, then there exists*

an m-dimensional subspace $\phi([3]^m)$ of $[3]^n$ on which A has density at least $\delta + c(\delta)$, where $c(\delta) > 0$ depends only on δ (and is bounded away from zero on any compact range of δ), and $m \geq m_0(n, \delta)$ for some function $m_0(n, \delta)$ that goes to infinity as $n \to \infty$ for fixed δ.

It is easy to see that Proposition 4.1.4 implies Theorem 4.1.1 (for instance, one could consider the infimum of all δ for which the theorem holds, and show that having this infimum non-zero would lead to a contradiction).

Now we have to figure out how to get that density increment. The original argument of Roth relied on Fourier analysis, which in turn relies on an underlying translation-invariant structure which is not present in the density Hales-Jewett setting. (Arithmetic progressions are translation-invariant, but combinatorial lines are not.) It turns out that one can proceed instead by adapting a (modification of) an argument of Ajtai and Szemerédi, which gave the first proof of Theorem 4.1.2.

The (modified) Ajtai-Szemerédi argument uses the density increment method, assuming that A has no right-angled triangles and showing that A has an increased density on a *subgrid* - a product $P \times Q$ of fairly long arithmetic progressions with the same spacing. The argument proceeds in two stages, which we describe slightly informally (in particular, glossing over some technical details regarding quantitative parameters such as ε) as follows:

- Step 1. If $A \subset [n]^2$ is dense but has no right-angled triangles, then A has an increased density on a cartesian product $U \times V$ of dense sets $U, V \subset [n]$ (which are not necessarily arithmetic progressions).
- Step 2. Any Cartesian product $U \times V$ in $[n]^2$ can be partitioned into reasonably large grids $P \times Q$, plus a remainder term of small density.

From Step 1, Step 2 and the pigeonhole principle we obtain the desired density increment of A on a grid $P \times Q$, and then the density increment argument gives us the corners theorem.

Step 1 is actually quite easy. If A is dense, then it must also be dense on some diagonal $D = \{(x, y) : x + y = const\}$, by the

pigeonhole principle. Let U and V denote the rows and columns that $A \cap D$ occupies. Every pair of points in $A \cap D$ forms the hypotenuse of some corner, whose third vertex lies in $U \times V$. Thus, if A has no corners, then A must avoid all the points formed by $U \times V$ (except for those of the diagonal D). Thus A has a significant density *decrease* on the Cartesian product $U \times V$. Dividing the remainder $[n]^2 \backslash (U \times V)$ into three further Cartesian products $U \times ([n] \backslash V)$, $([n] \backslash U) \times V$, $([n] \backslash U) \times ([n] \backslash V)$ and using the pigeonhole principle we obtain the claim (after redefining U, V appropriately).

Step 2 can be obtained by iterating a one-dimensional version:

- Step 2a. Any set $U \subset [n]$ can be partitioned into reasonably long arithmetic progressions P, plus a remainder term of small density.

Indeed, from Step 2a, one can partition $U \times [n]$ into products $P \times [n]$ (plus a small remainder), which can be easily repartitioned into grids $P \times Q$ (plus small remainder). This partitions $U \times V$ into sets $P \times (V \cap Q)$ (plus small remainder). Applying Step 2a again, each $V \cap Q$ can be partitioned further into progressions Q' (plus small remainder), which allows us to partition each $P \times (V \cap Q)$ into grids $P' \times Q'$ (plus small remainder).

So all one has left to do is establish Step 2a. But this can be done by the greedy algorithm: locate one long arithmetic progression P in U and remove it from U, then locate another to remove, and so forth until no further long progressions remain in the set. But *Szemerédi's theorem* then tells us the remaining set has low density, and one is done!

This argument has the apparent disadvantage of requiring a deep theorem (Szemerédi's theorem) in order to complete the proof. However, interestingly enough, when one adapts the argument to the density Hales-Jewett theorem, one gets to replace Szemerédi's theorem by a more elementary result - one which in fact follows from the (easy) $k = 2$ version of the density Hales-Jewett theorem, i.e. *Sperner's theorem*.

We first need to understand the analogue of the Cartesian products $U \times V$. Note that $U \times V$ is the intersection of a "vertically

insensitive set" $U \times [n]$ and a "horizontally insensitive set" $[n] \times V$. By "vertically insensitive" we mean that membership of a point (x, y) in that set is unaffected if one moves that point in a vertical direction, and similarly for "horizontally insensitive". In a similar fashion, define a "12-insensitive set" to be a subset of $[3]^n$, membership in which is unaffected if one flips a coordinate from a 1 to a 2 or vice versa (e.g. if 1223 lies in the set, then so must 1213, 1113, 2113, etc.). Similarly define the notion of a "13-insensitive set". We then define a "complexity 1 set" to be the intersection $E_{12} \cap E_{13}$ of a 12-insensitive set E_{12} and a 13-insensitive set E_{13}; these are analogous to the Cartesian products $U \times V$.

(For technical reasons, one actually has to deal with *local* versions of insensitive sets and complexity 1 sets, in which one is only allowed to flip a moderately small number of the n coordinates rather than all of them. But to simplify the discussion let me ignore this (important) detail, which is also a major issue to address in the other two proofs of this theorem.)

The analogues of Steps 1, 2 for the density Hales-Jewett theorem are then

- Step 1. If $A \subset [3]^n$ is dense but has no combinatorial lines, then A has an increased density on a (local) complexity 1 set $E_{12} \cap E_{13}$.
- Step 2. Any (local) complexity 1 set $E_{12} \cap E_{13} \subset [3]^n$ can be partitioned into moderately large combinatorial subspaces (plus a small remainder).

We can sketch how Step 1 works as follows. Given any $x \in [3]^n$, let $\pi_{1\to 2}(x)$ denote the string formed by replacing all 1s with 2s, e.g. $\pi_{1\to 2}(1321) = 2322$. Similarly define $\pi_{1\to 3}(x)$. Observe that $x, \pi_{1\to 2}(x), \pi_{1\to 3}(x)$ forms a combinatorial line (except in the rare case when x doesn't contain any 1s). Thus if we let $E_{12} := \{x : \pi_{1\to 2}(x) \in A\}$, $E_{13} := \{x : \pi_{1\to 3}(x) \in A\}$, we see that A must avoid essentially all of $E_{12} \cap E_{13}$. On the other hand, observe that E_{12} and E_{13} are 12-insensitive and 13-insensitive sets respectively. Taking complements and using the same sort of pigeonhole argument as before, we obtain the claim. (Actually, this argument doesn't quite work because E_{12},

E_{13} could be very sparse; this problem can be fixed, but requires one to use local complexity 1 sets rather than global ones, and also to introduce the concept of "equal-slices measure"; I will not discuss these issues here.)

Step 2 can be reduced, much as before, to the following analogue of Step 2a:

- Step 2a. Any 12-insensitive set $E_{12} \subset [3]^n$ can be partitioned into moderately large combinatorial subspaces (plus a small remainder).

Identifying the letters 1 and 2 together, one can quotient $[3]^n$ down to $[2]^n$; the preimages of this projection are precisely the 12-insensitive sets. Because of this, Step 2a is basically equivalent (modulo some technicalities about measure) to

- Step 2a'. Any $E \subset [2]^n$ can be partitioned into moderately large combinatorial subspaces (plus a small remainder).

By the greedy algorithm, we will be able to accomplish this step if we can show that every dense subset of $[2]^n$ contains moderately large subspaces. But this turns out to be possible by carefully iterating Sperner's theorem (which shows that every dense subset of $[2]^n$ contains combinatorial lines).

This proof of Theorem 4.1.1 extends without major difficulty to the case of higher k; see [**Po2009**].

4.1.3. The triangle removal argument. The *triangle removal lemma* of Ruzsa and Szemerédi[**RuSz1978**] is a graph-theoretic result which implies the corners theorem (and hence Roth's theorem). It asserts the following:

Lemma 4.1.5 (Triangle removal lemma)**.** *For every $\varepsilon > 0$ there exists $\delta > 0$ such that if a graph G on n vertices has fewer than δn^3 triangles, then the triangles can be deleted entirely by removing at most εn^2 edges.*

Let's see how the triangle removal lemma implies the corners theorem. A corner is, of course, already a triangle in the geometric sense, but we need to convert it to a triangle in the graph-theoretic

sense, as follows. Let A be a subset of $[n]^2$ with no corners; the aim is to show that A has small density. Let V_h be the set of all horizontal lines in $[n]^2$, V_v the set of vertical lines, and V_d the set of diagonal lines (thus all three sets have size about n). We create a tripartite graph G on the vertex sets $V_h \cup V_v \cup V_d$ by joining a horizontal line $h \in V_h$ to a vertical line $v \in V_v$ whenever h and v intersect at a point in A, and similarly connecting V_h or V_v to V_d. Observe that a triangle in G corresponds either to a corner in A, or to a "degenerate" corner in which the horizontal, vertical, and diagonal line are all concurrent. In particular, there are very few triangle in G, which can then be deleted by removing a small number of edges from G by the triangle removal lemma. But each edge removed can delete at most one degenerate corner, and the number of degenerate corners is $|A|$, and so $|A|$ is small as required.

All known proofs of the triangle removal lemma proceed by some version of the following three steps:

- "Regularity lemma step": Applying tools such as the *Szemerédi regularity lemma*, one can partition the graph G into components G_{ij} between cells V_i, V_j of vertices, such that most of the G_{ij} are "pseudorandom". One way to define what pseudorandom means is to view each graph component G_{ij} as a subset of the Cartesian product $V_i \times V_j$, in which case G_{ij} is pseudorandom if it does not have a significant density increment on any smaller Cartesian product $V_i' \times V_j'$ of non-trivial size.

- "Counting lemma step": By exploiting the pseudorandomness property, one shows that if G has a triple G_{ij}, G_{jk}, G_{ki} of dense pseudorandom graphs between cells V_i, V_j, V_k of non-trivial size, then this triple must generate a large number of triangles; hence, if G has very few triangles, then one cannot find such a triple of dense pseudorandom graphs.

- "Cleaning step": If one then removes all components of G which are too sparse or insufficiently pseudorandom, one can thus eliminate all triangles.

Pulling this argument back to the corners theorem, we see that cells such as V_i, V_j, V_k will correspond either to horizontally insensitive sets, vertically insensitive sets, or diagonally insensitive sets. Thus this proof of the corners theorem proceeds by partitioning $[n]^2$ in three different ways into insensitive sets in such a way that A is pseudorandom with respect to many of the cells created by any two of these partitions, counting the corners generated by any triple of large cells in which A is pseudorandom and dense, and cleaning out all the other cells.

It turns out that a variant of this argument can give Theorem 4.1.1; this was in fact the original approach studied by the polymath1 project, though it was only after a detour through ergodic theory (as well as the development of the density-increment argument discussed above) that the triangle-removal approach could be properly executed. In particular, an ergodic argument based on the infinitary analogue of the triangle removal lemma (and its hypergraph generalisations) was developed by Austin[**Au2009b**], which then inspired the combinatorial version sketched here.

The analogue of the vertex cells V_i are given by certain 12-insensitive sets E^a_{12}, 13-insensitive sets E^b_{13}, and 23-insensitive sets E^c_{23}. Roughly speaking, a set $A \subset [3]^n$ would be said to be pseudorandom with respect to a cell $E^a_{12} \cap E^b_{13}$ if $A \cap E^a_{12} \cap E^b_{13}$ has no further density increment on any smaller cell $E'_{12} \cap E'_{13}$ with E'_{12} a 12-insensitive subset of E^a_{12}, and E'_{13} a 13-insensitive subset of E^b_{13}. (This is an oversimplification, glossing over an important refinement of the concept of pseudorandomness involving the discrepancy between global densities in $[3]^n$ and local densities in subspaces of $[3]^n$.) There is a similar notion of A being pseudorandom with respect to a cell $E^b_{13} \cap E^c_{23}$ or $E^c_{23} \cap E^a_{12}$.

We briefly describe the "regularity lemma" step. By modifying the proof of the regularity lemma, one can obtain three partitions

$$[3]^n = E^1_{12} \cup \ldots \cup E^{M_{12}}_{12} = E^1_{13} \cup \ldots \cup E^{M_{13}}_{13} = E^1_{23} \cup \ldots \cup E^{M_{23}}_{23}$$

into 12-insensitive, 13-insensitive, and 23-insensitive components respectively, where M_{12}, M_{13}, M_{23} are not too large, and A is pseudorandom with respect to most cells $E^a_{12} \cap E^b_{13}$, $E^b_{13} \cap E^c_{23}$, and $E^c_{23} \cap E^a_{12}$.

In order for the counting step to work, one also needs an additional "stationarity" reduction, which is difficult to state precisely, but roughly speaking asserts that the "local" statistics of sets such as E_{12}^a on medium-dimensional subspaces are close to the corresponding "global" statistics of such sets; this can be achieved by an additional pigeonholing argument. We will gloss over this issue, pretending that there is no distinction between local statistics and global statistics. (Thus, for instance, if E_{12}^a has large global density in $[3]^n$, we shall assume that E_{12}^a also has large density on most medium-sized subspaces of $[3]^n$.)

Now for the "counting lemma" step. Suppose we can find a, b, c such that the cells $E_{12}^a, E_{13}^b, E_{23}^c$ are large, and that A intersects $E_{12}^a \cap E_{13}^b$, $E_{13}^b \cap E_{23}^c$, and $E_{23}^c \cap E_{12}^a$ in a dense pseudorandom manner. We claim that this will force A to have a large number of combinatorial lines ℓ, with $\ell(1)$ in $A \cap E_{12}^a \cap E_{13}^b$, $\ell(2)$ in $A \cap E_{23}^c \cap E_{12}^a$, and $\ell(3)$ in $A \cap E_{13}^b \cap E_{23}^c$. Because of the dense pseudorandom nature of A in these cells, it turns out that it will suffice to show that there are a lot of lines $\ell(1)$ with $\ell(1) \in E_{12}^a \cap E_{13}^b$, $\ell(2) \in E_{23}^c \cap E_{12}^a$, and $\ell(3) \in E_{13}^b \cap E_{23}^c$.

One way to generate a line ℓ is by taking the triple $\{x, \pi_{1\to 2}(x), \pi_{1\to 3}(x)\}$, where $x \in [3]^n$ is a generic point. (Actually, as we will see below, we would have to to a subspace of $[3]^n$ before using this recipe to generate lines.) Then we need to find many x obeying the constraints

$$x \in E_{12}^a \cap E_{13}^b; \quad \pi_{1\to 2}(x) \in E_{23}^c \cap E_{12}^a; \quad \pi_{1\to 3}(x) \in E_{13}^b \cap E_{23}^c.$$

Because of the various insensitivity properties, many of these conditions are redundant, and we can simplify to

$$x \in E_{12}^a \cap E_{13}^b; \quad \pi_{1\to 2}(x) \in E_{23}^c.$$

Now note that the property "$\pi_{1\to 2}(x) \in E_{23}^c$" is 123-insensitive; it is simultaneously 12-insensitive, 23-insensitive, and 13-insensitive. As E_{23}^c is assumed to be large, there will be large combinatorial subspaces on which (a suitably localised version of) this property "$\pi_{1\to 2}(x) \in E_{23}^c$" will be always true. Localising to this space (taking advantage of the stationarity properties alluded to earlier), we are now looking for solutions to

$$x \in E_{12}^a \cap E_{13}^b.$$

We'll pick x to be of the form $\pi_{2\to 1}(y)$ for some y. We can then rewrite the constraints on y as

$$y \in E_{12}^a; \quad \pi_{2\to 1}(y) \in E_{13}^b.$$

The property "$\pi_{2\to 1}(y) \in E_{13}^b$" is 123-invariant, and E_{13}^b is large, so by arguing as before we can pass to a large subspace where this property is always true. The largeness of E_{12}^a then gives us a large number of solutions.

Taking contrapositives, we conclude that if A in fact has no combinatorial lines, then there do not exist any triple $E_{12}^a, E_{13}^b, E_{23}^c$ of large cells with respect to which A is dense and pseudorandom. This forces A to be confined either to very small cells, or to very sparse subsets of cells, or to the rare cells which fail to be pseudorandom. None of these cases can contribute much to the density of A, and so A itself is very sparse - contradicting the hypothesis in Theorem 4.1.1 that A is dense (this is the "cleaning step"). This concludes the sketch of the triangle-removal proof of this theorem.

The ergodic version of this argument in [**Au2009b**] works for all values of k, so I expect the combinatorial version to do so as well.

4.1.4. The finitary Furstenberg-Katznelson argument. In [**FuKa1989**], Furstenberg and Katznelson gave the first proof of Theorem 4.1.1, by translating it into a recurrence statement about a certain type of stationary process indexed by an infinite cube $[3]^\omega := \bigcup_{n=1}^\infty [3]^n$. This argument was inspired by a long string of other successful proofs of density Ramsey theorems via ergodic means, starting with the initial paper of Furstenberg[**Fu1977**] giving an ergodic theory proof of Szemerédi's theorem. The latter proof was transcribed into a finitary language in [**Ta2006b**], so it was reasonable to expect that the Furstenberg-Katznelson argument could similarly be translated into a combinatorial framework.

Let us first briefly describe the original strategy of Furstenberg to establish Roth's theorem, but phrased in an informal, and vaguely combinatorial, language. The basic task is to get a non-trivial lower bound on averages of the form

$$\mathbf{E}_{a,r} f(a) f(a+r) f(a+2r) \tag{4.1}$$

where we will be a bit vague about what a, r are ranging over, and where f is some non-negative function of positive mean. It is then natural to study more general averages of the form

$$\mathbf{E}_{a,r} f(a)g(a+r)h(a+2r). \tag{4.2}$$

Now, it turns out that certain types of functions f, g, h give a negligible contribution to expressions such as (4.2). In particular, if f is *weakly mixing*, which roughly means that the pair correlations

$$\mathbf{E}_a f(a)f(a+r)$$

are small for most r, then the average (4.2) is small no matter what g, h are (so long as they are bounded). This can be established by some applications of the Cauchy-Schwarz inequality (or its close cousin, the *van der Corput lemma*). As a consequence of this, all weakly mixing components of f can essentially be discarded when considering an average such as (4.1).

After getting rid of the weakly mixing components, what is left? Being weakly mixing is like saying that almost all the shifts $f(\cdot + r)$ of f are close to orthogonal to each other. At the other extreme is that of *periodicity* - the shifts $f(\cdot + r)$ periodically recur to become equal to f again. There is a slightly more general notion of *almost periodicity* - roughly, this means that the shifts $f(\cdot + r)$ don't have to recur exactly to f again, but they are forced to range in a precompact set, which basically means that for every $\varepsilon > 0$, that $f(\cdot + r)$ lies within ε (in some suitable norm) of some finite-dimensional space. A good example of an almost periodic function is an *eigenfunction*, in which we have $f(a + r) = \lambda_r f(a)$ for each r and some quantity λ_r independent of a (e.g. one can take $f(a) = e^{2\pi i \alpha a}$ for some $\alpha \in \mathbf{R}$). In this case, the finite-dimensional space is simply the scalar multiples of $f(a)$ (and one can even take $\varepsilon = 0$ in this special case).

It is easy to see that non-trivial almost periodic functions are not weakly mixing; more generally, any function which correlates non-trivially with an almost periodic function can also be seen to not be weakly mixing. In the converse direction, it is also fairly easy to show that any function which is not weakly mixing must have non-trivial correlation with an almost periodic function. Because of this, it turns out that one can basically decompose *any* function into almost

periodic and weakly mixing components. For the purposes of getting lower bounds on (4.1), this allows us to essentially reduce matters to the special case when f is almost periodic. But then the shifts $f(\cdot + r)$ are almost ranging in a finite-dimensional set, which allows one to essentially assign each shift r a colour from a finite range of colours. If one then applies the *van der Waerden theorem*, one can find many arithmetic progressions $a, a+r, a+2r$ which have the same colour, and this can be used to give a non-trivial lower bound on (4.1). (Thus we see that the role of a compactness property such as almost periodicity is to reduce density Ramsey theorems to colouring Ramsey theorems.)

This type of argument can be extended to more advanced recurrence theorems, but certain things become more complicated. For instance, suppose one wanted to count progressions of length 4; this amounts to lower bounding expressions such as

$$\mathbf{E}_{a,r} f(a) f(a+r) f(a+2r) f(a+3r). \tag{4.3}$$

It turns out that f being weakly mixing is no longer enough to give a negligible contribution to expressions such as (4.3). For that, one needs the stronger property of being *weakly mixing relative to almost periodic functions*; roughly speaking, this means that for most r, the expression $f(\cdot)f(\cdot+r)$ is not merely of small mean (which is what weak mixing would mean), but that this expression furthermore does not correlate strongly with any almost periodic function (i.e. $\mathbf{E}_a f(a) f(a+r) g(a)$ is small for any almost periodic g). Once one has this stronger weak mixing property, then one can discard all components of f which are weakly mixing relative to almost periodic functions.

One then has to figure out what is left after all these components are discarded. Because we strengthened the notion of weak mixing, we have to weaken the notion of almost periodicity to compensate. The correct notion is no longer that of almost periodicity - in which the shifts $f(\cdot+r)$ almost take values in a finite-dimensional vector space - but that of almost periodicity *relative* to almost periodic functions, in which the shifts almost take values in a finite-dimensional *module* over the algebra of almost periodic functions. A good example of such a beast is that of a *quadratic eigenfunction*, in which we have $f(a+r) = \lambda_r(a) f(a)$ where $\lambda_r(a)$ is itself an ordinary eigenfunction, and thus

almost periodic in the ordinary sense; here, the relative module is the one-dimensional module formed by almost periodic multiples of f. (A typical example of a quadratic eigenfunction is $f(a) = e^{2\pi i \alpha a^2}$ for some $\alpha \in \mathbf{R}$.)

It turns out that one can "relativise" all of the previous arguments to the almost periodic "factor", and decompose an arbitrary f into a component which is weakly mixing relative to almost periodic functions, and another component which is almost periodic relative to almost periodic functions. The former type of components can be discarded. For the latter, we can once again start colouring the shifts $f(\cdot + r)$ with a finite number of colours, but with the caveat that the colour assigned is no longer independent of a, but depends in an almost periodic fashion on a. Nevertheless, it is still possible to combine the van der Waerden colouring Ramsey theorem with the theory of recurrence for ordinary almost periodic functions to get a lower bound on (4.3) in this case. One can then iterate this argument to deal with arithmetic progressions of longer length, but one now needs to consider even more intricate notions of almost periodicity, e.g. almost periodicity relative to (almost periodic functions relative to almost periodic functions), etc.

It turns out that these types of ideas can be adapted (with some effort) to the density Hales-Jewett setting. It's simplest to begin with the $k = 2$ situation rather than the $k = 3$ situation. Here, we are trying to obtain non-trivial lower bounds for averages of the form

$$\mathbf{E}_\ell f(\ell(1)) f(\ell(2)) \tag{4.4}$$

where ℓ ranges in some fashion over combinatorial lines in $[2]^n$, and f is some non-negative function with large mean.

The analogues of weakly mixing and almost periodic in this setting are the 12-uniform and 12-low influence functions respectively. Roughly speaking, a function is 12-low influence if its value usually doesn't change much if a 1 is flipped to a 2 or vice versa (e.g. the indicator function of a 12-insensitive set is 12-low influence); conversely, a 12-uniform function is a function g such that $\mathbf{E}_\ell f(\ell(1)) g(\ell(2))$ is small for all (bounded) f. One can show that any function can be decomposed, more or less orthogonally, into a 12-uniform function

and a 12-low influence function, with the upshot being that one can basically reduce the task of lower bounding (4.4) to the case when f is 12-low influence. But then $f(\ell(1))$ and $f(\ell(2))$ are approximately equal to each other, and it is straightforward to get a lower-bound in this case.

Now we turn to the $k = 3$ setting, where we are looking at lower-bounding expressions such as

$$\mathbf{E}_\ell f(\ell(1))g(\ell(2))h(\ell(3)) \tag{4.5}$$

with $f = g = h$.

It turns out that g (say) being 12-uniform is no longer enough to give a negligible contribution to the average (4.5). Instead, one needs the more complicated notion of g being 12-uniform relative to 23-low influence functions; this means that not only are the averages $\mathbf{E}_\ell f(\ell(1))g(\ell(2))$ small for all bounded f, but furthermore $\mathbf{E}_\ell f(\ell(1))g(\ell(2))h(\ell)$ is small for all bounded f and all 23-low influence h (there is a minor technical point here that h is a function of a line rather than of a point, but this should be ignored). Any component of g in (4.5) which is 12-uniform relative to 23-low influence functions are negligible and so can be removed.

One then needs to figure out what is left in g when these components are removed. The answer turns out to be functions g that are 12-almost periodic relative to 23-low influence. The precise definition of this concept is technical, but very roughly speaking it means that if one flips a digit from a 1 to a 2, then the value of g changes in a manner which is controlled by 23-low influence functions. Anyway, the upshot is that one can reduce g in (4.5) from f to the components of f which are 12-almost periodic relative to 23-low influence. Similarly, one can reduce h in (4.5) from f to the components of f which are 13-almost periodic relative to 23-low influence.

At this point, one has to use a colouring Ramsey theorem - in this case, the *Graham-Rothschild theorem* - in conjunction with the relative almost periodicity to locate lots of places in which $g(\ell(2))$ is close to $g(\ell(1))$ while $h(\ell(3))$ is simultaneously close to $h(\ell(1))$. This turns (4.5) into an expression of the form $\mathbf{E}_x f(x)g(x)h(x)$, which

turns out to be relatively easy to lower bound (because g, h, being projections of f, tend to be large wherever f is large).

Notes. This article first appeared at `terrytao.wordpress.com/2009/04/02`. Thanks to Ben, Daniel, Kevin O'Bryant, Sune Kristian Jakobsen, and anonymous commenters for corrections.

More information about the Polymath1 project can be found at `http://michaelnielsen.org/polymath1/index.php?title=Main_Page`.

4.2. Szemerédi's regularity lemma via random partitions

In the theory of dense graphs on n vertices, where n is large, a fundamental role is played by the *Szemerédi regularity lemma*:

Lemma 4.2.1 (Regularity lemma, standard version)**.** *Let $G = (V, E)$ be a graph on n vertices, and let $\varepsilon > 0$ and $k_0 \geq 0$. Then there exists a partition of the vertices $V = V_1 \cup \ldots \cup V_k$, with $k_0 \leq k \leq C(k_0, \varepsilon)$ bounded below by k_0 and above by a quantity $C(k_0, \varepsilon)$ depending only on k_0, ε, obeying the following properties:*

- *(Equitable partition) For any $1 \leq i, j \leq k$, the cardinalities $|V_i|, |V_j|$ of V_i and V_j differ by at most 1.*
- *(Regularity) For all but at most εk^2 pairs $1 \leq i < j \leq k$, the portion of the graph G between V_i and V_j is ε-*regular *in the sense that one has*

$$|d(A, B) - d(V_i, V_j)| \leq \varepsilon$$

for any $A \subset V_i$ and $B \subset V_j$ with $|A| \geq \varepsilon |V_i|, |B| \geq \varepsilon |V_j|$, where $d(A, B) := |E \cap (A \times B)|/|A||B|$ is the density of edges between A and B.

This lemma becomes useful in the regime when n is very large compared to k_0 or $1/\varepsilon$, because all the conclusions of the lemma are uniform in n. Very roughly speaking, it says that "up to errors of size ε", a large graph can be more or less described completely by a bounded number of quantities $d(V_i, V_j)$. This can be interpreted as saying that the space of all graphs is *totally bounded* (and hence *precompact*) in a suitable metric space, thus allowing one to take

formal limits of sequences (or subsequences) of graphs; see for instance [**LoSz2007**] for a discussion.

For various technical reasons it is easier to work with a slightly weaker version of the lemma, which allows for the cells $V_1, \ldots, V_k$ to have unequal sizes:

Lemma 4.2.2 (Regularity lemma, weighted version)**.** *Let $G = (V, E)$ be a graph on n vertices, and let $\varepsilon > 0$. Then there exists a partition of the vertices $V = V_1 \cup \ldots \cup V_k$, with $1 \leq k \leq C(\varepsilon)$ bounded above by a quantity $C(\varepsilon)$ depending only on ε, obeying the following properties:*

- *(Regularity) One has*

$$\sum_{(V_i, V_j) \text{ not } \varepsilon-\text{regular}} |V_i||V_j| = O(\varepsilon |V|^2) \tag{4.6}$$

  where the sum is over all pairs $1 \leq i \leq j \leq k$ for which G is not ε-regular between V_i and V_j.

While Lemma 4.2.2 is, strictly speaking, weaker than Lemma 4.2.1 in that it does not enforce the equitable size property between the atoms, in practice it seems that the two lemmas are roughly of equal utility; most of the combinatorial consequences of Lemma 4.2.1 can also be proven using Lemma 4.2.2. The point is that one always has to remember to weight each cell V_i by its density $|V_i|/|V|$, rather than by giving each cell an equal weight as in Lemma 4.2.1. Lemma 4.2.2 also has the advantage that one can easily generalise the result from finite vertex sets V to other probability spaces (for instance, one could weight V with something other than the uniform distribution). For applications to hypergraph regularity, it turns out to be slightly more convenient to have *two* partitions (coarse and fine) rather than just one; see for instance [**Ta2006c**]. In any event the arguments below that we give to prove Lemma 4.2.2 can be modified to give a proof of Lemma 4.2.1 also.

The proof of the regularity lemma is usually conducted by a *greedy algorithm*. Very roughly speaking, one starts with the trivial partition of V. If this partition already regularises the graph, we are done; if not, this means that there are some sets A and B in which there is a significant density fluctuation beyond what has already been detected

by the original partition. One then adds these sets to the partition and iterates the argument. Every time a new density fluctuation is incorporated into the partition that models the original graph, this increases a certain "index" or "energy" of the partition. On the other hand, this energy remains bounded no matter how complex the partition, so eventually one must reach a long "energy plateau" in which no further refinement is possible, at which point one can find the regular partition.

One disadvantage of the greedy algorithm is that it is not efficient in the limit $n \to \infty$, as it requires one to search over *all* pairs of subsets A, B of a given pair V_i, V_j of cells, which is an exponentially long search. There are more algorithmically efficient ways to regularise, for instance a polynomial time algorithm was given in [**AlDuLeRoYu1994**]. However, one can do even better, if one is willing to (a) allow cells of unequal size, (b) allow a small probability of failure, (c) have the ability to sample vertices from G at random, and (d) allow for the cells to be defined "implicitly" (via their relationships with a fixed set of reference vertices) rather than "explicitly" (as a list of vertices). In that case, one can regularise a graph in a number of operations *bounded* in n. Indeed, one has

Lemma 4.2.3 (Regularity lemma via random neighbourhoods)**.** *Let $\varepsilon > 0$. Then there exists integers $M_1, \ldots, M_m$ with the following property: whenever $G = (V, E)$ be a graph on finitely many vertices, if one selects one of the integers M_r at random from $M_1, \ldots, M_m$, then selects M_r vertices $v_1, \ldots, v_{M_r} \in V$ uniformly from V at random, then the 2^{M_r} vertex cells $V_1^{M_r}, \ldots, V_{2^{M_r}}^{M_r}$ (some of which can be empty) generated by the vertex neighbourhoods $A_t := \{v \in V : (v, v_t) \in E\}$ for $1 \leq t \leq M_r$, will obey the conclusions of Lemma 4.2.2 with probability at least $1 - O(\varepsilon)$.*

Thus, roughly speaking, one can regularise a graph simply by taking a large number of random vertex neighbourhoods, and using the partition (or Venn diagram) generated by these neighbourhoods as the partition. The intuition is that if there is any non-uniformity in the graph (e.g. if the graph exhibits bipartite behaviour), this will bias the random neighbourhoods to seek out the partitions that would

regularise that non-uniformity (e.g. vertex neighbourhoods would begin to fill out the two vertex cells associated to the bipartite property); if one takes sufficiently many such random neighbourhoods, the probability that all detectable non-uniformity is captured by the partition should converge to 1. (It is more complicated than this, because the finer one makes the partition, the finer the types of non-uniformity one can begin to detect, but this is the basic idea.)

This fact seems to be reasonably well-known folklore, discovered independently by many authors; it is for instance quite close to the graph property testing results in [**AlSh2008**], and also appears in [**Is2006**] and [**Au2008**] (and implicitly in [**Ta2007**]); I will present a proof of the lemma below.

4.2.1. Warmup: a weak regularity lemma. To motivate the idea, let's first prove a weaker but simpler (and more quantitatively effective) regularity lemma, analogous to that established by Frieze and Kannan:

Lemma 4.2.4 (Weak regularity lemma via random neighbourhoods)**.** *Let $\varepsilon > 0$. Then there exists an integer M with the following property: whenever $G = (V, E)$ be a graph on finitely many vertices, if one selects $1 \leq t \leq M$ at random, then selects t vertices $v_1, \ldots, v_t \in V$ uniformly from V at random, then the 2^t vertex cells $V_1^t, \ldots, V_{2^t}^t$ (some of which can be empty) generated by the vertex neighbourhoods $A_{t'} := \{v \in V : (v, v_{t'}) \in E\}$ for $1 \leq t' \leq t$, obey the following property with probability at least $1 - O(\varepsilon)$: for any vertex sets $A, B \subset V$, the number of edges $|E \cap (A \times B)|$ connecting A and B can be approximated by the formula*

$$|E \cap (A \times B)| = \sum_{i=1}^{2^t} \sum_{j=1}^{2^t} d(V_i^t, V_j^t) |A \cap V_i^t| |B \cap V_j^t| + O(\varepsilon |V|^2). \tag{4.7}$$

This weaker lemma only lets us count "macroscopic" edge densities $d(A, B)$, when A, B are dense subsets of V, whereas the full regularity lemma is stronger in that it also controls "microscopic" edge densities $d(A, B)$ where A, B are now dense subsets of the cells $V_i^{M_r}, V_j^{M_r}$. Nevertheless this weaker lemma is easier to prove and already illustrates many of the ideas.

Let's now prove this lemma. Fix $\varepsilon > 0$, let M be chosen later, let $G = (V, E)$ be a graph, and select $v_1, \ldots, v_M$ at random. (There can of course be many vertices selected more than once; this will not bother us.) Let A_t and $V_1^t, \ldots, V_{2^t}^t$ be as in the above lemma. For notational purposes it is more convenient to work with the (random) σ-algebra $\mathcal{B}_t$ generated by the $A_1, \ldots, A_t$ (i.e. the collection of all sets that can be formed from $A_1, \ldots, A_t$ by boolean operations); this is an atomic σ-algebra whose atoms are precisely the (non-empty) cells $V_1^t, \ldots, V_{2^t}^t$ in the partition. Observe that these σ-algebras are nested: $\mathcal{B}_t \subset \mathcal{B}_{t+1}$.

We will use the trick of turning sets into functions, and view the graph as a function $1_E : V \times V \to \mathbf{R}$. One can then form the *conditional expectation* $\mathbf{E}(1_E|\mathcal{B}_t \times \mathcal{B}_t) : V \times V \to \mathbf{R}$ of this function to the product σ-algebra $\mathcal{B}_t \times \mathcal{B}_t$, whose value on $V_i^t \times V_j^t$ is simply the average value of 1_E on the product set $V_i^t \times V_j^t$. (When i and j are different, this is simply the edge density $d(V_i^t, V_j^t)$). One can view $\mathbf{E}(1_E|\mathcal{B}_t \times \mathcal{B}_t)$ more combinatorially, as a weighted graph on V such that all edges between two distinct cells V_i^t, V_j^t have the same constant weight of $d(V_i^t, V_j^t)$.

We give V (and $V \times V$) the uniform probability measure, and define the *energy* e_t at time t to be the (random) quantity

$$e_t := \|\mathbf{E}(1_E|\mathcal{B}_t \times \mathcal{B}_t)\|_{L^2(V\times V)}^2 = \frac{1}{|V|^2} \sum_{v,w\in V} \mathbf{E}(1_E|\mathcal{B}_t \times \mathcal{B}_t)^2.$$

one can interpret this as the mean square of the edge densities $d(V_i^t, V_j^t)$, weighted by the size of the cells V_i^t, V_j^t. From Pythagoras' theorem we have the identity

$$e_{t'} = e_t + \|\mathbf{E}(1_E|\mathcal{B}_{t'} \times \mathcal{B}_{t'}) - \mathbf{E}(1_E|\mathcal{B}_t \times \mathcal{B}_t)\|_{L^2(V\times V)}^2$$

for all $t' > t$; in particular, the e_t are increasing in t. This implies that the expectations $\mathbf{E}e_t$ are also increasing in t. On the other hand, these expectations are bounded between 0 and 1. Thus, if we select $1 \leq t \leq M$ at random, expectation of

$$\mathbf{E}(e_{t+2} - e_t)$$

telescopes to be $O(1/M)$. Thus, by Markov's inequality, with probability $1 - O(\varepsilon)$ we can freeze $v_1, \ldots, v_t$ such that we have the conditional expectation bound

$$\mathbf{E}(e_{t+2} - e_t | v_1, \ldots, v_t) = O(\frac{1}{M\varepsilon}). \tag{4.8}$$

Suppose $v_1, \ldots, v_t$ have this property. We split

$$1_E = f_{U^\perp} + f_U$$

where

$$f_{U^\perp} := \mathbf{E}(1_E | \mathcal{B}_t \times \mathcal{B}_t)$$

and

$$f_U := 1_E - \mathbf{E}(1_E | \mathcal{B}_t \times \mathcal{B}_t).$$

We now assert that the partition $V_1^t, \ldots, V_{2^t}^t$ induced by $\mathcal{B}_t$ obeys the conclusions of Lemma 4.2.3. For this, we observe various properties on the two components of 1_E:

Lemma 4.2.5 ($f_{U^\perp}$ is structured)**.** *$f_{U^\perp}$ is constant on each product set $V_i^t \times V_j^t$.*

Proof. This is clear from construction. □

Lemma 4.2.6 (f_U is pseudorandom)**.** *The expression*

$$\frac{1}{|V|^4} \sum_{v,w,v',w' \in V} f_U(v,w) f_U(v,w') f_U(v',w) f_U(v',w')$$

is of size $O(\frac{1}{\sqrt{M\varepsilon}})$.

Proof. The left-hand side can be rewritten as

$$\mathbf{E} \frac{1}{|V|^2} \sum_{v,w \in V} f_U(v,w) f_U(v,v_{t+2}) f_U(v_{t+1},w) f_U(v_{t+1},v_{t+2}).$$

Observe that the function $(v,w) \mapsto f_U(v,v_{t+2}) f_U(v_{t+1},w) f_U(v,w)$ is measurable with respect to $\mathcal{B}_{t+2} \times \mathcal{B}_{t+2}$, so we can rewrite this expression as

$$\mathbf{E} \frac{1}{|V|^2} \sum_{v,w \in V} \mathbf{E}(f_U | \mathcal{B}_{t+2} \times \mathcal{B}_{t+2})(v,w) f_U(v,v_{t+2}) f_U(v_{t+1},w) f_U(v_{t+1},v_{t+2}).$$

Applying Cauchy-Schwarz, one can bound this by

$$\mathbf{E} \| \mathbf{E}(f_U | \mathcal{B}_{t+2} \times \mathcal{B}_{t+2}) \|_{L^2(V \times V)}.$$

But from Pythagoras we have

$$\mathbf{E}(f_U|\mathcal{B}_{t+2} \times \mathcal{B}_{t+2})^2 = e_{t+2} - e_t$$

and so the claim follows from (4.8) and another application of Cauchy-Schwarz. □

Now we can prove Lemma 4.2.4. Observe that

$$|E \cap (A \times B)| - \sum_{i=1}^{2^t} \sum_{j=1}^{2^t} d(V_i^t, V_j^t)|A \cap V_i^t||B \cap V_j^t|$$

$$= \sum_{v,w \in V} 1_A(v) 1_B(w) f_U(v,w).$$

Applying Cauchy-Schwarz twice in v, w and using Lemma 4.2.6, we see that the RHS is $O((M\varepsilon)^{-1/8})$; choosing $M \gg \varepsilon^{-9}$ we obtain the claim.

4.2.2. Strong regularity via random neighbourhoods. We now prove Lemma 4.2.3, which of course implies Lemma 4.2.2.

Fix $\varepsilon > 0$ and a graph $G = (V, E)$ on n vertices. We randomly select an infinite sequence $v_1, v_2, \ldots \in V$ of vertices in V, drawn uniformly and independently at random. We define $A_t, V_i^t, \mathcal{B}_t$, e_t, as before.

Now let m be a large number depending on $\varepsilon > 0$ to be chosen later, let $F : \mathbf{Z}^+ \to \mathbf{Z}^+$ be a rapidly growing function (also to be chosen later), and set $M_1 := F(1)$ and $M_r := 2(M_{r-1} + F(M_{r-1}))$ for all $1 \leq r \leq m$, thus $M_1 < M_2 < \ldots < M_{m+1}$ grows rapidly to infinity. The expected energies $\mathbf{E}e_{M_r}$ are increasing from 0 to 1, thus if we pick $1 \leq r \leq m$ uniformly at random, the expectation of

$$\mathbf{E}e_{M_{r+1}} - e_{M_r}$$

telescopes to be $O(1/m)$. Thus, by Markov's inequality, with probability $1 - O(\varepsilon)$ we will have

$$\mathbf{E}e_{M_{r+1}} - e_{M_r} = O(\frac{1}{m\varepsilon}).$$

Assume that r is chosen to obey this. Then, by another application of the pigeonhole principle, we can find $M_{r+1}/2 \leq t < M_{r+1}$ such

that

$$\mathbf{E}(e_{t+2} - e_t) = O(\frac{1}{m\varepsilon M_{r+1}}) = O(\frac{1}{m\varepsilon F(M_r)}).$$

Fix this t. We have

$$\mathbf{E}(e_t - e_{M_r}) = O(\frac{1}{m\varepsilon}),$$

so by Markov's inequality, with probability $1 - O(\varepsilon)$, $v_1, \ldots, v_t$ are such that

$$e_t - e_{M_r} = O(\frac{1}{m\varepsilon^2}) \tag{4.9}$$

and also obey the conditional expectation bound

$$\mathbf{E}(e_{t+2} - e_t | v_1, \ldots, v_t) = O(\frac{1}{m\varepsilon F(M_r)}). \tag{4.10}$$

Assume that this is the case. We split

$$1_E = f_{U^\perp} + f_{err} + f_U$$

where

$$f_{U^\perp} := \mathbf{E}(1_E | \mathcal{B}_{M_r} \times \mathcal{B}_{M_r})$$
$$f_{err} := \mathbf{E}(1_E | \mathcal{B}_t \times \mathcal{B}_t) - \mathbf{E}(1_E | \mathcal{B}_{M_r} \times \mathcal{B}_{M_r})$$
$$f_U := 1_E - \mathbf{E}(1_E | \mathcal{B}_t \times \mathcal{B}_t).$$

We now assert that the partition $V_1^{M_r}, \ldots, V_{2^{M_r}}^{M_r}$ induced by $\mathcal{B}_{M_r}$ obeys the conclusions of Lemma 4.2.2. For this, we observe various properties on the three components of 1_E:

Lemma 4.2.7 ($f_{U^\perp}$ locally constant)**.** *$f_{U^\perp}$ is constant on each product set $V_i^{M_r} \times V_j^{M_r}$.*

Proof. This is clear from construction. □

Lemma 4.2.8 (f_{err} small)**.** *We have $\|f_{err}\|_{L^2(V \times V)}^2 = O(\frac{1}{m\varepsilon^2})$.*

Proof. This follows from (4.9) and Pythagoras' theorem. □

Lemma 4.2.9 (f_U uniform)**.** *The expression*

$$\frac{1}{|V|^4} \sum_{v,w,v',w' \in V} f_U(v,w) f_U(v,w') f_U(v',w) f_U(v',w')$$

is of size $O(\frac{1}{\sqrt{m\varepsilon F(M_r)}})$.

Proof. This follows by repeating the proof of Lemma 4.2.6, but using (4.10) instead of (4.8). □

Now we verify the regularity.

First, we eliminate *small atoms*: the pairs (V_i, V_j) for which $|V_i^{M_r}| \leq \varepsilon|V|/2^{M_r}$ clearly give a net contribution of at most $O(\varepsilon|V|^2)$ and are acceptable; similarly for those pairs for which $|V_j^{M_r}| \leq \varepsilon|V|/2^{M_r}$. So we may henceforth assume that

$$|V_i^{M_r}|, |V_j^{M_r}| \leq \varepsilon|V|/2^{M_r}. \tag{4.11}$$

Now, let $A \subset V_i^{M_r}$, $B \subset V_i^{M_r}$ have densities

$$\alpha := |A|/|V_i^{M_r}| \geq \varepsilon; \beta := |B|/|V_j^{M_r}| \geq \varepsilon,$$

then

$$\alpha\beta d(A,B) = \frac{1}{|V_i^{M_r}||V_j^{M_r}|} \sum_{v \in V_i^{M_r}} \sum_{w \in V_i^{M_r}} 1_A(v) 1_B(w) 1_E(v,w).$$

We divide 1_E into the three pieces $f_{U^\perp}$, f_{err}, f_U.

The contribution of $f_{U^\perp}$ is exactly $\alpha\beta d(V_i^{M_r}, V_j^{M_r})$.

The contribution of f_{err} can be bounded using Cauchy-Schwarz as

$$O(\frac{1}{|V_i^{M_r}||V_j^{M_r}|} \sum_{v \in V_i^{M_r}} \sum_{w \in V_i^{M_r}} |f_{err}(v,w)|^2)^{1/2}.$$

Using Lemma 4.2.8 and Chebyshev's inequality, we see that the pairs (V_i, V_j) for which this quantity exceeds ε^3 will contribute at most ε^{-8}/m to (4.6), which is acceptable if we choose m so that $m \gg \varepsilon^{-9}$. Let us now discard these bad pairs.

Finally, the contribution of f_U can be bounded by two applications of Cauchy-Schwarz and (4.2.9) as

$$O(\frac{|V|^2}{|V_i^{M_r}||V_j^{M_r}|} \frac{1}{(m\varepsilon F(M_r))^{1/8}})$$

which by (4.11) is bounded by

$$O(2^{2M_r}\varepsilon^{-2}/(m\varepsilon F(M_r))^{1/8}).$$

This can be made $O(\varepsilon^3)$ by selecting F sufficiently rapidly growing depending on ε. Putting this all together we see that

$$\alpha\beta d(A,B) = \alpha\beta d(V_i^{M_r}, V_j^{M_r}) + O(\varepsilon^3)$$

which (since $\alpha, \beta \geq \varepsilon$) gives the desired regularity.

Remark 4.2.10. Of course, this argument gives tower-exponential bounds (as F is exponential and needs to be iterated m times), which will be familiar to any reader already acquainted with the regularity lemma.

Remark 4.2.11. One can take the partition induced by random neighbourhoods here and carve it up further to be both equitable and (mostly) regular, thus recovering a proof of Lemma 1, by following the arguments in [**Ta2006c**]. Of course, when one does so, one no longer has a partition created purely from random neighbourhoods, but it is pretty clear that one is not going to be able to make an equitable partition just from boolean operations applied to a few random neighbourhoods.

Notes. This article first appeared at `terrytao.wordpress.com/2009/04/26`. Thanks to Anup for corrections.

Asaf Shapira noted that in [**FiMaSh2007**] a similar (though not identical) regularisation algorithm was given which explicitly regularises a graph or hypergraph in linear time.

4.3. Szemerédi's regularity lemma via the correspondence principle

In the previous section, we discussed the *Szemerédi regularity lemma*, and how a given graph could be regularised by partitioning the vertex set into random neighbourhoods. More precisely, we gave a proof of

Lemma 4.3.1 (Regularity lemma via random neighbourhoods)**.** *Let $\varepsilon > 0$. Then there exists integers $M_1, \ldots, M_m$ with the following property: whenever $G = (V, E)$ be a graph on finitely many vertices, if one selects one of the integers M_r at random from $M_1, \ldots, M_m$, then selects M_r vertices $v_1, \ldots, v_{M_r} \in V$ uniformly from V at random, then the 2^{M_r} vertex cells $V_1^{M_r}, \ldots, V_{2^{M_r}}^{M_r}$ (some of which can be empty)*

generated by the vertex neighbourhoods $A_t := \{v \in V : (v, v_t) \in E\}$ *for* $1 \leq t \leq M_r$*, will obey the regularity property*

$$\sum_{(V_i,V_j)\,\text{not}\,\varepsilon-\text{regular}} |V_i||V_j| \leq \varepsilon |V|^2 \tag{4.12}$$

with probability at least $1 - O(\varepsilon)$*, where the sum is over all pairs* $1 \leq i \leq j \leq k$ *for which* G *is not* ε*-regular between* V_i *and* V_j*. [Recall that a pair* (V_i, V_j) *is* ε-regular *for* G *if one has*

$$|d(A, B) - d(V_i, V_j)| \leq \varepsilon$$

for any $A \subset V_i$ *and* $B \subset V_j$ *with* $|A| \geq \varepsilon |V_i|, |B| \geq \varepsilon |V_j|$*, where* $d(A, B) := |E \cap (A \times B)|/|A||B|$ *is the density of edges between* A *and* B*.]*

The proof was a combinatorial one, based on the standard energy increment argument.

In this article I would like to discuss an alternate approach to the regularity lemma, which is an infinitary approach passing through a graph-theoretic version of the Furstenberg correspondence principle. While this approach superficially looks quite different from the combinatorial approach, it in fact uses many of the same ingredients, most notably a reliance on random neighbourhoods to regularise the graph. This approach was introduced in [**Ta2007**], and used in [**Au2008, AuTa2010**] to establish some property testing results for hypergraphs; more recently, a closely related infinitary hypergraph removal lemma developed in [**Ta2007**] was also used in [**Au2009, Au2009b**] to give new proofs of the multidimensional Szemeredi theorem and of the density Hales-Jewett theorem (the latter being a spinoff of the polymath1 project, see Section 4.1).

For various technical reasons we will not be able to use the correspondence principle to recover Lemma 4.3.1 in its full strength; instead, we will establish the following slightly weaker variant.

Lemma 4.3.2 (Regularity lemma via random neighbourhoods, weak version)**.** *Let* $\varepsilon > 0$*. Then there exist an integer* M_* *with the following property: whenever* $G = (V, E)$ *be a graph on finitely many vertices, there exists* $1 \leq M \leq M_*$ *such that if one selects* M *vertices* $v_1, \ldots, v_M \in V$ *uniformly from* V *at random, then the* 2^M *vertex cells*

$V_1^M, \ldots, V_{2^M}^M$ generated by the vertex neighbourhoods $A_t := \{v \in V : (v, v_t) \in E\}$ for $1 \leq t \leq M$, will obey the regularity property (4.12) *with probability at least* $1 - \varepsilon$.

Roughly speaking, Lemma 4.3.1 asserts that one can regularise a large graph G with high probability by using M_r random neighbourhoods, where M_r is chosen at random from one of a number of choices $M_1, \ldots, M_m$; in contrast, the weaker Lemma 4.3.2 asserts that one can regularise a large graph G with high probability by using *some* integer M from $1, \ldots, M_*$, but the exact choice of M depends on G, and it is not guaranteed that a randomly chosen M will be likely to work. While Lemma 4.3.2 is strictly weaker than Lemma 4.3.1, it still implies the (weighted) Szemerédi regularity lemma (Lemma 4.2.2).

4.3.1. The graph correspondence principle. The first key tool in this argument is the *graph correspondence principle*, which takes a sequence of (increasingly large) graphs and uses random sampling to extract an infinitary limit object, which will turn out to be an infinite but random (and, crucially, *exchangeable*) graph. This concept of a graph limit is related to (though slightly different from) the "graphons" used as graph limits in [**LoSz2007**], or the ultraproducts used in [**ElSz2008**]. It also seems to be related to the concept of an elementary limit that I discussed in Section 3.4, though this connection is still rather tentative.

The correspondence works as follows. We start with a finite, deterministic graph $G = (V, E)$. We can then form an infinite, random graph $\hat{G} = (\mathbf{Z}, \hat{E})$ from this graph by the following recipe:

- The vertex set of $\hat{G}$ will be the integers $\mathbf{Z} = \{-2, -1, 0, 1, 2, \ldots\}$.
- For every integer n, we randomly select a vertex v_n in V, uniformly and independently at random. (Note that there will be many collisions, i.e. integers n, m for which $v_n = v_m$, but these collisions will become asymptotically negligible in the limit $|V| \to \infty$.)
- We then define the edge set $\hat{E}$ of $\hat{G}$ by declaring (n, m) to be an edge on $\hat{E}$ if and only if (v_n, v_m) is an edge in E (which in particular requires $v_n \neq v_m$).

More succinctly, $\hat{G}$ is the pullback of G under a random map from $\mathbf{Z}$ to V.

The random graph $\hat{G}$ captures all the "local" information of G, while obscuring all the "global" information. For instance, the edge density of G is essentially just the probability that a given edge, say $(1,2)$, lies in $\hat{G}$. (There is a small error term due to the presence of collisions, but this goes away in the limit $|V| \to \infty$.) Similarly, the triangle density of G is essentially the probability that a given triangle, say $\{(1,2),(2,3),(3,1)\}$, lies in $\hat{G}$. On the other hand, it is difficult to read off global properties of G, such as being connected or 4-colourable, just from $\hat{G}$.

At first glance, it may seem a poor bargain to trade in a finite deterministic graph G for an infinite random graph $\hat{G}$, which is a more complicated and less elementary object. However, there are three major advantages of working with $\hat{G}$ rather than G:

- **Exchangeability**. The probability distribution of $\hat{G}$ has a powerful symmetry or *exchangeability* property: if one takes the random graph $\hat{G}$ and interchanges any two vertices in $\mathbf{Z}$, e.g. 3 and 5, one obtains a new graph which is not equal to $\hat{G}$, but nevertheless has the same probability distribution as $\hat{G}$, basically because the v_n were selected in an iid (independent and identically distributed) manner. More generally, given any permutation $\sigma : \mathbf{Z} \to \mathbf{Z}$, the pullback $\sigma^*(\hat{G})$ of $\hat{G}$ by σ has the same probability distribution as $\hat{G}$; thus we have a measure-preserving action of the symmetric group S_∞, which places us in the general framework of ergodic theory.
- **Limits**. The space of probability measures on the space $2^{\binom{\mathbf{Z}}{2}}$ of infinite graphs is sequentially compact; given any sequence $\hat{G}_n = (\mathbf{Z}, \hat{E}_n)$ of infinite random graphs, one can find a subsequence $\hat{G}_{n_j}$ which converges in the *vague topology* to another infinite random graph. What this means is that given any event E on infinite graphs that involve only finitely many of the edges, the probability that $\hat{G}_{n_j}$ obeys E converges to the probability that $\hat{G}$ obeys E. (Thus, for instance, the probability that $\hat{G}_{n_j}$ contains the triangle

$\{(1,2),(2,3),(3,1)\}$ will converge to the probability that $\hat{G}$ contains the same triangle.) Note that properties that involve infinitely many edges (e.g. connectedness) need not be preserved under vague limits.

- **Factors**. The underlying probability space for the random variable $\hat{G}$ is the space $2^{\binom{\mathbf{Z}}{2}}$ of infinite graphs, and it is natural to give this space the Borel σ-algebra $\mathcal{B}_{\mathbf{Z}}$, which is the σ-algebra generated by the *cylinder events* "$(i,j) \in \hat{G}$" for $i,j \in \mathbf{Z}$. But this σ-algebra also has a number of useful sub-σ-algebras or *factors*, representing various partial information on the graph $\hat{G}$. In particular, given any subset I of $\mathbf{Z}$, one can create the factor $\mathcal{B}_I$, defined as the σ-algebra generated by the events "$(i,j) \in \hat{G}$" for $i,j \in I$. Thus for instance, the event that $\hat{G}$ contains the triangle is measurable in $\mathcal{B}_{\{1,2,3\}}$, but not in $\mathcal{B}_{\{1,2\}}$. One can also look at compound factors such as $\mathcal{B}_I \wedge \mathcal{B}_J$, the factor generated by the union of $\mathcal{B}_I$ and $\mathcal{B}_J$. For instance, the event that $\hat{G}$ contains the edges $(1,2),(1,3)$ is measurable in $\mathcal{B}_{\{1,2\}} \vee \mathcal{B}_{\{1,3\}}$, but the event that $\hat{G}$ contains the triangle $\{(1,2),(2,3),(3,1)\}$ is not.

The connection between the infinite random graph $\hat{G}$ and partitioning through random neighbourhoods comes when contemplating the relative difference between a factor such as $\mathcal{B}_{\{-n,\ldots,-1\}}$ and $\mathcal{B}_{\{-n,\ldots,-1\}\cup\{1\}}$ (say). The latter factor is generated by the former factor, together with the events "$(1,-i) \in \hat{E}$" for $i = 1,\ldots,n$. But observe if $\hat{G} = (\mathbf{Z}, \hat{E})$ is generated from a finite deterministic graph $G = (V,E)$, then $(1,-i)$ lies in $\hat{E}$ if and only if v_1 lies in the vertex neighbourhood of v_{-i}. Thus, if one uses the vertex neighbourhoods of $v_{-1},\ldots,v_{-n}$ to subdivide the original vertex set V into 2^n cells of varying sizes, the factor $\mathcal{B}_{\{-n,\ldots,-1\}\cup\{1\}}$ is generated from $\mathcal{B}_{\{-n,\ldots,-1\}}$, together with the random variable that computes which of these 2^n cells the random vertex v_1 falls into. We will see this connection in more detail later in this post, when we use the correspondence principle to prove Lemma 4.3.2.

Combining the exchangeability and limit properties (and noting that the vague limit of exchangeable random graphs is still exchangeable), we obtain

Lemma 4.3.3 (Graph correspondence principle)**.** *Let $G_n = (V_n, E_n)$ be a sequence of finite deterministic graphs, and let $\hat{G}_n = (\mathbf{Z}, \hat{E}_n)$ be their infinite random counterparts. Then there exists a subsequence n_j such that $\hat{G}_{n_j}$ converges in the vague topology to an exchangeable infinite random graph $\hat{G} = (\mathbf{Z}, \hat{E})$.*

We can illustrate this principle with three main examples, two from opposing extremes of the "dichotomy between structure and randomness", and one intermediate one.

Example 4.3.4 (Random example)**.** Let $G_n = (V_n, E_n)$ be a sequence of ε_n-regular graphs of edge density p_n, where $|V_n| \to \infty$, $\varepsilon_n \to 0$, and $p_n \to p$ as $n \to \infty$. Then any graph limit $\hat{G} = (\mathbf{Z}, \hat{G})$ of this sequence will be an *Erdös-Rényi graph* $\hat{G} = G(\infty, p)$, where each edge (i, j) lies in $\hat{G}$ with an independent probability of p.

Example 4.3.5 (Structured example)**.** Let $G_n = (V_n, E_n)$ be a sequence of complete bipartite graphs, where the two cells of the bipartite graph have vertex density q_n and $1 - q_n$ respectively, with $|V_n| \to \infty$ and $q_n \to q$. Then any graph limit $\hat{G} = (\mathbf{Z}, \hat{E})$ of this sequence will be a random complete bipartite graph, constructed as follows: first, randomly colour each vertex n of $\mathbf{Z}$ red with probability q and blue with probability $1 - q$, independently for each vertex. Then define $\hat{G}$ to be the complete bipartite graph between the red vertices and the blue vertices.

Example 4.3.6 (Random+structured example)**.** Let $G_n = (V_n, E_n)$ be a sequence of *incomplete* bipartite graphs, where the two cells of the bipartite graph have vertex density p_n and $1-p_n$ respectively, and the graph G_n is ε_n-regular between these two cells with edge density p_n, with $|V_n| \to \infty$, $\varepsilon_n \to 0$, $p_n \to p$, and $q_n \to q$. Then any graph limit $\hat{G} = (\mathbf{Z}, \hat{E})$ of this sequence will be a random bipartite graph, constructed as follows: first, randomly colour each vertex n of $\mathbf{Z}$ red with probability q and blue with probability $1 - q$, independently for each vertex. Then define $\hat{G}$ to be the bipartite graph between the red

vertices and the blue vertices, with each edge between red and blue having an independent probability of p of lying in $\hat{E}$.

One can use the graph correspondence principle to prove statements about finite deterministic graphs, by the usual *compactness and contradiction approach*: argue by contradiction, create a sequence of finite deterministic graph counterexamples, use the correspondence principle to pass to an infinite random exchangeable limit, and obtain the desired contradiction in the infinitary setting. This will be how we shall approach the proof of Lemma 4.3.2.

4.3.2. The infinitary regularity lemma. To prove the finitary regularity lemma via the correspondence principle, one must first develop an infinitary counterpart. We will present this infinitary regularity lemma (first introduced in this paper) shortly, but let us motivate it by a discussion based on the three model examples of infinite exchangeable graphs $\hat{G} = (\mathbf{Z}, \hat{E})$ from the previous section.

First, consider the "random" graph $\hat{G}$ from Example 4.3.4. Here, we observe that the events "$(i,j) \in \hat{E}$" are jointly independent of each other, thus for instance

$$\mathbf{P}((1,2),(2,3),(3,1) \in \hat{E}) = \prod_{(i,j)=(1,2),(2,3),(3,1)} \mathbf{P}((i,j) \in \hat{E}).$$

More generally, we see that the factors $\mathcal{B}_{\{i,j\}}$ for all distinct $i,j \in \mathbf{Z}$ are independent, which means that

$$\mathbf{P}(E_1 \wedge \ldots \wedge E_n) = \mathbf{P}(E_1) \ldots \mathbf{P}(E_n)$$

whenever $E_1 \in \mathcal{B}_{\{i_1,j_1\}}, \ldots, E_n \in \mathcal{B}_{\{i_n,j_n\}}$ and the $\{i_1,j_1\}, \ldots, \{i_n,j_n\}$ are distinct.

Next, we consider the "structured" graph $\hat{G}$ from Example 4.3.5, where we take $0 < p < 1$ to avoid degeneracies. In contrast to the preceding example, the events "$(i,j) \in \hat{E}$" are now highly dependent; for instance, if $(1,2) \in \hat{E}$ and $(1,3) \in \hat{E}$, then this forces $(2,3)$ to lie outside of $\hat{E}$, despite the fact that the events "$(i,j) \in \hat{E}$" each occur with a non-zero probability of $p(1-p)$. In particular, the factors $\mathcal{B}_{\{1,2\}}, \mathcal{B}_{\{1,3\}}, \mathcal{B}_{\{2,3\}}$ are not jointly independent.

However, one can recover a *conditional* independence by introducing some new factors. Specifically, let $\mathcal{B}_i$ be the factor generated

by the event that the vertex i is coloured red. Then we see that the factors $\mathcal{B}_{\{1,2\}}, \mathcal{B}_{\{1,3\}}, \mathcal{B}_{\{2,3\}}$ now become *conditionally* jointly independent, relative to the base factor $\mathcal{B}_1 \vee \mathcal{B}_2 \vee \mathcal{B}_3$, which means that we have conditional independence identities such as

$$\mathbf{P}((1,2),(2,3),(3,1) \in \hat{E} | \mathcal{B}_1 \vee \mathcal{B}_2 \vee \mathcal{B}_3) = \prod_{(i,j)=(1,2),(2,3),(3,1)} \mathbf{P}((i,j) \in \hat{E} | \mathcal{B}_1 \vee \mathcal{B}_2 \vee \mathcal{B}_3).$$

Indeed, once one fixes (conditions) the information in $\mathcal{B}_1 \vee \mathcal{B}_2 \vee \mathcal{B}_3$ (i.e. once one knows what colour the vertices $1, 2, 3$ are), the events "$(i,j) \in \hat{E}$" for $(i,j) = (1,2), (2,3), (3,1)$ either occur with probability 1 (if i, j have distinct colours) or probability 0 (if i, j have the same colour), and so the conditional independence is trivially true.

A similar phenomenon holds for the "random+structured" graph $\hat{G}$ from Example 4.3.6, with $0 < p, q < 1$. Again, the factors $\mathcal{B}_{\{i,j\}}$ are not jointly independent in an absolute sense, but once one introduces the factors $\mathcal{B}_i$ based on the colour of the vertex i, we see once again that the $\mathcal{B}_{\{i,j\}}$ become conditionally jointly independent relative to the $\mathcal{B}_i$.

These examples suggest, more generally, that we should be able to *regularise* the graph $\hat{G}$ (or more precisely, the system of edge factors $\mathcal{B}_{\{i,j\}}$) by introducing some single-vertex factors $\mathcal{B}_i$, with respect to which the edge factors become conditionally independent; this is the infinitary analogue of a finite graph becoming ε-regular relative to a suitably chosen partition of the vertex set into cells.

Now, in Examples 4.3.5, 4.3.6 we were able to obtain this regularisation because the vertices of the graph were conveniently coloured for us (red or blue). But for general infinite exchangeable graphs $\hat{G}$, such a vertex colouring is not provided to us, so how is one to generate the vertex factors $\mathcal{B}_i$?

The key trick - which is the infinitary analogue of using random neighbourhoods to regularise a finitary graph - is to sequester half of the infinite vertices in $\mathbf{Z}$ - e.g. the negative vertices $-1, -2, \ldots$ - away as "reference" or "training" vertices, and then and colorise the remaining vertices i of the graph based on how that vertex interacts with the reference vertices. More formally, we define $\mathcal{B}_i$ for

$i = 0, 1, 2, \ldots$ by the formula

$$\mathcal{B}_i := \mathcal{B}_{\{-1,-2,\ldots\}\cup\{i\}}.$$

We then have

Lemma 4.3.7 (Infinitary regularity lemma)**.** *Let $\hat{G} = (\mathbf{Z}, \hat{E})$ be a infinite exchangeable random graph. Then the $\mathcal{B}_{\{i,j\}} \vee \mathcal{B}_i \vee \mathcal{B}_j$ for natural numbers i, j are conditionally jointly independent relative to the $\mathcal{B}_i$. More precisely, if I is a set of natural numbers, E is a subset of $\binom{I}{2}$, and E_e is a $\mathcal{B}_e \wedge \bigwedge_{i\in e} \mathcal{B}_i$-measurable event for all $e \in E$, then*

$$\mathbf{P}(\bigwedge_{e\in E} E_e | \bigwedge_{i\in I} \mathcal{B}_i) = \prod_{e\in E} \mathbf{P}(E_e | \bigwedge_{i\in I} \mathcal{B}_i).$$

Proof. By induction on E, it suffices to show that for any $e_0 \in E$, the event E_{e_0} and the event $\bigwedge_{e\in E\setminus\{e_0\}} E_e$ are independent relative to $\bigwedge_{i\in I} \mathcal{B}_i$.

By relabeling we may take $I = \{1, \ldots, n\}$ and $e_0 = \{1, 2\}$ for some $n \geq 2$. We use the exchangeability of $\hat{G}$ (and *Hilbert's hotel*) to observe that the random variables

$$\mathbf{E}(1_{E_{e_0}} | \mathcal{B}_{\{-1,-2,\ldots\}\cup\{1\}} \vee \mathcal{B}_{\{-1,-2,\ldots\}\cup\{2\}})$$

and

$$\mathbf{E}(1_{E_{e_0}} | \mathcal{B}_{\{-1,-2,\ldots\}\cup\{1\}\cup\{3,\ldots,n\}} \vee \mathcal{B}_{\{-1,-2,\ldots\}\cup\{2\}\cup\{3,\ldots,n\}})$$

have the same distribution; in particular, they have the same L^2 norm. By Pythagoras' theorem, they must therefore be equal almost surely; furthermore, for any intermediate σ-algebra $\mathcal{B}$ between $\mathcal{B}_{\{-1,-2,\ldots\}\cup\{1\}} \vee \mathcal{B}_{\{-1,-2,\ldots\}\cup\{2\}}$ and $\mathcal{B}_{\{-1,-2,\ldots\}\cup\{1\}\cup\{3,\ldots,n\}} \vee \mathcal{B}_{\{-1,-2,\ldots\}\cup\{2\}\cup\{3,\ldots,n\}}$, $\mathbf{E}(1_{E_{e_0}} | \mathcal{B})$ is also equal almost surely to the above two expressions. (The astute reader will observe that we have just run the "energy increment argument"; in the infinitary world, it is somewhat slicker than in the finitary world, due to the convenience of the Hilbert's hotel trick, and the fact that the existence of orthogonal projections (and in particular, conditional expectation) is itself encoding an energy increment argument.)

As a special case of the above observation, we see that

$$\mathbf{E}(1_{E_{e_0}} | \bigwedge_{i\in I} \mathcal{B}_i) = \mathbf{E}(1_{E_{e_0}} | \bigwedge_{i\in I} \mathcal{B}_i \wedge \bigwedge_{e\in E\setminus\{e_0\}} \mathcal{B}_e).$$

In particular, this implies that E_0 is conditionally independent of every event measurable in $\bigwedge_{i\in I} \mathcal{B}_i \wedge \bigwedge_{e\in E\backslash\{e_0\}} \mathcal{B}_e$, relative to $\bigwedge_{i\in I} \mathcal{B}_i$, and the claim follows. □

Remark 4.3.8. The same argument also allows one to easily regularise infinite exchangeable hypergraphs; see [**Ta2007**]. In fact one can go further and obtain a structural theorem for these hypergraphs generalising *de Finetti's theorem*, and also closely related to the graphons of Lovasz and Szegedy; see [**Au2008**] for details.

4.3.3. Proof of finitary regularity lemma. Having proven the infinitary regularity lemma, we now use the correspondence principle and the compactness and contradiction argument to recover the finitary regularity lemma, Lemma 4.3.2.

Suppose this lemma failed. Carefully negating all the quantifiers, this means that there exists $\varepsilon > 0$, a sequence M_n going to infinity, and a sequence of finite deterministic graphs $G_n = (V_n, E_n)$ such that for every $1 \leq M \leq M_n$, if one selects vertices $v_1, \ldots, v_M \in V_n$ uniformly from V_n, then the 2^M vertex cells $V_1^M, \ldots, V_{2^M}^M$ generated by the vertex neighbourhoods $A_t := \{v \in V : (v, v_t) \in E\}$ for $1 \leq t \leq M$, will obey the regularity property (4.12) with probability less than $1 - \varepsilon$.

We convert each of the finite deterministic graphs $G_n = (V_n, E_n)$ to an infinite random exchangeable graph $\hat{G}_n = (\mathbf{Z}, \hat{E}_n)$; invoking the ocrrespondence principle and passing to a subsequence if necessary, we can assume that this graph converges in the vague topology to an exchangeable limit $\hat{G} = (\mathbf{Z}, \hat{E})$. Applying the infinitary regularity lemma to this graph, we see that the edge factors $\mathcal{B}_{\{i,j\}} \wedge \mathcal{B}_i \wedge \mathcal{B}_j$ for natural numbers i, j are conditionally jointly independent relative to the vertex factors $\mathcal{B}_i$.

Now for any distinct natural numbers i, j, let $f(i, j)$ be the indicator of the event "(i, j) lies in $\hat{E}$", thus $f = 1$ when (i, j) lies in $\hat{E}$ and $f(i, j) = 0$ otherwise. Clearly $f(i, j)$ is $\mathcal{B}_{\{i,j\}}$-measurable. We can write

$$f(i, j) = f_{U^\perp}(i, j) + f_U(i, j)$$

where

$$f_{U^\perp}(i, j) := \mathbf{E}(f(i, j)|\mathcal{B}_i \wedge \mathcal{B}_j)$$

and

$$f_U(i,j) := f(i,j) - f_{U^\perp}(i,j).$$

The exchangeability of $\hat{G}$ ensures that $f, f_U, f_{U^\perp}$ are exchangeable with respect to permutations of the natural numbers, in particular $f_U(i,j) = f_U(j,i)$ and $f_{U^\perp}(i,j) = f_{U^\perp}(j,i)$.

By the infinitary regularity lemma, the $f_U(i,j)$ are jointly independent relative to the $\mathcal{B}_i$, and also have mean zero relative to these factors, so in particular they are infinitely pseudorandom in the sense that

$$\mathbf{E} f_U(1,2) f_U(3,2) f_U(1,4) f_U(3,4) = 0.$$

Meanwhile, the random variable $f_{U^\perp}(1,2)$ is measurable with respect to the factor $\mathcal{B}_1 \vee \mathcal{B}_2$, which is the limit of the factors $\mathcal{B}_{\{-1,-2,\ldots,-M\}\cup\{1\}} \vee \mathcal{B}_{\{-1,-2,\ldots,-M\}\cup\{2\}}$ as M increases. Thus, given any $\tilde{\varepsilon} > 0$ (to be chosen later), one can find an approximation $\tilde{f}_{U^\perp}(1,2)$ to $f_{U^\perp}(1,2)$, bounded between 0 and 1, which is $\mathcal{B}_{\{-1,-2,\ldots,-M\}\cup\{1\}} \vee \mathcal{B}_{\{-1,-2,\ldots,-M\}\cup\{2\}}$-measurable for some M, and such that

$$\mathbf{E}|\tilde{f}_{U^\perp}(1,2) - f_{U^\perp}(1,2)| \leq \tilde{\varepsilon}.$$

We can also impose the symmetry condition $\tilde{f}_{U^\perp}(1,2) = \tilde{f}_{U^\perp}(2,1)$. Now let $\tilde{\varepsilon}' > 0$ be an extremely small number (depending on $\tilde{\varepsilon}, n$) to be chosen later. Then one can find an approximation $\tilde{f}_U(1,2)$ to $f_U(1,2)$, bounded between -1 and 1, which is $\mathcal{B}_{\{-1,-2,\ldots,-M'\}\cup\{1\}} \vee \mathcal{B}_{\{-1,-2,\ldots,-M'\}\cup\{2\}}$-measurable for some M', and such that

$$\mathbf{E}|\tilde{f}_U(1,2) - f_U(1,2)| \leq \tilde{\varepsilon}'.$$

Again we can impose the symmetry condition $\tilde{f}_U(1,2) = \tilde{f}_U(2,1)$. We can then extend $\tilde{f}_U$ by exchangeability, so that

$$\mathbf{E}|\tilde{f}_U(i,j) - f_U(i,j)| \leq \tilde{\varepsilon}'.$$

for all distinct natural numbers i, j. By the triangle inequality we then have

$$\mathbf{E}\tilde{f}_U(1,2)\tilde{f}_U(3,2)\tilde{f}_U(1,4)\tilde{f}_U(3,4) = O(\tilde{\varepsilon}') \tag{4.13}$$

and by a separate application of the triangle inequality

$$\mathbf{E}|f(i,j) - \tilde{f}_{U^\perp}(i,j) - \tilde{f}_U(i,j)| = O(\tilde{\varepsilon}). \tag{4.14}$$

The bounds (4.13), (4.14) apply to the limiting infinite random graph $\hat{G} = (\mathbf{Z}, \hat{E})$. On the other hand, all the random variables appearing in (4.13), (4.14) involve at most finitely many of the edges of the graph. Thus, by vague convergence, the bounds (4.13), (4.14) also apply to the graph $\hat{G}_n = (\mathbf{Z}, \hat{E}_n)$ for sufficiently large n.

Now we unwind the definitions to move back to the finite graphs $G_n = (V_n, E_n)$. Observe that, when applied to the graph $\hat{G}_n$, one has

$$\tilde{f}_{U^\perp}(1,2) = F_{U^\perp,n}(v_1, v_2)$$

where $F_{U,n} : V_n \times V_n \to [0,1]$ is a symmetric function which is constant on the pairs of cells $V_1^M, \ldots, V_{2^M}^M$ generated the vertex neighbourhoods of $v_{-1}, \ldots, v_{-M}$. Similarly,

$$\tilde{f}_U(1,2) = F_{U,n}(v_1, v_2)$$

for some symmetric function $F_{U,n} : V_n \times V_n \to [-1,1]$. The estimate (4.13) can then be converted to a uniformity estimate on $F_{U,n}$

$$\mathbf{E} F_{U,n}(v_1,v_2) F_{U,n}(v_3,v_2) F_{U,n}(v_1,v_4) F_{U,n}(v_3,v_4) = O(\tilde{\varepsilon}')$$

while the estimate (4.14) can be similarly converted to

$$\mathbf{E}|1_{E_n}(v_1,v_2) - F_{U^\perp,n}(v_1,v_2) - F_{U,n}(v_1,v_2)| = O(\tilde{\varepsilon}).$$

If one then repeats the arguments in the preceding blog post, we conclude (if $\tilde{\varepsilon}$ is sufficiently small depending on ε, and $\tilde{\varepsilon}'$ is sufficiently small depending on ε, $\tilde{e}ps$, M) that for $1-\varepsilon$ of the choices for $v_{-1}, \ldots, v_{-M}$, the partition $V_1^M, \ldots, V_{2^M}^M$ induced by the corresponding vertex neighbourhoods will obey (4.12). But this contradicts the construction of the G_n, and the claim follows.

Notes. This article first appeared at `terrytao.wordpress.com/2009/05/08`.

4.4. The two-ends reduction for the Kakeya maximal conjecture

In this articleI would like to make some technical notes on a standard reduction used in the (Euclidean, maximal) Kakeya problem, known as the *two ends reduction*. This reduction (which takes advantage of the approximate scale-invariance of the Kakeya problem) was introduced by Wolff[**Wo1995**], and has since been used many times,

both for the Kakeya problem and in other similar problems (e.g. in [**TaWr2003**] to study curved Radon-like transforms). I was asked about it recently, so I thought I would describe the trick here. As an application I give a proof of the $d = \frac{n+1}{2}$ case of the Kakeya maximal conjecture.

The *Kakeya maximal function conjecture* in $\mathbf{R}^n$ can be formulated as follows:

Conjecture 4.4.1 (Kakeya maximal function conjecture)**.** *If $0 < \delta < 1$, $1 \leq d \leq n$, and $T_1, \ldots, T_N$ is a collection of $\delta \times 1$ tubes oriented in a δ-separated set of directions, then*

$$\| \sum_{i=1}^N 1_{T_i} \|_{L^{d/(d-1)}(\mathbf{R}^n)} \ll_\varepsilon (\frac{1}{\delta})^{\frac{n}{d}-1+\varepsilon} \tag{4.15}$$

for any $\varepsilon > 0$.

A standard duality argument shows that (4.15) is equivalent to the estimate

$$\sum_{i=1}^N \int_{T_i} F \ll_\varepsilon (\frac{1}{\delta})^{\frac{n}{d}-1+\varepsilon} \|F\|_{L^d(\mathbf{R}^n)}$$

for arbitrary non-negative measurable functions F; breaking F up into level sets via dyadic decomposition, this estimate is in turn equivalent to the estimate

$$\sum_{i=1}^N |E \cap T_i| \ll_\varepsilon (\frac{1}{\delta})^{\frac{n}{d}-1+\varepsilon} |E|^{1/d} \tag{4.16}$$

for arbitrary measurable sets E. This estimate is then equivalent to the following:

Conjecture 4.4.2 (Kakeya maximal function conjecture, second version)**.** *If $0 < \delta, \lambda < 1$, $1 \leq d \leq n$, $T_1, \ldots, T_N$ is a collection of $\delta \times 1$ tubes oriented in a δ-separated set of directions, and E is a measurable set such that $|E \cap T_i| \geq \lambda |T_i|$ for all i, then*

$$|E| \gg_\varepsilon (N\delta^{n-1})\lambda^d \delta^{n-d+\varepsilon}$$

for all $\varepsilon > 0$.

Indeed, to deduce (4.16) from Conjecture 4.4.2 one can perform another dyadic decomposition, this time based on the dyadic range of

the densities $|E \cap T_i|/|T_i|$. Conversely, (4.16) implies Conjecture 4.4.2 in the case $N\delta^{n-1} \sim 1$, and the remaining case $N\delta^{n-1} \ll 1$ can then be deduced by the *random rotations trick* (see e.g. [**ElObTa2009**]).

We can reformulate the conjecture again slightly:

Conjecture 4.4.3 (Kakeya maximal function conjecture, third version)**.** *Let $0 < \delta, \lambda < 1$, $1 \leq d \leq n$, and $T_1, \ldots, T_N$ is a collection of $\delta \times 1$ tubes oriented in a δ-separated set of directions with $N \sim \delta^{1-n}$. For each $1 \leq i \leq N$, let $E_i \subset T_i$ be a set with $|E_i| \geq \lambda |T_i|$. Then*

$$|\bigcup_{i=1}^{N} E_i| \gg_\varepsilon \lambda^d \delta^{n-d+\varepsilon}$$

for all $\varepsilon > 0$.

We remark that (the Minkowski dimension version of) the Kakeya set conjecture essentially corresponds to the $\lambda = 1$ case of Conjecture 4.4.3, while the Hausdorff dimension can be shown to be implied by the case where $\lambda \gg \frac{1}{\log^2 1/\delta}$ (actually any lower bound here which is dyadically summable in δ would suffice). Thus, while the Kakeya set conjecture is concerned with how small one can make unions of tubes T_i, the Kakeya maximal function conjecture is concerned with how small one can make unions of *portions* E_i of tubes T_i, where the density λ of the tubes are fixed.

A key technical problem in the Euclidean setting (which is not present in the finite field case), is that the portions E_i of T_i may be concentrated in only a small portion of the tube, e.g. they could fill up a $\delta \times \lambda$ subtube, rather than being dispersed uniformly throughout the tube. Because of this, the set $\bigcup_{i=1}^N E_i$ could be crammed into a far tighter space than one would ideally like. Fortunately, the *two ends reduction* allows one to eliminate this possibility, letting one only consider portions E_i which are not concentrated on just one end of the tube or another, but occupy both ends of the tube in some sense. A more precise version of this is as follows.

Definition 4.4.4 (Two ends condition)**.** Let E be a subset of $\mathbf{R}^n$, and let $\varepsilon > 0$. We say that E obeys the *two ends condition* with exponent ε if one has the bound

$$|E \cap B(x,r)| \ll_\varepsilon r^\varepsilon |E|$$

for all balls $B(x,r)$ in $\mathbf{R}^n$ (note that the bound is only nontrivial when $r \ll 1$).

Informally, the two ends condition asserts that E cannot concentrate in a small ball; it implies for instance that the diameter of E is $\gg_\varepsilon 1$.

We now have

Proposition 4.4.5 (Two ends reduction)**.** *To prove Conjecture 4.4.3 for a fixed value of d and n, it suffices to prove it under the assumption that the sets E_i all obey the two ends condition with exponent ε, for any fixed value of $\varepsilon > 0$.*

The key tool used to prove this proposition is

Lemma 4.4.6 (Every set has a large rescaled two-ends piece)**.** *Let $E \subset \mathbf{R}^n$ be a set of positive measure and diameter $O(1)$, and let $0 < \varepsilon < n$. Then there exists a ball $B(x,r)$ of radius $r = O(1)$ such that*

$$|E \cap B(x,r)| \gg r^\varepsilon |E|$$

and

$$|E \cap B(x',r')| \ll (r'/r)^\varepsilon |E \cap B(x,r)|$$

for all other balls $B(x',r')$.

Proof. Consider the problem of maximising the quantity $|E \cap B(x,r)|/r^\varepsilon$ among al balls $B(x,r)$ of radius at most the diameter of E. On the one hand, this quantity can be at least $\gg |E|$, simply by taking $B(x,r)$ equal to the smallest ball containing E. On the other hand, using the trivial bound $|E \cap B(x,r)| \leq |B(x,r)| \ll r^n$ we see that the quantity $|E \cap B(x,r)|/r^\varepsilon$ is bounded. Thus the supremum of the $|E \cap B(x,r)|/r^\varepsilon$ is finite. If we pick a ball $B(x,r)$ which comes within a factor of 2 (say) of realising this supremum then the claim easily follows. (Actually one can even attain the supremum exactly by a compactness argument, though this is not necessary for our applications.) □

One can view the quantity r in the above lemma as describing the "width" of the set E; this is the viewpoint taken for instance in [**TaWr2003**].

Now we prove Proposition 4.4.5.

Proof. Suppose Conjecture 4.4.3 has already been proven (assuming the two ends condition with exponent ε) for some value of d, n, and some small value of ε. Now suppose we have the setup of Conjecture 4.4.3 without the two-ends condition.

The first observation is that the claim is easy when $\lambda \ll \delta$. Indeed, in this case we can just bound $|\bigcup_{i=1}^N E_i|$ from below the volume $\lambda|T_i| \sim \lambda\delta^{n-1}$ of a single tube. So we may assume that λ is much greater than δ.

Let $\varepsilon > 0$ be arbitrary. We apply Lemma 4.4.6 to each E_i, to find a ball $B(x_i, r_i)$ such that

$$|E_i \cap B(x_i, r_i)| \gg r_i^\varepsilon |E_i| \tag{4.17}$$

and

$$|E_i \cap B(x', r')| \ll (r'/r_i)^\varepsilon |E_i \cap B(x_i, r_i)|$$

for all $B(x', r')$. From (4.17) and the fact that $|E_i| = \lambda|T_i| \gg \lambda\delta^{n-1} \gg \delta^n$, as well as the trivial bound $|E_i \cap B(x_i, r_i)| \leq |B(x_i, r_i)| \ll r_i^n$, we obtain the lower bound $r_i \gg \delta^{1+O(\varepsilon)}$. Thus there are only about $O(\log \frac{1}{\delta})$ possible dyadic ranges $\rho \leq r_i \leq 2\rho$. Using the pigeonhole principle (refining the number N of tubes by a factor of $\log \frac{1}{\delta}$), we may assume that there is a single $\delta^{1+O(\varepsilon)} \leq \rho \ll 1$ such that *all* of the r_i lie in the same dyadic range $[\rho, 2\rho]$.

The intersection of T_i with $B(x_i, r_i)$ is then contained in a $\delta \times O(\rho)$ tube $\tilde{T}_i$, and $\tilde{E}_i := E_i \cap \tilde{T}_i$ occupies a fraction

$$|\tilde{E}_i|/|\tilde{T}_i| \gg r_i^\varepsilon |E_i|/|\tilde{T}_i| \gg \delta^{O(\varepsilon)} \lambda/\rho$$

of $\tilde{T}_i$. If we then rescale each of the $\tilde{E}_i$ and $\tilde{T}_i$ by $O(1/\rho)$, we can locate subsets E_i' of $O(\delta/\rho) \times 1$-tubes T_i' of density $\gg \delta^{O(\varepsilon)}\lambda/\rho$. These tubes T_i' have cardinality $\delta^{1-n+O(\varepsilon)}$ (the loss here is due to the use of the pigeonhole principle earlier) and occupy a δ-separated set of directions, but after refining these tubes a bit we may assume that they instead occupy a δ/ρ-separated set of directions, at the expense of cutting the cardinality down to $\delta^{O(\varepsilon)}(\delta/\rho)^{1-n}$ or so. Furthermore, by construction the E_i' obey the two-ends condition at exponent ε. Applying the hypothesis that Conjecture 4.4.3 holds for such sets, we

conclude that

$$|\bigcup_i E_i'| \gg_\varepsilon \delta^{O(\varepsilon)}[\lambda/\rho]^d[\delta/\rho]^{n-d},$$

which on undoing the rescaling by $1/\rho$ gives

$$|\bigcup_i \tilde{E}_i| \gg_\varepsilon \delta^{O(\varepsilon)}\lambda^d\delta^{n-d}.$$

Since $\varepsilon > 0$ was arbitrary, the claim follows. □

To give an idea of how this two-ends reduction is used, we give a quick application of it:

Proposition 4.4.7. *The Kakeya maximal function conjecture is true for $d \leq \frac{n+1}{2}$.*

Proof. We use the "bush" argument of Bourgain. By the above reductions, it suffices to establish the bound

$$|\bigcup_{i=1}^N E_i| \gg_\varepsilon \lambda^{\frac{n+1}{2}}\delta^{\frac{n-1}{2}-\varepsilon}$$

whenever $N \sim \delta^{1-n}$, and $E_i \subset T_i$ are subsets of $\delta \times 1$ tubes T_i in δ-separated directions with density λ and obeying the two-ends condition with exponent ε.

Let μ be the maximum multiplicity of the E_i, i.e. $\mu := \|\sum_{i=1}^N 1_{E_i}\|_{L^\infty(\mathbf{R}^n)}$. On the one hand, we clearly have

$$|\bigcup_{i=1}^N E_i| \geq \frac{1}{\mu}\|\sum_{i=1}^N 1_{E_i}\|_{L^1(\mathbf{R}^n)} \gg \frac{1}{\mu}\lambda N\delta^{n-1} \gg \frac{\lambda}{\mu}.$$

This bound is good when μ is small. What if μ is large? Then there exists a point x_0 which is contained in μ of the E_i, and hence also contained in (at least) μ of the tubes T_i. These tubes form a "bush" centred at x_0, but the portions of that tube near the centre x_0 of the bush have high overlap. However, the two-ends condition can be used to finesse this issue. Indeed, that condition ensures that for each E_i involved in this bush, we have

$$|E_i \cap B(x_0, r)| \leq \frac{1}{2}|E_i|$$

for some $r \sim 1$, and thus

$$|E_i \backslash B(x_0, r)| \geq \frac{1}{2}|E_i| \gg \lambda\delta^{n-1}.$$

The δ-separated nature of the tubes T_i implies that the maximum overlap of the portion $T_i \backslash B(x_0, r)$ of the μ tubes in the bush away from the origin is $O(1)$, and so

$$|\bigcup_i E_i \backslash B(x_0, r)| \gg \mu\lambda\delta^{n-1}.$$

Thus we have two different lower bounds for $\bigcup_i E_i$, namely $\frac{\lambda}{\mu}$ and $\mu\lambda\delta^{n-1}$. Taking the geometric mean of these bounds to eliminate the unknown multiplicity μ, we obtain

$$|\bigcup_i E_i| \gg \lambda\delta^{(n-1)/2},$$

which certainly implies the desired bound since $\lambda \leq 1$. □

Remark 4.4.8. Note that the two-ends condition actually proved a *better* bound than what was actually needed for the Kakeya conjecture, in that the power of λ was more favourable than necessary. However, this gain disappears under the rescaling argument used in the proof of Proposition 4.4.5. Nevertheless, this does illustrate one of the advantages of employing the two-ends reduction; the bounds one gets upon doing so tend to be better (especially for small values of λ) than what one would have had without it, and so getting the right bound tends to be a bit easier in such cases. Note though that for the Kakeya set problem, where λ is essentially 1, the two-ends reduction is basically redundant.

Remark 4.4.9. One technical drawback to using the two-ends reduction is that if at some later stage one needs to refine the sets E_i to smaller sets, then one may lose the two-ends property. However, one could invoke the arguments used in Proposition 4.4.5 to recover this property again by refining E_i further. One may then lose some other property by this further refinement, but one convenient trick that allows one to take advantage of multiple refinements simultaneously is to iteratively refine the various sets involved and use the pigeonhole principle to find some place along this iteration where all relevant statistics of the system (e.g. the "width" r of the E_i) stabilise (here

one needs some sort of monotonicity property to obtain this stabilisation). This type of trick was introduced in [**Wo1998**] and has been used in several subsequent papers, for instance in [**LaTa2001**].

Notes. This article first appeared at `terrytao.wordpress.com/2009/05/15`. Thanks to Arie Israel, Josh Zahl, Shuanglin Shao and an anonymous commenter for corrections.

4.5. The least quadratic nonresidue, and the square root barrier

A large portion of analytic number theory is concerned with the distribution of number-theoretic sets such as the primes, or *quadratic residues* in a certain modulus. At a local level (e.g. on a short interval $[x, x+y]$), the behaviour of these sets may be quite irregular. However, in many cases one can understand the *global* behaviour of such sets on very large intervals, (e.g. $[1, x]$), with reasonable accuracy (particularly if one assumes powerful additional conjectures, such as the Riemann hypothesis and its generalisations). For instance, in the case of the primes, we have the *prime number theorem*, which asserts that the number of primes in a large interval $[1, x]$ is asymptotically equal to $x/\log x$; in the case of quadratic residues modulo a prime p, it is clear that there are exactly $(p-1)/2$ such residues in $[1, p]$. With elementary arguments, one can also count statistics such as the number of pairs of consecutive quadratic residues; and with the aid of deeper tools such as the Weil sum estimates, one can count more complex patterns in these residues also (e.g. k-point correlations).

One is often interested in converting this sort of "global" information on long intervals into "local" information on short intervals. If one is interested in the behaviour on a *generic* or *average* short interval, then the question is still essentially a global one, basically because one can view a long interval as an average of a long sequence of short intervals. (This does not mean that the problem is automatically easy, because not every global statistic about, say, the primes is understood. For instance, we do not know how to rigorously establish the conjectured asymptotic for the number of *twin primes* $n, n+2$

in a long interval $[1, N]$, and so we do not fully understand the local distribution of the primes in a typical short interval $[n, n+2]$.)

However, suppose that instead of understanding the *average-case* behaviour of short intervals, one wants to control the *worst-case* behaviour of such intervals (i.e. to establish bounds that hold for *all* short intervals, rather than *most* short intervals). Then it becomes substantially harder to convert global information to local information. In many cases one encounters a "square root barrier", in which global information at scale x (e.g. statistics on $[1, x]$) cannot be used to say anything non-trivial about a fixed (and possibly worst-case) short interval at scales $x^{1/2}$ or below. (Here we ignore factors of $\log x$ for simplicity.) The basic reason for this is that even randomly distributed sets in $[1, x]$ (which are basically the most uniform type of global distribution one could hope for) exhibit random fluctuations of size $x^{1/2}$ or so in their global statistics (as can be seen for instance from the *central limit theorem*). Because of this, one could take a random (or pseudorandom) subset of $[1, x]$ and delete all the elements in a short interval of length $o(x^{1/2})$, without anything suspicious showing up on the global statistics level; the edited set still has essentially the same global statistics as the original set. On the other hand, the worst-case behaviour of this set on a short interval has been drastically altered.

One stark example of this arises when trying to control the largest gap between consecutive prime numbers in a large interval $[x, 2x]$. There are convincing heuristics that suggest that this largest gap is of size $O(\log^2 x)$ (*Cramér's conjecture*). But even assuming the Riemann hypothesis, the best upper bound on this gap is only of size $O(x^{1/2} \log x)$, basically because of this square root barrier.

On the other hand, in some cases one can use additional tricks to get past the square root barrier. The key point is that many number-theoretic sequences have special structure that distinguish them from being exactly like random sets. For instance, quadratic residues have the basic but fundamental property that the product of two quadratic residues is again a quadratic residue. One way to use this sort of structure to amplify bad behaviour in a single short interval into bad behaviour across many short intervals (cf. Section 1.9 of *Structure*

and Randomness). Because of this amplification, one can sometimes get new worst-case bounds by tapping the average-case bounds.

In this post I would like to indicate a classical example of this type of amplification trick, namely Burgess's bound on short character sums. To narrow the discussion, I would like to focus primarily on the following classical problem:

Problem 4.5.1. What are the best bounds one can place on the first quadratic non-residue n_p in the interval $[1, p-1]$ for a large prime p?

(The first quadratic residue is, of course, 1; the more interesting problem is the first quadratic non-residue.)

Probabilistic heuristics (presuming that each non-square integer has a 50-50 chance of being a quadratic residue) suggests that n_p should have size $O(\log p)$, and indeed Vinogradov conjectured that $n_p = O_\varepsilon(p^\varepsilon)$ for any $\varepsilon > 0$. Using the *Pólya-Vinogradov inequality*, one can get the bound $n_p = O(\sqrt{p}\log p)$ (and can improve it to $n_p = O(\sqrt{p})$ using *smoothed sums*); combining this with a sieve theory argument (exploiting the multiplicative nature of quadratic residues) one can boost this to $n_p = O(p^{\frac{1}{2\sqrt{e}}}\log^2 p)$. Inserting Burgess's amplification trick one can boost this to $n_p = O_\varepsilon(p^{\frac{1}{4\sqrt{e}}+\varepsilon})$ for any $\varepsilon > 0$. Apart from refinements to the ε factor, this bound has stood for five decades as the "world record" for this problem, which is a testament to the difficulty in breaching the square root barrier.

Note: in order not to obscure the presentation with technical details, I will be using asymptotic notation $O()$ in a somewhat informal manner.

4.5.1. Character sums. To approach the problem, we begin by fixing the large prime p and introducing the *Legendre symbol* $\chi(n) = \left(\frac{n}{p}\right)$, defined to equal 0 when n is divisible by p, $+1$ when n is an invertible quadratic residue modulo p, and -1 when n is an invertible quadratic non-residue modulo p. Thus, for instance, $\chi(n) = +1$ for all $1 \leq n < n_p$. One of the main reasons one wants to work with the function χ is that it enjoys two easily verified properties:

- χ is periodic with period p.

- One has the total multiplicativity property $\chi(nm) = \chi(n)\chi(m)$ for all integers n, m.

In the jargon of number theory, χ is a *Dirichlet character* with conductor p. Another important property of this character is of course the law of *quadratic reciprocity*, but this law is more important for the *average-case* behaviour in p, whereas we are concerned here with the *worst-case* behaviour in p, and so we will not actually use this law here.

An obvious way to control n_p is via the character sum

$$\sum_{1 \leq n \leq x} \chi(n). \tag{4.18}$$

From the triangle inequality, we see that this sum has magnitude at most x. If we can then obtain a non-trivial bound of the form

$$\sum_{1 \leq n \leq x} \chi(n) = o(x) \tag{4.19}$$

for some x, this forces the existence of a quadratic residue less than or equal to x, thus $n_p \leq x$. So one approach to the problem is to bound the character sum (4.18).

As there are just as many residues as non-residues, the sum (4.18) is periodic with period p and we obtain a trivial bound of p for the magnitude of the sum. One can achieve a non-trivial bound by Fourier analysis. One can expand

$$\chi(n) = \sum_{a=0}^{p-1} \hat{\chi}(a) e^{2\pi i a n/p}$$

where $\hat{\chi}(a)$ are the Fourier coefficients of χ:

$$\hat{\chi}(a) := \frac{1}{p} \sum_{n=0}^{p-1} \chi(n) e^{-2\pi i a n/p}.$$

As there are just as many quadratic residues as non-residues, $\hat{\chi}(0) = 0$, so we may drop the $a = 0$ term. From summing the geometric series we see that

$$\sum_{1 \leq n \leq x} e^{2\pi i a n/p} = O(1/\|a/p\|), \tag{4.20}$$

where $\|a/p\|$ is the distance from a/p to the nearest integer (0 or 1); inserting these bounds into (4.18) and summing what is essentially a harmonic series in a we obtain

$$\sum_{1 \leq n \leq x} \chi(n) = O(p \log p \sup_{a \neq 0} |\hat{\chi}(a)|).$$

Now, how big is $\hat{\chi}(a)$? Taking absolute values, we get a bound of 1, but this gives us something worse than the trivial bound. To do better, we use the Plancherel identity

$$\sum_{a=0}^{p-1} |\hat{\chi}(a)|^2 = \frac{1}{p} \sum_{n=0}^{p-1} |\chi(n)|^2$$

which tells us that

$$\sum_{a=0}^{p-1} |\hat{\chi}(a)|^2 = O(1).$$

This tells us that $\hat{\chi}$ is small *on the average*, but does not immediately tell us anything new about the *worst-case* behaviour of χ, which is what we need here. But now we use the multiplicative structure of χ to relate average-case and worst-case behaviour. Note that if b is coprime to p, then $\chi(bn)$ is a scalar multiple of $\chi(n)$ by a quantity $\chi(b)$ of magnitude 1; taking Fourier transforms, this implies that $\hat{\chi}(a/b)$ and $\hat{\chi}(a)$ also differ by this factor. In particular, $|\hat{\chi}(a/b)| = |\hat{\chi}(a)|$. As b was arbitrary, we thus see that $|\hat{\chi}(a)|$ is constant for all a coprime to p; in other words, the worst case is the *same* as the average case. Combining this with the Plancherel bound one obtains $|\hat{\chi}(a)| = O(1/\sqrt{p})$, leading to the *Pólya-Vinogradov inequality*

$$\sum_{1 \leq n \leq x} \chi(n) = O(\sqrt{p} \log p).$$

(In fact, a more careful computation reveals the slightly sharper bound $|\sum_{1 \leq n \leq x} \chi(n)| \leq \sqrt{p} \log p$; this is non-trivial for $x > \sqrt{p} \log p$.)

Remark 4.5.2. Up to logarithmic factors, this is consistent with what one would expect if χ fluctuated like a random sign pattern (at least for x comparable to p; for smaller values of x, one expects instead a bound of the form $O(\sqrt{x})$, up to logarithmic factors). It is conjectured that the $\log p$ factor can be replaced with a $O(\log \log p)$ factor, which would be consistent with the random fluctuation model

and is best possible; this is known for GRH, but unconditionally the Pólya-Vinogradov inequality is still the best known. (See however *http://arxiv.org/abs/math/0503113*this paper of Granville and Soundararajan for an improvement for non-quadratic characters χ.)

A direct application of the Pólya-Vinogradov inequality gives the bound $n_p \leq \sqrt{p} \log p$. One can get rid of the logarithmic factor (which comes from the harmonic series arising from (4.20)) by replacing the sharp cutoff $1_{1 \leq n \leq x}$ by a smoother sum, which has a better behaved Fourier transform. But one can do better still by exploiting the multiplicativity of χ again, by the following trick of Vinogradov. Observe that not only does one have $\chi(n) = +1$ for all $n \leq n_p$, but also $\chi(n) = +1$ for any n which is $n_p - 1$-smooth, i.e. is the product of integers less than n_p. So even if n_p is significantly less than x, one can show that the sum (4.18) is large if the majority of integers less than x are $n_p - 1$-smooth.

Since every integer n less than x is either n_p-smooth (in which case $\chi(n) = +1$), or divisible by a prime q between n_p and x (in which case $\chi(n)$ is at least -1), we obtain the lower bound

$$\sum_{1 \leq n \leq x} \chi(n) \geq \sum_{1 \leq n \leq x} 1 - \sum_{n_p < q \leq x} \sum_{1 \leq n \leq x: q|n} 2.$$

Clearly, $\sum_{1 \leq n \leq x} 1 = x + O(1)$ and $\sum_{1 \leq n \leq x: q|n} 2 = 2\frac{x}{q} + O(1)$. The total number of primes less than x is $O(\frac{x}{\log x}) = o(x)$ by the prime number theorem, thus

$$\sum_{1 \leq n \leq x} \chi(n) \geq x - \sum_{n_p < q \leq x} 2\frac{x}{q} + o(x).$$

Using the classical asymptotic $\sum_{q \leq y} \frac{1}{q} = \log\log y + C + o(1)$ for some absolute constant C (which basically follows from the prime number theorem, but also has an elementary proof), we conclude that

$$\sum_{1 \leq n \leq x} \chi(n) \geq x[1 - 2\log\frac{\log x}{\log n_p} + o(1)].$$

If $n_p \geq x^{\frac{1}{\sqrt{e}}+\varepsilon}$ for some fixed $\varepsilon > 0$, then the expression in brackets is bounded away from zero for x large; in particular, this is incompatible with (4.19) for x large enough. As a consequence, we see that if we have a bound of the form (4.19), then we can conclude $n_p =$

$O_\varepsilon(x^{\frac{1}{\sqrt{e}}+\varepsilon})$ for all $\varepsilon > 0$; in particular, from the Pólya-Vinogradov inequality one has

$$n_p = O_\varepsilon(p^{\frac{1}{2\sqrt{e}}+\varepsilon})$$

for all $\varepsilon > 0$, or equivalently that $n_p \leq p^{\frac{1}{2\sqrt{e}}+o(1)}$. (By being a bit more careful, one can refine this to $n_p = O(p^{\frac{1}{2\sqrt{e}}} \log^{2/\sqrt{e}} p)$.)

Remark 4.5.3. The estimates on the Gauss-type sums $\hat{\chi}(a) := \frac{1}{p}\sum_{n=0}^{p-1} \chi(n)e^{-2\pi ian/p}$ are sharp; nevertheless, they fail to penetrate the square root barrier in the sense that no non-trivial estimates are provided below the scale $\sqrt{p}$. One can also see this barrier using the *Poisson summation formula* (Exercise 1.12.41), which basically gives a formula that (very roughly) takes the form

$$\sum_{n=O(x)} \chi(n) \sim \frac{x}{\sqrt{p}} \sum_{n=O(p/x)} \chi(n)$$

for any $1 < x < p$, and is basically the limit of what one can say about character sums using Fourier analysis alone. In particular, we see that the Pólya-Vinogradov bound is basically the Poisson dual of the trivial bound. The scale $x = \sqrt{p}$ is the crossing point where Poisson summation does not achieve any non-trivial modification of the scale parameter.

4.5.2. Average-case bounds. The Pólya-Vinogradov bound establishes a non-trivial estimate (4.18) for x significantly larger than $\sqrt{p}\log p$. We are interested in extending (4.18) to shorter intervals.

Before we address this issue for a fixed interval $[1, x]$, we first study the *average-case* bound on short character sums. Fix a short length y, and consider the shifted sum

$$\sum_{a \leq n \leq a+y} \chi(n), \tag{4.21}$$

where a is a parameter. The analogue of (4.18) for such intervals would be

$$\sum_{a \leq n \leq a+y} \chi(n) = o(y). \tag{4.22}$$

For y very small (e.g. $y = p^\varepsilon$ for some small $\varepsilon > 0$), we do not know how to establish (4.22) for *all* a; but we can at least establish (4.22)

for *almost all* a, with only about $O(\sqrt{p})$ exceptions (here we see the square root barrier again!).

More precisely, we will establish the moment estimates

$$\frac{1}{p}\sum_{a=0}^{p-1}|\sum_{a\leq n\leq a+y}\chi(n)|^k = O_k(y^{k/2}+y^k p^{-1/2}) \tag{4.23}$$

for any positive even integer $k = 2, 4, \ldots$. If y is not too tiny, say $y \geq p^\varepsilon$ for some $\varepsilon > 0$, then by applying (4.23) for a sufficiently large k and using Chebyshev's inequality (or Markov's inequality), we see (for any given $\delta > 0$) that one has the non-trivial bound

$$|\sum_{a\leq n\leq a+y}\chi(n)| \leq \delta y$$

for all but at most $O_{\delta,\varepsilon}(\sqrt{p})$ values of $a \in [1,p]$.

To see why (4.23) is true, let us just consider the easiest case $k = 2$. Squaring both sides, we expand (4.23) as

$$\frac{1}{p}\sum_{a=0}^{p-1}\sum_{a\leq n,m\leq a+y}\chi(n)\chi(m) = O(y) + O(y^2 p^{-1/2}).$$

We can write $\chi(n)\chi(m)$ as $\chi(nm)$. Writing $m = n+h$, and using the periodicity of χ, we can rewrite the left-hand side as

$$\sum_{h=-y}^{y}(y-|h|)[\frac{1}{p}\sum_{n\in F_p}\chi(n(n+h))]$$

where we have abused notation and identified the finite field F_p with $\{0,1,\ldots,p-1\}$.

For $h = 0$, the inner average is $O(1)$. For h non-zero, we claim the bound

$$\sum_{n\in F_p}\chi(n(n+h)) = O(\sqrt{p}) \tag{4.24}$$

which is consistent with (and is in fact slightly stronger than) what one would get if χ was a random sign pattern; assuming this bound gives (4.23) for $k = 2$ as required.

The bound (4.24) can be established by quite elementary means (as it comes down to counting points on the hyperbola $y^2 = x(x+h)$,

which can be done by transforming the hyperbola to be rectangular), but for larger values of k we will need the more general estimate

$$\sum_{n \in F_p} \chi(P(n)) = O_k(\sqrt{p}) \tag{4.25}$$

whenever P is a polynomial over F of degree k which is not a constant multiple of a perfect square; this can be easily seen to give (4.23) for general k.

An equivalent form of (4.25) is that the *hyperelliptic curve*

$$\{(x, y) \in F_p \times F_p : y^2 = P(x)\} \tag{4.26}$$

contains $p + O_k(\sqrt{p})$ points. This fact follows from a general theorem of Weil establishing the Riemann hypothesis for curves over function fields, but can also be deduced by a more elementary argument of Stepanov[**St1969**], using the polynomial method, which we now give here. (This arrangement of the argument is based on the exposition in [**IwKo2004**].)

By translating the x variable we may assume that $P(0)$ is non-zero. The key lemma is the following. Assume p large, and take l to be an integer comparable to $\sqrt{p}$ (other values of this parameter are possible, but this is the optimal choice). All polynomials $Q(x)$ are understood to be over the field F_p (i.e. they lie in the polynomial ring $F_p[X]$), although indeterminate variables x need not lie in this field.

Lemma 4.5.4. *There exists a non-zero polynomial $Q(x)$ of one indeterminate variable x over F_p of degree at most $lp/2 + O_k(p)$ which vanishes to order at least l at every point $x \in F_p$ for which $P(x)$ is a quadratic residue.*

Note from the factor theorem that Q can vanish to order at least l at at most $\deg(Q)/l \leq p/2 + O_k(\sqrt{p})$ points, and so we see that $P(x)$ is an invertible quadratic residue for at most $p/2 + O_k(\sqrt{p})$ values of F_p. Multiplying P by a quadratic non-residue and running the same argument, we also see that $P(x)$ is an invertible quadratic non-residue for at most $p/2 + O_k(\sqrt{p})$ values of F_p, and (4.25) (or the asymptotic for the number of points in (4.26)) follows.

We now prove the lemma. The polynomial Q will be chosen to be of the form

$$Q(x) = P^l(x)(R(x, x^p) + P^{\frac{p-1}{2}}(x)S(x, x^p))$$

where $R(x, z), S(x, z)$ are polynomials of degree at most $\frac{p-k-1}{2}$ in x, and degree at most $\frac{l}{2}+C$ in z, where C is a large constant (depending on k) to be chosen later (these parameters have been optimised for the argument that follows). Since P has degree at most k, such a Q will have degree

$$\leq kl + \frac{p-k-1}{2} + \frac{p-1}{2}k + p(\frac{l}{2} + C') = \frac{lp}{2} + O_k(p)$$

as required. We claim (for suitable choices of C, C') that

(a) The degrees are small enough that $Q(x)$ is a non-zero polynomial whenever $R(x, z), S(x, z)$ are non-zero polynomials; and

(b) The degrees are large enough that there exists a non-trivial choice of $R(x, z)$ and $S(x, z)$ that $Q(x)$ vanishes to order at least l whenever $x \in F_p$ is such that $P(x)$ is a quadratic residue.

Claims (a) and (b) together establish the lemma.

We first verify (a). We can cancel off the initial P^l factor, so that we need to show that $R(x, x^p)+P^{\frac{p-1}{2}}(x)S(x, x^p)$ does not vanish when at least one of $R(x, z), Q(x, z)$ is not vanishing. We may assume that R, Q are not both divisible by z, since we could cancel out a common factor of x^p otherwise.

Suppose for contradiction that the polynomial $R(x, x^p)+P^{\frac{p-1}{2}}S(x, x^p)$ vanished, which implies that $R(x, 0) = -P^{\frac{p-1}{2}}(x)S(x, 0)$ modulo x^p. Squaring and multiplying by P, we see that

$$R(x, 0)^2 P(x) = P(x)^p S(x, 0)^2 \bmod x^p.$$

But over F_p and modulo x^p, $P(x)^p = P(0)$ by Fermat's little theorem. Observe that $R(x, 0)^2 P(x)$ and $P(0)S(x, 0)^2$ both have degree at most $p - 1$, and so we can remove the x^p modulus and conclude that $R(x, 0)^2 P(x) = P(0)S(x, 0)^2$ over F_p. But this implies (by the fundamental theorem of arithmetic for $F_p[X]$) that P is a constant

multiple of a square, a contradiction. (Recall that $P(0)$ is non-zero, and that $R(x,0)$ and $S(x,0)$ are not both zero.)

Now we prove (b). Let $x \in F_p$ be such that $P(x)$ is a quadratic residue, thus $P(x)^{\frac{p-1}{2}} = +1$ by Fermat's little theorem. To get vanishing to order l, we need

$$\frac{d^j}{dx^j}[P^l(x)(R(x,x^p) + P^{\frac{p-1}{2}}(x)S(x,x^p))] = 0 \tag{4.27}$$

for all $0 \leq j < l$. (Of course, we cannot define derivatives using limits and Newton quotients in this finite characteristic setting, but we can still define derivatives of polynomials formally, thus for instance $\frac{d}{dx}x^n := nx^{n-1}$, and enjoy all the usual rules of calculus, such as the product rule and chain rule.)

Over F_p, the polynomial x^p has derivative zero. If we then compute the derivative in (4.27) using many applications of the product and chain rule, we see that the left-hand side of (4.27) can be expressed in the form

$$P^{l-j}(x)[R_j(x,x^p) + P^{\frac{p-1}{2}}(x)S_j(x,x^p))]$$

where $R_j(x,z), S_j(x,z)$ are polynomials of degree at most $\frac{p-k-1}{2} + O_k(j)$ in x and at most $\frac{l}{2} + C$ in z, whose coefficients depend in some linear fashion on the coefficients of R and S. (The exact nature of this linear relationship will depend on k, p, P, but this will not concern us.) Since we only need to evaluate this expression when $P(x)^{\frac{p-1}{2}} = +1$ and $x^p = p$ (by Fermat's little theorem), we thus see that we can verify (4.27) provided that the polynomial

$$P^{l-j}(x)[R_j(x,x) + S_j(x,x))]$$

vanishes identically. This is a polynomial of degree at most

$$O(l-j) + \frac{p-k-1}{2} + O_k(j) + \frac{l}{2} + C = \frac{p}{2} + O_k(p^{1/2}) + C,$$

and there are $l+1$ possible values of j, so this leads to

$$\frac{lp}{2} + O_k(p) + O(C\sqrt{p})$$

linear constraints on the coefficients of R and S to satisfy. On the other hand, the total number of these coefficients is

$$2 \times (\frac{p-k-1}{2} + O(1)) \times (\frac{l}{2} + C + O(1)) = \frac{lp}{2} + Cp + O_k(p).$$

For C large enough, there are more coefficients than constraints, and so one can find a non-trivial choice of coefficients obeying the constraints (4.27), and (b) follows.

Remark 4.5.5. If one optimises all the constants here, one gets an upper bound of basically $8k\sqrt{p}$ for the deviation in the number of points in (4.26). This is only a little worse than the sharp bound of $2g\sqrt{p}$ given from Weil's theorem, where $g = \lfloor \frac{k-1}{2} \rfloor$ is the genus; however, it is possible to boost the former bound to the latter by using a version of the tensor power trick (generalising F_p to F_{p^m} and then letting $m \to \infty$) combined with the theory of *Artin L-functions* and the *Riemann-Roch theorem*. This is (very briefly!) sketched in Section 1.9 of *Structure and Randomness*.

Remark 4.5.6. Once again, the global estimate (4.25) is very sharp, but cannot penetrate below the square root barrier, in that one is allowed to have about $O(\sqrt{p})$ exceptional values of a for which no cancellation exists. One expects that these exceptional values of a in fact do not exist, but we do not know how to do this unless y is larger than $x^{1/4}$ (so that the Burgess bounds apply).

4.5.3. The Burgess bound. The average case bounds in the previous section give an alternate demonstration of a non-trivial estimate (4.18) for $x > p^{1/2+\varepsilon}$, which is just a bit weaker than what the Pólya-Vinogradov inequality gives. Indeed, if (4.18) failed for such an x, thus

$$|\sum_{n \in [1,x]} \chi(n)| \gg x,$$

then by taking a small y (e.g. $y = p^{\varepsilon/2}$) and covering $[1,x]$ by intervals of length y, we see (from a first moment method argument) that

$$|\sum_{a \leq n \leq a+y} \chi(n)| \gg y$$

for a positive fraction of the a in $[1,x]$. But this contradicts the results of the previous section.

Burgess observed that by exploiting the multiplicativity of χ one last time to amplify the above argument, one can extend the range for which (4.18) can be proved from $x > p^{1/2+\varepsilon}$ to also cover the range $p^{1/4+\varepsilon} < x < p^{1/2}$. The idea is not to cover $[1, x]$ by intervals of length y, but rather by *arithmetic progressions* $\{a, a + r, \ldots, a + yr\}$ of length y, where $a = O(x)$ and $r = O(x/y)$. Another application of the first moment method then shows that

$$|\sum_{0 \leq j \leq y} \chi(a + jr)| \gg y$$

for a positive fraction of the a in $[1, x]$ and r in $[1, x/y]$ (i.e. $\gg x^2/y$ such pairs (a, r)).

For technical reasons, it will be inconvenient if a and r have too large of a common factor, so we pause to remove this possibility. Observe that for any $d \geq 1$, the number of pairs (a, r) which have d as a common factor is $O(\frac{1}{d^2}x^2/y)$. As $\sum_{d=1}^{\infty} \frac{1}{d^2}$ is convergent, we may thus remove those pairs which have too large of a common factor, and assume that all pairs (a, r) have common factor $O(1)$ at most (so are "almost coprime").

Now we exploit the multiplicativity of χ to write $\chi(a + jr)$ as $\chi(r)\chi(b+j)$, where b is a residue which is equal to $a/r \bmod q$. Dividing out by $\chi(r)$, we conclude that

$$|\sum_{0 \leq j \leq y} \chi(b + j)| \gg y \tag{4.28}$$

for $\gg x^2/y$ pairs (a, r).

Now for a key observation: the $\gg x^2/y$ values of b arising from the pairs (a, r) are mostly disjoint. Indeed, suppose that two pairs $(a, r), (a', r')$ generated the same value of b, thus $a/r = a'/r' \bmod p$. This implies that $ar' = a'r \bmod p$. Since $x < p^{1/2}$, we see that $ar', a'r$ do not exceed p, so we may remove the modulus and conclude that $ar' = a'r$. But since we are assuming that a, r and a', r' are almost coprime, we see that for each (a, r) there are at most $O(1)$ values of a', r' for which $ar' = a'r$. So the b's here only overlap with multiplicity $O(1)$, and we conclude that (4.28) holds for $\gg x^2/y$ values of b. But comparing this with the previous section, we obtain a contradiction unless $x^2/y \ll \sqrt{p}$. Setting y to be a sufficiently small power of p, we

obtain Burgess's result that (4.18) holds for $x > p^{1/4+\varepsilon}$ for any fixed $\varepsilon > 0$.

Combining Burgess's estimate with Vinogradov's sieving trick we conclude the bound $n_p = O_\varepsilon(p^{1/4\sqrt{e}+\varepsilon})$ for all $\varepsilon > 0$, which is the best bound known today on the least quadratic non-residue except for refinements of the p^ε error term.

Remark 4.5.7. There are many generalisations of this result, for instance to more general characters (with possibly composite conductor), or to shifted sums (4.21). However, the $p^{1/4}$ type exponent has not been improved except with the assistance of powerful conjectures such as the *generalised Riemann hypothesis*.

Notes. This article first appeared at `terrytao.wordpress.com/2009/08/18`. Thanks to Efthymios Sofos, Joshua Zelinsky, K, and Seva Lev for corrections.

Boris noted the similarity between the use of the Frobenius map $x \mapsto x^p$ in Stepanov's argument and Thues trick from the proof of his famous result on the Diophantine approximations to algebraic numbers, where instead of the exact equality $x = x^p$ that is used here, he used two very good approximations to the same algebraic number.

4.6. Determinantal processes

Given a set S, a (simple) *point process* is a random subset A of S. (A non-simple point process would allow multiplicity; more formally, A is no longer a subset of S, but is a Radon measure on S, where we give S the structure of a locally compact Polish space, but I do not wish to dwell on these sorts of technical issues here.) Typically, A will be finite or countable, even when S is uncountable. Basic examples of point processes include

- (Bernoulli point process) S is an at most countable set, $0 \leq p \leq 1$ is a parameter, and A a random set such that the events $x \in A$ for each $x \in S$ are jointly independent and occur with a probability of p each. This process is automatically simple.

- (Discrete *Poisson point process*) S is an at most countable space, λ is a measure on S (i.e. an assignment of a non-negative number $\lambda(\{x\})$ to each $x \in S$), and A is a *multiset* where the multiplicity of x in A is a *Poisson random variable* with intensity $\lambda(\{x\})$, and the multiplicities of $x \in A$ as x varies in S are jointly independent. This process is usually not simple.

- (Continuous *Poisson point process*) S is a locally compact Polish space with a Radon measure λ, and for each $\Omega \subset S$ of finite measure, the number of points $|A \cap \Omega|$ that A contains inside Ω is a *Poisson random variable* with intensity $\lambda(\Omega)$. Furthermore, if $\Omega_1, \ldots, \Omega_n$ are disjoint sets, then the random variables $|A \cap \Omega_1|, \ldots, |A \cap \Omega_n|$ are jointly independent. (The fact that Poisson processes exist at all requires a non-trivial amount of measure theory, and will not be discussed here.) This process is almost surely simple iff all points in S have measure zero.

- (Spectral point processes) The spectrum of a random matrix is a point process in $\mathbf{C}$ (or in $\mathbf{R}$, if the random matrix is Hermitian). If the spectrum is almost surely simple, then the point process is almost surely simple. In a similar spirit, the zeroes of a random polynomial are also a point process.

A remarkable fact is that many natural (simple) point processes are *determinantal processes*. Very roughly speaking, this means that there exists a positive semi-definite kernel $K : S \times S \to \mathbf{R}$ such that, for any $x_1, \ldots, x_n \in S$, the probability that $x_1, \ldots, x_n$ all lie in the random set A is proportional to the determinant $\det((K(x_i, x_j))_{1 \leq i,j \leq n})$. Examples of processes known to be determinantal include non-intersecting random walks, spectra of random matrix ensembles such as GUE, and zeroes of polynomials with gaussian coefficients.

I would be interested in finding a good explanation (even at the heuristic level) as to why determinantal processes are so prevalent in practice. I do have a very weak explanation, namely that determinantal processes obey a large number of rather pretty algebraic identities, and so it is plausible that any other process which has a

very algebraic structure (in particular, any process involving gaussians, characteristic polynomials, etc.) would be connected in some way with determinantal processes. I'm not particularly satisfied with this explanation, but I thought I would at least describe some of these identities below to support this case. (This is partly for my own benefit, as I am trying to learn about these processes, particularly in connection with the spectral distribution of random matrices.) The material here is partly based on [**HoKrPeVi2006**].

4.6.1. Discrete determinantal processes. In order to ignore all measure-theoretic distractions and focus on the algebraic structure of determinantal processes, we will first consider the discrete case when the space S is just a finite set $S = \{1, \ldots, N\}$ of cardinality $|S| = N$. We say that a process $A \subset S$ is a *determinantal process* with kernel K, where K is an $k \times k$ symmetric real matrix, if one has

$$\mathbf{P}(\{i_1, \ldots, i_k\} \subset A) = \det(K(i_a, i_b))_{1 \leq a,b \leq k} \tag{4.29}$$

for all distinct $i_1, \ldots, i_k \in S$.

To build determinantal processes, let us first consider point processes of a fixed cardinality n, thus $0 \leq n \leq N$ and A is a random subset of S of size n, or in other words a random variable taking values in the set $\binom{S}{n} := \{B \subset S : |B| = n\}$.

In this simple model, an n-element point processes is basically just a collection of $\binom{N}{n}$ probabilities $p_B = \mathbf{P}(A = B)$, one for each $B \in \binom{S}{n}$, which are non-negative numbers which add up to 1. For instance, in the uniform point process where A is drawn uniformly at random from $\binom{S}{n}$, each of these probabilities p_B would equal $1/\binom{N}{n}$. How would one generate other interesting examples of n-element point processes?

For this, we can borrow the idea from quantum mechanics that probabilities can arise as the square of coefficients of unit vectors, though unlike quantum mechanics it will be slightly more convenient here to work with real vectors rather than complex ones. To formalise this, we work with the n^{th} *exterior power* $\bigwedge^n \mathbf{R}^N$ of the Euclidean space $\mathbf{R}^N$; this space is sort of a "quantisation" of $\binom{S}{n}$, and is analogous to the space of quantum states of n identical *fermions*, if each fermion can exist classically in one of N states (or "spins"). (The

requirement that the process be simple is then analogous to the *Pauli exclusion principle*.)

This space of n-vectors in $\mathbf{R}^N$ is spanned by the wedge products $e_{i_1} \wedge \ldots \wedge e_{i_n}$ with $1 \leq i_1 < \ldots < i_n \leq N$, where $e_1, \ldots, e_N$ is the standard basis of $\mathbf{R}^N$. There is a natural inner product to place on $\bigwedge^n \mathbf{R}^N$ by declaring all the $e_{i_1} \wedge \ldots \wedge e_{i_n}$ to be orthonormal.

Lemma 4.6.1. *If $f_1, \ldots, f_N$ is any orthonormal basis of $\mathbf{R}^N$, then the $f_{i_1} \wedge \ldots \wedge f_{i_n}$ for $1 \leq i_1 < \ldots < i_n \leq N$ are an orthonormal basis for $\bigwedge^n \mathbf{R}^N$.*

Proof. By definition, this is true when $(f_1, \ldots, f_N) = (e_1, \ldots, e_N)$. If the claim is true for some orthonormal basis $f_1, \ldots, f_N$, it is not hard to see that the claim also holds if one rotates f_i and f_j in the plane that they span by some angle θ, where $1 \leq i < j \leq n$ are arbitrary. But any orthonormal basis can be rotated into any other by a sequence of such rotations (e.g. by using *Euler angles*), and the claim follows. □

Corollary 4.6.2. *If $v_1, \ldots, v_n$ are vectors in $\mathbf{R}^N$, then the magnitude of $v_1 \wedge \ldots \wedge v_n$ is equal to the n-dimensional volume of the parallelopiped spanned by $v_1, \ldots, v_n$.*

Proof. Observe that applying row operations to v_i (i.e. modifying one v_i by a scalar multiple of another v_j) does not affect either the wedge product or the volume of the parallelopiped. Thus by using the *Gram-Schmidt process*, we may assume that the v_i are orthogonal; by normalising we may assume they are orthonormal. The claim now follows from the preceding lemma. □

From this and the ordinary Pythagorean theorem in the inner product space $\bigwedge^n \mathbf{R}^N$, we conclude the *multidimensional Pythagorean theorem*: the square of the n-dimensional volume of a parallelopiped in $\mathbf{R}^N$ is the sum of squares of the n-dimensional volumes of the projection of that parallelopiped to each of the $\binom{N}{n}$ coordinate subspaces $\mathrm{span}(e_{i_1}, \ldots, e_{i_n})$. (I believe this theorem was first observed in this generality by Donchian and Coxeter.) We also note another related fact:

Lemma 4.6.3 (Gram identity)**.** *If $v_1, \ldots, v_n$ are vectors in $\mathbf{R}^N$, then the square of the magnitude of $v_1 \wedge \ldots \wedge v_n$ is equal to the determinant of the* Gram matrix $(v_i \cdot v_j)_{1 \leq i,j \leq n}$.

Proof. Again, the statement is invariant under row operations, and one can reduce as before to the case of an orthonormal set, in which case the claim is clear. (Alternatively, one can proceed via the *Cauchy-Binet formula*.) □

If we define $e_{\{i_1,\ldots,i_n\}} := e_{i_1} \wedge \ldots \wedge e_{i_n}$, then we have identified the standard basis of $\bigwedge^n \mathbf{R}^N$ with $\binom{S}{n}$ by identifying e_B with B. As a consequence of this and the multidimensional Pythagorean theorem, every unit n-vector ω in $\bigwedge^n \mathbf{R}^N$ determines an n-element point process A on S, by declaring the probability p_B of A taking the value B to equal $|\omega \cdot e_B|^2$ for each $B \in \binom{S}{n}$. Note that multiple n-vectors can generate the same point process, because only the magnitude of the coefficients $\omega \cdot e_B$ are of interest; in particular, ω and $-\omega$ generate the same point process. (This is analogous to how multiplying the wave function in quantum mechanics by a complex phase has no effect on any physical observable.)

Now we can introduce determinantal processes. If V is an n-dimensional subspace of $\mathbf{R}^N$, we can define the (projection) *determinantal process* $A = A_V$ associated to V to be the point process associated to the *volume form* of V, i.e. to the wedge product of an orthonormal basis of V. (This volume form is only determined up to sign, because the orientation of V has not been fixed, but as observed previously, the sign of the form has no impact on the resulting point process.)

By construction, the probability that the point process A is equal to a set $\{i_1, \ldots, i_n\}$ is equal to the square of the determinant of the $n \times n$ matrix consisting of the $i_1, \ldots, i_n$ coordinates of an arbitrary orthonormal basis of V. By extending such an orthonormal basis to the rest of $\mathbf{R}^N$, and representing $e_{i_1}, \ldots, e_{i_n}$ in this basis, it is not hard to see that $\mathbf{P}(A = \{i_1, \ldots, i_n\})$ can be interpreted geometrically as the square of the volume of the parallelopiped generated by $Pe_{i_1}, \ldots, Pe_{i_n}$, where P is the orthogonal projection onto V.

In fact we have the more general fact:

Lemma 4.6.4. *If $k \geq 1$ and $i_1, \ldots, i_k$ are distinct elements of S, then $\mathbf{P}(\{i_1, \ldots, i_k\} \subset A)$ is equal to the square of the k-dimensional volume of the parallelopiped generated by the orthogonal projections of $Pe_{i_1}, \ldots, Pe_{i_k}$ to V.*

Proof. We can assume that $k \leq n$, since both expressions in the lemma vanish otherwise.

By (anti-)symmetry we may assume that $\{i_1, \ldots, i_k\} = \{1, \ldots, k\}$. By the Gram-Schmidt process we can find an orthonormal basis $v_1, \ldots, v_n$ of V such that each v_i is orthogonal to $e_1, \ldots, e_{i-1}$.

Now consider the $n \times N$ matrix M with rows $v_1, \ldots, v_n$, thus M vanishes below the diagonal. The probability $\mathbf{P}(\{1, \ldots, k\} \in A)$ is equal to the sum of squares of the determinants of all the $n \times n$ minors of M that contain the first k rows. As M vanishes below the diagonal, we see from cofactor expansion that this is equal to the product of the squares of the first k diagonal entries, times the sum of squares of the determinants of all the $n-k \times n-k$ minors of the bottom $n-k$ rows. But by the generalised Pythagorean theorem, this latter factor is the square of the volume of the parallelopiped generated by $v_{k+1}, \ldots, v_n$, which is 1. Meanwhile, by the base times height formula, we see that the product of the first k diagonal entries of M is equal in magnitude to the k-dimensional volume of the orthogonal projections of $e_1, \ldots, e_k$ to V. The claim follows. □

As a special case of Lemma 4.6.4, we have $\mathbf{P}(i \in A) = \|Pe_i\|^2$ for any i. In particular, if e_i lies in V, then i almost surely lies in A, and when e_i is orthogonal to V, i almost surely is disjoint from A.

Let $K(i,j) = Pe_i \cdot e_j = Pe_i \cdot Pe_j$ denote the matrix coefficients of the orthogonal projection P. From Lemma 4.6.4 and the Gram identity, we conclude that A is a determinantal process (see (4.29)) with kernel K. Also, by combining Lemma 4.6.4 with the generalised Pythagorean theorem, we conclude a monotonicity property:

Lemma 4.6.5 (Monotonicity property)**.** *If $V \subset W$ are nested subspaces of $\mathbf{R}^N$, then $\mathbf{P}(B \subset A_V) \leq \mathbf{P}(B \subset A_W)$ for every $B \subset S$.*

This seems to suggest that there is some way of representing A_W as the union of A_V with another process coupled with A_V, but I was

not able to build a non-artificial example of such a representation. On the other hand, if $V \subset \mathbf{R}^N$ and $V' \subset \mathbf{R}^{N'}$, then the process $A_{V \oplus V'}$ associated with the direct sum $V \oplus V' \subset \mathbf{R}^{N+N'}$ has the same distribution of the disjoint union of A_V with an independent copy of $A_{V'}$.

The determinantal process interacts nicely with complements:

Lemma 4.6.6 (Hodge duality)**.** *Let V be an n-dimensional subspace of $\mathbf{R}^N$. The $N-n$-element determinantal process $A_{V^\perp}$ associated to the orthogonal complement $V^\perp$ of V has the same distribution as the complement $S \backslash A_V$ of the n-element determinantal process A_V associated to V.*

Proof. We need to show that $\mathbf{P}(A_V = B) = \mathbf{P}(A_{V^\perp} = S \backslash B)$ for all $B \in \binom{N}{n}$. By symmetry we can take $B = \{1, \ldots, n\}$. Let $v_1, \ldots, v_n$ and $v_{n+1}, \ldots, v_N$ be an orthonormal basis for V and $V^\perp$ respectively, and let M be the resulting $N \times N$ orthogonal matrix; then the task is to show that the top $n \times n$ minor X of M has the same determinant squared as the bottom $N-n \times N-n$ minor Y. But if one splits $M = \begin{pmatrix} X & Z \\ W & Y \end{pmatrix}$, we see from the orthogonality property that $XX^* = I_n - ZZ^*$ and $Y^*Y = I_{N-n} - Z^*Z$, where I_n is the $n \times n$ identity matrix. But from the *singular value decomposition* we see that $I_n - ZZ^*$ and $I_{N-n} - Z^*Z$ have the same determinant, and the claim follows. (One can also establish this lemma using the *Hodge star operation.*) □

From this lemma we see that $S \backslash A$ is a determinantal process with kernel $I_N - K$. In particular, we have

$$\mathbf{P}(\{i_1, \ldots, i_k\} \cap A = \emptyset) = \det(I_k - (K(i_a, i_b))_{1 \leq a,b \leq k}). \tag{4.30}$$

The construction of the determinantal process given above is somewhat indirect. A more direct way to build the process exploits the following lemma:

Lemma 4.6.7. *Let V be an n-dimensional subspace of $\mathbf{R}^N$, let A_V be the corresponding n-element determinantal process, and let $1 \leq i_1 < \ldots < i_k \leq N$ for some $1 \leq k \leq n$. Then the if one conditions on the event that $\{i_1, \ldots, i_k\} \in A_V$ (assuming this event has non-zero probability), the resulting $n-k$-element process $A_V \backslash \{i_1, \ldots, i_k\}$ has*

the same distribution as the $n-k$-element determinantal process A_W associated to the $n-k$-dimensional subspace W of V that is orthogonal to $e_{i_1}, \ldots, e_{i_k}$.

Proof. By symmetry it suffices to consider the case $\{i_1, \ldots, i_k\} = \{1, \ldots, k\}$. By a further application of symmetry it suffices to show that

$$\mathbf{P}(A_V = \{1, \ldots, n\}) = \mathbf{P}(\{1, \ldots, k\} \subset A_V)\mathbf{P}(A_W = \{k+1, \ldots, n\}).$$

By the Gram-Schmidt process, we can find an orthonormal basis $v_1, \ldots, v_n$ of V whose $n \times N$ matrix of coefficients vanishes below the diagonal. One then easily verifies (using Lemma 4.6.4) that $\mathbf{P}(A_V = \{1, \ldots, n\})$ is the product of the n diagonal entries, $\mathbf{P}(\{1, \ldots, k\} \subset A_V)$ is the product of the first k, and $\mathbf{P}(A_W = \{k+1, \ldots, n\})$ is the product of the last $n-k$, and the claim follows. □

Remark 4.6.8. There is a dual version of this lemma: if one conditions on the event that $\{i_1, \ldots, i_k\}$ is disjoint from A_V, then the resulting process is the determinantal process associated to the orthogonal projection of V to the orthogonal complement of $e_{i_1}, \ldots, e_{i_k}$.

From this lemma, it is not difficult to see that one can build A_V recursively as $A_V = \{a\} \cup A_{V_a}$, where a is a random variable drawn from S with a $\mathbf{P}(a = i) = \|Pe_i\|^2 / \dim(V)$ for each i, and V_a is the subspace of V orthogonal to e_a. Another consequence of this lemma and the monotonicity property is the negative dependence inequality

$$\mathbf{P}(B_1 \cup B_2 \subset A) \leq \mathbf{P}(B_1 \subset A)\mathbf{P}(B_2 \subset A)$$

for any disjoint $B_1, B_2 \subset S$; thus the presence of A on one set B_1 reduces the chance of A being present on a disjoint set B_2 (not surprising, since A has fixed size).

Thus far, we have only considered point processes with a fixed number n of points. As a consequence, the determinantal kernel K involved here is of a special form, namely the coefficients of an orthogonal projection matrix to an n-dimensional space (or equivalently, a symmetric matrix whose eigenvalues consist of n ones and $N-n$ zeroes). But one can create more general point processes by taking a *mixture* of the fixed-number processes, e.g. first picking a projection

kernel K (or a subspace V) by some random process, and then sampling A from the point process associated to that kernel or subspace.

For instance, let $\phi_1, \ldots, \phi_N$ be an orthonormal basis of $\mathbf{R}^N$, and let $0 \leq \lambda_1, \ldots, \lambda_N \leq 1$ be weights. Then we can create a random subspace V of $\mathbf{R}^N$ by setting V equal to the span V_J of some random subset $\{\phi_j : j \in J\}$ of the basis $v_1, \ldots, v_N$, where each j lies in J with an independent probability of λ_j, and then sampling A from A_V. Then A will be a point process whose cardinality can range from 0 to N. Given any set $\{i_1, \ldots, i_k\} \subset S$, we can then compute the probability $\mathbf{P}(\{i_1, \ldots, i_k\} \subset A)$ as

$$\mathbf{P}(\{i_1, \ldots, i_k\} \subset A) = \mathbf{E}_J \mathbf{P}(\{i_1, \ldots, i_k\} \subset A_{V_J})$$

where J is selected as above. Using (4.29), we have

$$\mathbf{P}(\{i_1, \ldots, i_k\} \subset A_{V_J}) = \det(K_{V_J}(i_a, i_b))_{1 \leq a,b \leq k}.$$

But $K_{V_J}(i_a, i_b) = \sum_{j \in J} \phi_j(i_a)\phi_j(i_b)$, where $\phi_j(i)$ is the i^{th} coordinate of ϕ_j. Thus we can write

$$(K_{V_J}(i_a, i_b))_{1 \leq a,b \leq k} = \sum_{j=1}^{N} \mathbf{I}(j \in J) R_j$$

where $\mathbf{I}(j \in J)$ is the indicator of the event $j \in J$, and R_j is the rank one matrix $(\phi_j(i_a)\phi_j(i_b))_{1 \leq a,b \leq k}$. Using multilinearity of the determinant, and the fact that any determinant involving two or more rows of the same rank one matrix automatically vanishes, we see that we can express

$$\det((K_{V_J}(i_a, i_b))_{1 \leq a,b \leq k}) = \sum_{1 \leq j_1, \ldots, j_k \leq N, \text{distinct}} \mathbf{I}(j_1, \ldots, j_k \in J) \det(R_{j_1, \ldots, j_k})$$

wheree $R_{j_1,\ldots,j_k}$ is the matrix whose first row is the same as that of R_{j_1}, the second row is the same as that of R_{j_2}, and so forth. Taking expectations in J, the quantity $\mathbf{I}(j_1, \ldots, j_k \in J)$ becomes $\lambda_{j_1} \ldots \lambda_{j_k}$. Undoing the multilinearity step, we conclude that

$$\mathbf{E}_J \det(K_{V_J}(i_a, i_b))_{1 \leq a,b \leq k} = \det(\sum_{j=1}^{N} \lambda_j R_j)$$

and thus A is a determinantal process with kernel

$$K(x,y) := \sum_{j=1}^{N} \lambda_j \phi_j(x) \phi_j(y).$$

To summarise, we have created a determinantal process A whose kernel K is now an arbitrary symmetric matrix with eigenvalues $\lambda_1, \ldots, \lambda_n \in [0,1]$, and it is a mixture of constant-size processes A_{V_J}. In particular, the cardinality $|A|$ of this process has the same distribution as the cardinality $|J|$ of the random subset of $\{1, \ldots, N\}$, or in other words $|A| \equiv I_{\lambda_1} + \ldots + I_{\lambda_k}$, where $I_{\lambda_1}, \ldots, I_{\lambda_k}$ are independent Bernoulli variables with expectation $\lambda_1, \ldots, \lambda_k$ respectively.

Observe that if one takes a determinantal process $A \subset S$ with kernel K, and restricts it to a subset S' of S, then the resulting process $A \cap S' \subset S'$ is a determinantal process whose kernel K' is simply the restriction of K to the $S' \times S'$ block of $S \times S$. Applying the previous observation, we conclude that the random variable $|A \cap S'|$ has the same distribution as the sum of $|S'|$ independent Bernoulli variables, whose expectations are the eigenvalues of the restriction of K to S'. (Compare this to the Poisson point process A with some intensity measure λ, where the distribution of $|A \cap \Omega|$ is a Poisson process with intensity $\lambda(\Omega)$.) Note that most point processes do not obey this property (e.g. the uniform distribution on $\binom{S}{n}$ does not unless $n = 0, 1$ or $n = N, N-1$), and so most point processes are not determinantal.

It is known that increasing a positive semi-definite matrix by another positive semi-definite matrix does not decrease the determinant (indeed, it does not decrease any eigenvalue, by the minimax characterisation of those eigenvalues). As a consequence, if the kernel K' of a determinantal process A' is larger than the kernel K of another determinantal process A in the sense that $K - K'$ is positive semi-definite, then A' is "larger" than A in the sense that $\mathbf{P}(B \subset A') \geq \mathbf{P}(B \subset A)$ for all $B \subset S$. A particularly nice special case is when $K = cK'$ for some $0 \leq c \leq 1$, then $\mathbf{P}(B \subset A) = c^{|B|}\mathbf{P}(B \subset A')$ for all B, and one can interpret A as the process obtained from A' by deleting each element of A' independently at random with probability $1 - c$ (i.e. keeping that element independently at random with probability c).

As a consequence of this, one can obtain a converse to our previous construction of determinantal processes, and conclude that a determinantal process can be associated to a symmetric kernel K only if the eigenvalues of K lie between zero and one. The fact that K is positive semi-definite follows from the fact that all symmetric minors of K have non-negative determinant (thanks to (4.29)). Now suppose for contradiction that K has an eigenvalue larger than 1, then one can find $0 \leq c < 1$ such that the largest eigenvalue of cK is exactly 1. By our previous discussion, the process A_{cK} associated to cK is then formed from the process A_K by deleting each element of A with non-zero probability; in particular, A_K is empty with non-zero probability. On the other hand, we know that $|A_K|$ has the distribution of the sum of independent Bernoulli variables, at least one of which is 1 with probability one, a contradiction. (This proof is due to [**HoKrPeVi2006**], though the result is originally due to Soshnikov[**So2000**]. An alternate proof is to extend the identity (4.30) to all determinantal processes and conclude that $I - K$ is necessarily positive definite.)

4.6.2. Continuous determinantal processes. One can extend the theory of discrete determinantal processes to the continuous setting. For simplicity we restrict attention to (simple) point processes $A \subset \mathbf{R}$ on the real line. A process A is said to have *correlation functions* $\rho_k : \mathbf{R}^k \to \mathbf{R}$ for $k \geq 1$ if the ρ_k are symmetric, non-negative, and locally integrable, and one has the formula

$$\mathbf{E} \sum_{x_1,\ldots,x_k \in A, \text{distinct}} f(x_1,\ldots,x_k) = \int_{\mathbf{R}^k} f(x_1,\ldots,x_k)\rho_k(x_1,\ldots,x_k)\, dx_1 \ldots dx_k$$

for any bounded measurable symmetric f with compact support, where the left-hand side is summed over all k-tuples of distinct points in A (this sum is of course empty if $|A| \leq k$). Intuitively, the probability that A contains an element in the infinitesimal interval $[x_i, x_i + dx_i]$ for all $1 \leq i \leq k$ and distinct $x_1,\ldots,x_k$ is equal to $\rho_k(x_1,\ldots,x_k)dx_1 \ldots dx_k$. The ρ_k are not quite probability distributions; instead, the integral $\int_{\mathbf{R}^k} \rho_k$ is equal to $k!\mathbf{E}\binom{|A|}{k}$. Thus, for instance, if A is a constant-size process of cardinality n, then ρ_k has integral $\frac{n!}{(n-k)!}$ on $\mathbf{R}^n$ for $1 \leq k \leq n$ and vanishes for $k > n$.

If the correlation functions exist, it is easy to see that they are unique (up to almost everywhere equivalence), and can be used to compute various statistics of the process. For instance, an application of the inclusion-exclusion principle shows that for any bounded measurable set Ω, the probability that $A \cap \Omega = \emptyset$ is (formally) equal to

$$\sum_{k=0}^{\infty} \frac{(-1)^k}{k!} \int_{(\mathbf{R}\backslash\Omega)^k} \rho_k(x_1, \ldots, x_k)\ dx_1 \ldots dx_k.$$

A process is *determinantal* with some symmetric measurable kernel $K : \mathbf{R} \times \mathbf{R} \to \mathbf{R}$ if it has correlation functions ρ_k given by the formula

$$\rho_k(x_1, \ldots, x_k) = \det(K(x_i, x_j))_{1 \leq i,j \leq k}. \tag{4.31}$$

Informally, the probability that A intersects the infinitesimal intervals $[x_i, x_i + dx_i]$ for distinct $x_1, \ldots, x_k$ is $\det(K(x_i, x_j) dx_i^{1/2} dx_j^{1/2})_{1 \leq i,j \leq k}$. (Thus, K is most naturally interpreted as a half-density, or as an integral operator from $L^2(\mathbf{R})$ to $L^2(\mathbf{R})$.)

There are analogues of the discrete theory in this continuous setting. For instance, one can show that a symmetric measurable kernel K generates a determinantal process if and only if the associated integral operator $\mathcal{K}$ has spectrum lies in the interval $[0, 1]$. The analogue of (4.30) is the formula

$$\mathbf{P}(A \cap \Omega = \emptyset) = \det(I - \mathcal{K}|_\Omega);$$

more generally, the distribution of $|A \cap \Omega|$ is the sum of independent Bernoulli variables, whose expectations are the eigenvalues of $\mathcal{K}|_\Omega$. Finally, if $\mathcal{K}$ is an orthogonal projection onto an n-dimensional space, then the process has a constant size of n. Conversely, if A is a process of constant size n, whose n^{th} correlation function $\rho_n(x_1, \ldots, x_n)$ is given by (4.31), where $\mathcal{K}$ is an orthogonal projection onto an n-dimensional space, then (4.31) holds for all other values of k as well, and so A is a determinantal process with kernel K. (This is roughly the analogue of Lemma 4.6.4.)

These facts can be established either by approximating a continuous process as the limit of discrete ones, or by obtaining alternate proofs of several of the facts in the previous section which do not

rely as heavily on the discrete hypotheses. See [**HoKrPeVi2006**] for details.

A Poisson process can be viewed as the limiting case of a determinantal process in which $\mathcal{K}$ degenerates to a (normalisation of) a multiplication operator $f \mapsto \lambda f$, where λ is the intensity function.

4.6.3. The spectrum of GUE. Now we turn to a specific example of a continuous point process, namely the spectrum $A = \{\lambda_1, \ldots, \lambda_n\} \subset \mathbf{R}$ of the *Gaussian unitary ensemble* $M_n = (\zeta_{ij})_{1 \leq i,j \leq n}$, where the ζ_{ij} are independent for $1 \leq i \leq j \leq n$ with mean zero and variance 1, with ζ_{ij} being the standard complex gaussian for $i < j$ and the standard real gaussian $N(0,1)$ for $i = j$. The probability distribution of M_n can be expressed as

$$c_n \exp(-\frac{1}{2} \operatorname{trace}(M_n^2))\ dM_n$$

where dM_n is Lebesgue measure on the space of Hermitian $n \times n$ matrices, and $c_n > 0$ is some explicit normalising constant.

The n-point correlation function of A can be computed explicitly:

Lemma 4.6.9 (Ginibre formula)**.** *The n-point correlation function $\rho_n(x_1, \ldots, x_n)$ of the GUE spectrum A is given by*

$$\rho_n(x_1, \ldots, x_n) = c'_n(\prod_{1 \leq i < j \leq n} |x_i - x_j|^2) \exp(-\sum_{i=1}^n x_i^2/2) \tag{4.32}$$

where the normalising constant c'_n is chosen so that ρ_n has integral 1.

The constant $c'_n > 0$ is essentially the reciprocal of the *partition function* for this ensemble, and can be computed explicitly, but we will not do so here.

Proof. Let D be a diagonal random matrix $D = \operatorname{diag}(x_1, \ldots, x_n)$ whose entries are drawn using the distribution $\rho_n(x_1, \ldots, x_n)$ defined by (4.32), and let $U \in U(n)$ be a unitary matrix drawn uniformly at random (with respect to Haar measure on $U(n)$) and independently of D. It will suffice to show that the GUE M_n has the same probability distribution as $U^* D U$. Since probability distributions have total mass one, it suffices to show that their distributions differ up to multiplicative constants.

The distributions of M_n and U^*DU are easily seen to be continuous and invariant under unitary rotations. Thus, it will suffice to show that their probability density at a given diagonal matrix $D_0 = \mathrm{diag}(x_1^0, \ldots, x_n^0)$ are the same up to multiplicative constants. We may assume that the x_i^0 are distinct, since this occurs for almost every choice of D_0.

On the one hand, the probability density of M_n at D_0 is proportional to $\exp(-\sum_{i=1}^n (x_i^0)^2/2)$. On the other hand, a short computation shows that if U^*DU is within a distance $O(\varepsilon)$ of D_0 for some infinitesimal $\varepsilon > 0$, then (up to permutations) D must be a distance $O(\varepsilon)$ from D_0, and the ij entry of U must be a complex number of size $O(\varepsilon/|x_i^0 - x_j^0|)$ for $1 \leq i < j \leq n$, while the diagonal entries of U can be arbitrary phases. Pursuing this computation more rigorously (e.g. using the Harish-Chandra formula) and sending $\varepsilon \to 0$, one can show that the probability density of U^*DU at D_0 is a constant multiple of

$$\rho_n(x_1, \ldots, x_n) \prod_{1 \leq i < j \leq n} \frac{1}{|x_i^0 - x_j^0|^2}$$

(the square here arising because of the complex nature of the ij coefficient of U) and the claim follows. □

One can also represent the k-point correlation functions as a determinant:

Lemma 4.6.10 (Gaudin-Mehta formula)**.** *The k-point correlation function $\rho_k(x_1, \ldots, x_n)$ of the GUE spectrum A is given by*

$$\rho_k(x_1, \ldots, x_k) = \det(K_n(x_i, x_j))_{1 \leq i < j \leq k} \tag{4.33}$$

where $K_n(x, y)$ is the kernel of the orthogonal projection $\mathcal{K}$ in $L^2(\mathbf{R})$ to the space spanned by the polynomials $x^i e^{-x^2/4}$ for $i = 0, \ldots, n-1$. In other words, A is the n-point determinantal process with kernel K_n.

Proof. By the material in the preceding section, it suffices to establish this for $k = n$. As K is the kernel of an orthogonal projection to an n-dimensional space, it generates an n-point determinantal process and so $\det(K_n(x_i, x_j))_{1 \leq i < j \leq n}$ has integral $\binom{n}{n} = 1$. Thus it

will suffice to show that ρ_n and $\det(K_n(x_i,x_j))_{1\leq i<j\leq n}$ agree up to multiplicative constants.

By Gram-Schmidt, one can find an orthonormal basis $\phi_i(x)e^{-x^2/4}$, $i=0,\ldots,n-1$ for the range of $\mathcal{K}$, with each ϕ_i a polynomial of degree i (these are essentially the *Hermite polynomials*). Then we can write

$$K_n(x_i,x_j) = \sum_{k=0}^{n-1} \phi_k(x_i)\phi_k(x_j)e^{-(x_i^2+x_j^2)/4}.$$

Cofactor expansion then shows that $\det(K_n(x_i,x_j))_{1\leq i<j\leq n}$ is equal to $\exp(-\sum_{i=1}^n x_i^2/2)$ times a polynomial $P(x_1,\ldots,x_n)$ in $x_1,\ldots,x_n$ of degree at most $2\sum_{k=0}^{n-1} k = n(n-1)$. On the other hand, this determinant is always non-negative, and vanishes whenever $x_i = x_j$ for any $1\leq i<j\leq n$, and so must contain $(x_i-x_j)^2$ as a factor for all $1\leq i<j\leq n$. As the total degree of all these (relatively prime) factors is $n(n-1)$, the claim follows. □

This formula can be used to obtain asymptotics for the (renormalised) GUE eigenvalue spacings in the limit $n\to\infty$, by using asymptotics for (renormalised) Hermite polynomials; this was first established by Dyson[**Dy1970**].

Notes. This article first appeared at `terrytao.wordpress.com/2009/08/23`. Thanks to anonymous commenters for corrections.

Craig Tracy noted that some non-determinantal processes, such as TASEP, still enjoy many of the spacing distributions as their determinantal counterparts.

Manju Krishnapur raised the relevant question of how one could determine quickly whether a given process is determinantal.

Russell Lyons noted the open problem on coupling determinantal processes together was also raised in Question 10.1 of [**Ly2003**] (which also covers most of the other material in this article).

4.7. The Cohen-Lenstra distribution

At a conference recently, I learned of the recent work of Ellenberg, Venkatesh, and Westerland[**ElVeWe2009**], which concerned the conjectural behaviour of *class groups* of quadratic fields, and in particular

to explain the numerically observed phenomenon that about 75.4% of all quadratic fields $\mathbf{Q}[\sqrt{d}]$ (with d prime) enjoy unique factorisation (i.e. have trivial class group). (Class groups, as I learned at this conference, are arithmetic analogues of the (abelianised) fundamental groups in topology, with Galois groups serving as the analogue of the full fundamental group.) One thing I learned here was that there was a canonical way to randomly generate a (profinite) abelian group, by taking the product of randomly generated finite abelian *p-groups* for each prime p. The way to canonically randomly generate a finite abelian p-group is to take large integers n, d, and look at the cokernel of a random homomorphism from $(\mathbf{Z}/p^n\mathbf{Z})^d$ to $(\mathbf{Z}/p^n\mathbf{Z})^d$. In the limit $n, d \to \infty$ (or by replacing $\mathbf{Z}/p^n\mathbf{Z}$ with the p-adics and just sending $d \to \infty$), this stabilises and generates any given p-group G with probability

$$\frac{1}{|\operatorname{Aut}(G)|} \prod_{j=1}^{\infty} (1 - \frac{1}{p^j}), \tag{4.34}$$

where $\operatorname{Aut}(G)$ is the group of automorphisms of G. In particular this leads to the strange identity

$$\sum_G \frac{1}{|\operatorname{Aut}(G)|} = \prod_{j=1}^{\infty} (1 - \frac{1}{p^j})^{-1} \tag{4.35}$$

where G ranges over all p-groups; I do not know how to prove this identity other than via the above probability computation, the proof of which I give below.

Based on the heuristic that the class group should behave "randomly" subject to some "obvious" constraints, it is expected that a randomly chosen real quadratic field $\mathbf{Q}[\sqrt{d}]$ has unique factorisation (i.e. the class group has trivial p-group component for every p) with probability

$$\prod_{p \text{ odd}} \prod_{j=2}^{\infty} (1 - \frac{1}{p^j}) \approx 0.754,$$

whereas a randomly chosen imaginary quadratic field $\mathbf{Q}[\sqrt{-d}]$ has unique factorisation with probability

$$\prod_{p \text{ odd}} \prod_{j=1}^{\infty} (1 - \frac{1}{p^j}) = 0.$$

The former claim is conjectural, whereas the latter claim follows from (for instance) Siegel's theorem on the size of the class group, as discussed in Section 3.12.4. The work in [**ElVeWe2009**] establishes some partial results towards the function field analogues of these heuristics.

4.7.1. p-groups. Henceforth the prime p will be fixed. We will abbreviate "finite abelian p-group" as "p-group" for brevity. Thanks to the *classification of finite abelian groups*, the p-groups are all isomorphic to the products

$$(\mathbf{Z}/p^{n_1}\mathbf{Z}) \times \ldots \times (\mathbf{Z}/p^{n_d}\mathbf{Z})$$

of cyclic p-groups.

The cokernel of a random homomorphism from $(\mathbf{Z}/p^n\mathbf{Z})^d$ to $(\mathbf{Z}/p^n\mathbf{Z})^d$ can be written as the quotient of the p-group $(\mathbf{Z}/p^n\mathbf{Z})^d$ by the subgroup generated by d randomly chosen elements $x_1, \ldots, x_d$ from that p-group. One can view this quotient as a d-fold iterative process, in which one starts with the p-group $(\mathbf{Z}/p^n\mathbf{Z})^d$, and then one iterates d times the process of starting with a p-group G, and quotienting out by a randomly chosen element x of that group G. From induction, one sees that at the j^{th} stage of this process ($0 \leq j \leq d$), one ends up with a p-group isomorphic to $(\mathbf{Z}/p^n\mathbf{Z})^{d-j} \times G_j$ for some p-group G_j.

Let's see how the group $(\mathbf{Z}/p^n\mathbf{Z})^{d-j} \times G_j$ transforms to the next group $(\mathbf{Z}/p^n\mathbf{Z})^{d-j-1} \times G_{j+1}$. We write a random element of $(\mathbf{Z}/p^n\mathbf{Z})^{d-j} \times G_j$ as (x, y), where $x \in (\mathbf{Z}/p^n\mathbf{Z})^{d-j}$ and $y \in G_j$. Observe that for any $0 \leq i < n$, x is a multiple of p^i (but not p^{i+1}) with probability $(1 - p^{-(d-j)})p^{-i(d-j)}$. (The remaining possibility is that x is zero, but this event will have negligible probability in the limit $n \to \infty$.) If x is indeed divisible by p^i but not p^{i+1}, and i is not too close to n, a little thought will then reveal that $|G_{j+1}| = p^i|G_j|$. Thus the size of the p-groups G_j only grow as j increases. (Things go wrong when i gets close to n, e.g. $p^i \geq p^n/|G_j|$, but the total size of this event as j ranges from 0 to d sums to be $o(1)$ as $n \to \infty$ (uniformly in d), by using the tightness bounds on $|G_j|$ mentioned below. Alternatively, one can avoid a lot of technicalities by taking the limit $n \to \infty$ before taking the limit $d \to \infty$ (instead of studying

the double limit $n, d \to \infty$), or equivalently by replacing the cyclic group $\mathbf{Z}/p^n\mathbf{Z}$ with the p-adics $\mathbf{Z}_p$.)

The exponentially decreasing nature of the probability $(1-p^{-(d-j)})p^{-i(d-j)}$ in i (and in $d-j$) furthermore implies that the distribution of $|G_j|$ forms a *tight sequence* in n, j, d: for every $\varepsilon > 0$, one has an $R > 0$ such that the probability that $|G_j| \geq R$ is less than ε for all choices of n, j, d. (This tightness is necessary to prove the equality in (4.35) rather than just an inequality (from Fatou's lemma).) Indeed, the probability that $|G_j| = p^m$ converges as $n, d \to \infty$ to the t^m coefficient in the generating function

$$\prod_{k=1}^{\infty} \sum_{i=0}^{\infty} t^i (1-p^{-k}) p^{-ik} = \prod_{k=1}^{\infty} \frac{1-p^{-k}}{1-tp^{-k}}. \tag{4.36}$$

In particular, this claim is true for the final cokernel G_d. Note that this (and the geometric series formula) already yields (4.34) in the case of the trivial group $G = \{0\}$ and the order p group $G = \mathbf{Z}/p\mathbf{Z}$ (note that $\mathrm{Aut}(G)$ has order 1 and p in these respective cases). But it is not enough to deal with higher groups. For instance, up to isomorphism there are two p-groups of order p^2, namely $\mathbf{Z}/p^2\mathbf{Z}$ and $(\mathbf{Z}/p\mathbf{Z})^2$, whose automorphism group has order p^2-p and $(p^2-1)(p^2-p)$ respectively. Summing up the corresponding two expressions (4.34) one can observe that this matches the t^2 coefficient of (4.36) (after some applications of the geometric series formula). Thus we see that (4.36) is consistent with the claim (4.34), but does not fully imply that claim.

To get the full asymptotic (4.34) we try a slightly different tack. Fix a p-group G, and consider the event that the cokernel of a random map $T : (\mathbf{Z}/p^n\mathbf{Z})^d \to (\mathbf{Z}/p^n\mathbf{Z})^d$ is isomorphic to G. We assume n so large that all elements in G have order at most p^n. If this is the case, then there must be a surjective homomorphism $\phi : (\mathbf{Z}/p^n\mathbf{Z})^d \to G$ such that the range of T is equal to the kernel of ϕ. The number of homomorphisms from $(\mathbf{Z}/p^n\mathbf{Z})^d$ to G is $|G|^d$ (one has to pick d generators in G). If d is large, it is easy to see that most of these homomorphisms are surjective (the proportion of such homomorphisms is $1 - o(1)$ as $d \to \infty$). On the other hand, there is some multiplicity; the range of T can emerge as the kernel of ϕ in $|\mathrm{Aut}(G)|$ different ways (since any two surjective homomorphisms $\phi, \phi' : (\mathbf{Z}/p^n\mathbf{Z})^d \to G$ with the same kernel arise from an automorphism of G). So to

prove (4.34), it suffices to show that for any surjective homomorphism $\phi : (\mathbf{Z}/p^n\mathbf{Z})^d \to G$, the probability that the range of T equals the kernel of ϕ is

$$(1+o(1))|G|^{-d}\prod_{j=1}^{\infty}(1-\frac{1}{p^j}).$$

The range of T is the same thing as the subgroup of $(\mathbf{Z}/p^n\mathbf{Z})^d$ generated by d random elements $x_1,\ldots,x_d$ of that group. The kernel of ϕ has index $|G|$ inside $(\mathbf{Z}/p^n\mathbf{Z})^d$, so the probability that all of those random elements lie in the kernel of ϕ is $|G|^{-d}$. So it suffices to prove the following claim: if ϕ is a fixed surjective homomorphism from $(\mathbf{Z}/p^n\mathbf{Z})^d$ to G, and $x_1,\ldots,x_d$ are chosen randomly from the kernel of ϕ, then $x_1,\ldots,x_d$ will generate that kernel with probability

$$(1+o(1))\prod_{j=1}^{\infty}(1-\frac{1}{p^j}). \tag{4.37}$$

But from the classification of p-groups, the kernel of ϕ (which has bounded index inside $(\mathbf{Z}/p^n\mathbf{Z})^d$) is isomorphic to

$$(\mathbf{Z}/p^{n-O(1)}\mathbf{Z})\times\ldots\times(\mathbf{Z}/p^{n-O(1)}\mathbf{Z}) \tag{4.38}$$

where $O(1)$ means "bounded uniformly in n", and there are d factors here. As in the previous argument, one can now imagine starting with the group (4.38), and then iterating d times the operation of quotienting out by the group generated by a randomly chosen element; our task is to compute the probability that one ends up with the trivial group by applying this process.

As before, at the j^{th} stage of the iteration, one ends up with a group of the form

$$(\mathbf{Z}/p^{n-O(1)}\mathbf{Z})\times\ldots\times(\mathbf{Z}/p^{n-O(1)}\mathbf{Z})\times G_j \tag{4.39}$$

where there are $d-j$ factors of $(\mathbf{Z}/p^{n-O(1)}\mathbf{Z})$. The group G_j is increasing in size, so the only way in which one ends up with the trivial group is if all the G_j are trivial. But if G_j is trivial, the only way that G_{j+1} is trivial is if the randomly chosen element from (4.39) has a $(\mathbf{Z}/p^{n-O(1)}\mathbf{Z})\times\ldots\times(\mathbf{Z}/p^{n-O(1)}\mathbf{Z})$ component which is invertible (i.e. not a multiple of p), which occurs with probability $1-p^{-(d-j)}$

(assuming n is large enough). Multiplying all these probabilities together gives (4.37).

Notes. This article first appeared at `terrytao.wordpress.com/2009/10/02`. Thanks to David Speyer and an anonymous commenter for corrections.

4.8. An entropy Plünnecke-Ruzsa inequality

A handy inequality in additive combinatorics is the *Plünnecke-Ruzsa inequality*[**Ru1989**]:

Theorem 4.8.1 (Plünnecke-Ruzsa inequality)**.** *Let $A, B_1, \ldots, B_m$ be finite non-empty subsets of an additive group G, such that $|A+B_i| \leq K_i|A|$ for all $1 \leq i \leq m$ and some scalars $K_1, \ldots, K_m \geq 1$. Then there exists a subset A' of A such that $|A' + B_1 + \ldots + B_m| \leq K_1 \ldots K_m|A'|$.*

The proof uses graph-theoretic techniques. Setting $A = B_1 = \ldots = B_m$, we obtain a useful corollary: if A has small doubling in the sense that $|A + A| \leq K|A|$, then we have $|mA| \leq K^m|A|$ for all $m \geq 1$, where $mA = A + \ldots + A$ is the sum of m copies of A.

In a recent paper[**Ta2010c**], I adapted a number of sum set estimates to the entropy setting, in which finite sets such as A in G are replaced with discrete random variables X taking values in G, and (the logarithm of) cardinality $|A|$ of a set A is replaced by *Shannon entropy* $\mathbf{H}(X)$ of a random variable X. (Throughout this note I assume all entropies to be finite.) However, at the time, I was unable to find an entropy analogue of the Plünnecke-Ruzsa inequality, because I did not know how to adapt the graph theory argument to the entropy setting.

I recently discovered, however, that buried in a classic paper[**KaVe1983**] of Kaimonovich and Vershik (implicitly in Proposition 1.3, to be precise) there was the following analogue of Theorem 4.8.1:

Theorem 4.8.2 (Entropy Plünnecke-Ruzsa inequality)**.** *Let $X, Y_1, \ldots, Y_m$ be independent random variables of finite entropy taking values in an additive group G, such that $\mathbf{H}(X+Y_i) \leq \mathbf{H}(X)+\log K_i$ for all $1 \leq i \leq$*

m and some scalars $K_1, \ldots, K_m \geq 1$. Then $\mathbf{H}(X + Y_1 + \ldots + Y_m) \leq \mathbf{H}(X) + \log K_1 \ldots K_m$.

In fact Theorem 4.8.2 is a bit "better" than Theorem 4.8.1 in the sense that Theorem 4.8.1 needed to refine the original set A to a subset A', but no such refinement is needed in Theorem 4.8.2. One corollary of Theorem 4.8.2 is that if $\mathbf{H}(X_1 + X_2) \leq \mathbf{H}(X) + \log K$, then $\mathbf{H}(X_1 + \ldots + X_m) \leq \mathbf{H}(X) + (m-1)\log K$ for all $m \geq 1$, where $X_1, \ldots, X_m$ are independent copies of X; this improves slightly over the analogous combinatorial inequality. Indeed, the function $m \mapsto \mathbf{H}(X_1 + \ldots + X_m)$ is concave (this can be seen by using the $m = 2$ version of Theorem 4.8.2 (or (4.41) below) to show that the quantity $\mathbf{H}(X_1 + \ldots + X_{m+1}) - \mathbf{H}(X_1 + \ldots + X_m)$ is decreasing in m).

Theorem 4.8.2 is actually a quick consequence of the *submodularity inequality*

$$\mathbf{H}(W) + \mathbf{H}(X) \leq \mathbf{H}(Y) + \mathbf{H}(Z) \tag{4.40}$$

in information theory, which is valid whenever X, Y, Z, W are discrete random variables such that Y and Z each determine X (i.e. X is a function of Y, and also a function of Z), and Y and Z jointly determine W (i.e W is a function of Y and Z). To apply this, let X, Y, Z be independent discrete random variables taking values in G. Observe that the pairs $(X, Y + Z)$ and $(X + Y, Z)$ each determine $X + Y + Z$, and jointly determine (X, Y, Z). Applying (4.40) we conclude that

$$\mathbf{H}(X, Y, Z) + \mathbf{H}(X + Y + Z) \leq \mathbf{H}(X, Y + Z) + \mathbf{H}(X + Y, Z)$$

which after using the independence of X, Y, Z simplifies to the *sumset submodularity inequality*

$$\mathbf{H}(X + Y + Z) + \mathbf{H}(Y) \leq \mathbf{H}(X + Y) + \mathbf{H}(Y + Z) \tag{4.41}$$

(this inequality was also recently observed *http://www.stat.yale.edu/ mm888/Pubs/2008/ITW-sums08.pdf* by Madiman; it is the $m = 2$ case of Theorem 4.8.2). As a corollary of this inequality, we see that if $\mathbf{H}(X + Y_i) \leq \mathbf{H}(X) + \log K_i$, then

$$\mathbf{H}(X + Y_1 + \ldots + Y_i) \leq \mathbf{H}(X + Y_1 + \ldots + Y_{i-1}) + \log K_i,$$

and Theorem 4.8.2 follows by telescoping series.

The proof of Theorem 4.8.2 seems to be genuinely different from the graph-theoretic proof of Theorem 4.8.1. It would be interesting to see if the above argument can be somehow adapted to give a stronger version of Theorem 4.8.1. Note also that both Theorem 4.8.1 and Theorem 4.8.2 have extensions to more general combinations of $X, Y_1, \ldots, Y_m$ than $X + Y_i$; see [**GyMaRu2008**] and and *madiman* respectively.

It is also worth remarking that the above inequalities largely carry over to the non-abelian setting. For instance, if $X_1, X_2, \ldots$ are iid copies of a discrete random variable in a multiplicative group G, the above arguments show that the function $m \mapsto \mathbf{H}(X_1 \ldots X_m)$ is concave. In particular, the expression $\frac{1}{m}\mathbf{H}(X_1 \ldots X_m)$ decreases monotonically to a limit, the *asymptotic entropy* $\mathbf{H}(G, X)$. This quantity plays an important role in the theory of bounded harmonic functions on G, as observed by [**KaVe1983**]:

Proposition 4.8.3. *Let G be a discrete group, and let X be a discrete random variable in G with finite entropy, whose support generates G. Then there exists a non-constant bounded function $f : G \to \mathbf{R}$ which is harmonic with respect to X (which means that $\mathbf{E}f(Xx) = f(x)$ for all $x \in G$) if and only if $\mathbf{H}(G, X) \neq 0$.*

Proof. (Sketch) Suppose first that $\mathbf{H}(G, X) = 0$, then we see from concavity that the successive differences $\mathbf{H}(X_1 \ldots X_m) - \mathbf{H}(X_1 \ldots X_{m-1})$ converge to zero. From this it is not hard to see that the *mutual information*

$$\mathbf{I}(X_m, X_1 \ldots X_m) := \mathbf{H}(X_m) + \mathbf{H}(X_1 \ldots X_m) - \mathbf{H}(X_m | X_1 \ldots X_m)$$

goes to zero as $m \to \infty$. Informally, knowing the value of X_m reveals very little about the value of $X_1 \ldots X_m$ when m is large.

Now let $f : G \to \mathbf{R}$ be a bounded harmonic function, and let m be large. For any $x \in G$ and any value s in the support of X_m, we observe from harmonicity that

$$f(sx) = \mathbf{E}(f(X_1 \ldots X_m x) | X_m = s).$$

But from the asymptotic vanishing of mutual information and the boundedness of f, one can show that the right-hand side will converge

to $\mathbf{E}(f(X_1 \dots X_m x))$, which by harmonicity is equal to $f(x)$. Thus f is invariant with respect to the support of X, and is thus constant since this support generates G.

Conversely, if $\mathbf{H}(G, X)$ is non-zero, then the above arguments show that $\mathbf{I}(X_m, X_1 \dots X_m)$ stays bounded away from zero as $m \to \infty$, thus $X_1 \dots X_m$ reveals a non-trivial amount of information about X_m. This turns out to be true even if m is not deterministic, but is itself random, varying over some medium-sized range. From this, one can find a bounded function F such that the conditional expectation $\mathbf{E}(F(X_1 \dots X_m)|X_m = s)$ varies non-trivially with s. On the other hand, the bounded function $x \mapsto \mathbf{E}F(X_1 \dots X_{m-1}x)$ is approximately harmonic (because we are varying m), and has some non-trivial fluctuation near the identity (by the preceding sentence). Taking a limit as $m \to \infty$ (using Arzelá-Ascoli) we obtain a non-constant bounded harmonic function as desired. □

Notes. This article first appeared at `terrytao.wordpress.com/2009/10/27`. Thanks to Seva Lev and an anonymous commenter for corrections.

4.9. An elementary noncommutative Freiman theorem

Let X be a finite subset of a non-commutative group G. As mentioned in Section 3.2 of *Structure and Randomness*,, there is some interest in classifying those X which obey *small doubling* conditions such as $|X \cdot X| = O(|X|)$ or $|X \cdot X^{-1}| = O(|X|)$. A full classification here has still not been established. However, I wanted to record here an elementary argument of Freiman [**Fr1973**] (see also [**TaVu2006b**, Exercise 2.6.5], which in turn is based on an argument in [**La2001**]) that handles the case when $|X \cdot X|$ is very close to $|X|$:

Proposition 4.9.1. *If $|X^{-1} \cdot X| < \frac{3}{2}|X|$, then $X \cdot X^{-1}$ and $X^{-1} \cdot X$ are both finite groups, which are conjugate to each other. In particular, X is contained in the right-coset (or left-coset) of a group of order less than $\frac{3}{2}|X|$.*

Remark 4.9.2. The constant $\frac{3}{2}$ is completely sharp; consider the case when $X = \{e, x\}$ where e is the identity and x is an element of

order larger than 2. This is a small example, but one can make it as large as one pleases by taking the direct product of X and G with any finite group. In the converse direction, we see that whenever X is contained in the right-coset $S \cdot x$ (resp. left-coset $x \cdot S$) of a group of order less than $2|X|$, then $X \cdot X^{-1}$ (resp. $X^{-1} \cdot X$) is necessarily equal to all of S, by the inclusion-exclusion principle (see the proof below for a related argument).

Proof. We begin by showing that $S := X \cdot X^{-1}$ is a group. As S is symmetric and contains the identity, it suffices to show that this set is closed under addition.

Let $a, b \in S$. Then we can write $a = xy^{-1}$ and $b = zw^{-1}$ for $x, y, z, w \in X$. If y were equal to z, then $ab = xw^{-1} \in X \cdot X^{-1}$ and we would be done. Of course, there is no reason why y should equal z; but we can use the hypothesis $|X^{-1} \cdot X| < \frac{3}{2}|X|$ to boost this as follows. Observe that $x^{-1} \cdot X$ and $y^{-1} \cdot X$ both have cardinality $|X|$ and lie inside $X^{-1} \cdot X$, which has cardinality strictly less than $\frac{3}{2}|X|$. By the inclusion-exclusion principle, this forces $x^{-1} \cdot X \cap y^{-1} \cdot X$ to have cardinality greater than $\frac{1}{2}|X|$. In other words, there exist more than $\frac{1}{2}|X|$ pairs $x', y' \in X$ such that $x^{-1}x' = y^{-1}y'$, which implies that $a = x'(y')^{-1}$. Thus there are more than $\frac{1}{2}|X|$ elements $y' \in X$ such that $a = x'(y')^{-1}$ for some $x' \in X$ (since x' is uniquely determined by y'); similarly, there exists more than $\frac{1}{2}|X|$ elements $z' \in X$ such that $b = z'(w')^{-1}$ for some $w' \in X$. Again by inclusion-exclusion, we can thus find $y' = z'$ in X for which one has simultaneous representations $a = x'(y')^{-1}$ and $b = y'(z')^{-1}$, and so $ab = x'(z')^{-1} \in X \cdot X^{-1}$, and the claim follows.

In the course of the above argument we showed that every element of the group S has more than $\frac{1}{2}|X|$ representations of the form xy^{-1} for $x, y \in X$. But there are only $|X|^2$ pairs (x, y) available, and thus $|S| < 2|X|$.

Now let x be any element of X. Since $X \cdot x^{-1} \subset S$, we have $X \subset S \cdot x$, and so $X^{-1} \cdot X \subset x^{-1} \cdot S \cdot x$. Conversely, every element of $x^{-1} \cdot S \cdot x$ has exactly $|S|$ representations of the form $z^{-1}w$ where $z, w \in S \cdot x$. Since X occupies more than half of $S \cdot x$, we thus see from the inclusion-exclusion principle, there is thus at least one

representation $z^{-1}w$ for which z, w both lie in X. In other words, $x^{-1} \cdot S \cdot x = X^{-1} \cdot X$, and the claim follows. □

To relate this to the classical doubling constants $|X \cdot X|/|X|$, we first make an easy observation:

Lemma 4.9.3. *If $|X \cdot X| < 2|X|$, then $X \cdot X^{-1} = X^{-1} \cdot X$.*

Again, this is sharp; consider X equal to $\{x, y\}$ where x, y generate a free group.

Proof. Suppose that xy^{-1} is an element of $X \cdot X^{-1}$ for some $x, y \in X$. Then the sets $X \cdot x$ and $X \cdot y$ have cardinality $|X|$ and lie in $X \cdot X$, so by the inclusion-exclusion principle, the two sets intersect. Thus there exist $z, w \in X$ such that $zx = wy$, thus $xy^{-1} = z^{-1}w \in X^{-1} \cdot X$. This shows that $X \cdot X^{-1}$ is contained in $X^{-1} \cdot X$. The converse inclusion is proven similarly. □

Proposition 4.9.4. *If $|X \cdot X| < \frac{3}{2}|X|$, then $S := X \cdot X^{-1}$ is a finite group of order $|X \cdot X|$, and $X \subset S \cdot x = x \cdot S$ for some x in the normaliser of S.*

The factor $\frac{3}{2}$ is sharp, by the same example used to show sharpness of Proposition 4.9.1. However, there seems to be some room for further improvement if one weakens the conclusion a bit; see below the fold.

Proof. Let $S = X^{-1} \cdot X = X \cdot X^{-1}$ (the two sets being equal by Lemma 4.9.3). By the argument used to prove Lemma 4.9.3, every element of S has more than $\frac{1}{2}|X|$ representations of the form xy^{-1} for $x, y \in X$. By the argument used to prove Proposition 4.9.1, this shows that S is a group; also, since there are only $|X|^2$ pairs (x, y), we also see that $|S| < 2|X|$.

Pick any $x \in X$; then $x^{-1} \cdot X, X \cdot x^{-1} \subset S$, and so $X \subset x \cdot S, S \cdot x$. Because every element of $x \cdot S \cdot x$ has $|S|$ representations of the form yz with $y \in x \cdot S$, $z \in S \cdot x$, and X occupies more than half of $x \cdot S$ and of $S \cdot x$, we conclude that each element of $x \cdot S \cdot x$ lies in $X \cdot X$, and so $X \cdot X = x \cdot S \cdot x$ and $|S| = |X \cdot X|$.

The intersection of the groups S and $x \cdot S \cdot x^{-1}$ contains $X \cdot x^{-1}$, which is more than half the size of S, and so we must have $S = x \cdot S \cdot x^{-1}$, i.e. x normalises S, and the proposition follows. □

Because the arguments here are so elementary, they extend easily to the infinitary setting in which X is now an infinite set, but has finite measure with respect to some translation-invariant Kiesler measure μ. We omit the details. (I am hoping that this observation may help simplify some of the theory in that setting.)

4.9.1. Beyond the 3/2 barrier. It appears that one can push the arguments a bit beyond the 3/2 barrier, though of course one has to weaken the conclusion in view of the counterexample in Remark 4.9.2. Here I give a result that increases $3/2 = 1.5$ to the golden ratio $\phi := (1+\sqrt{5})/2 = 1.618\ldots$:

Proposition 4.9.5 (Weak non-commutative Kneser theorem)**.** *If $|X^{-1} \cdot X|, |X \cdot X^{-1}| \leq K|X|$ for some $1 < K < \phi$, then $X \cdot X^{-1} = H \cdot Z$ for some finite subgroup H, and some finite set Z with $|Z| \leq C(K)$ for some $C(K)$ depending only on K.*

Proof. Write $S := X \cdot X^{-1}$. Let us say that h *symmetrises* S if $h \cdot S = S$, and let H be the set of all h that symmetrise S. It is clear that H is a finite group with $H \cdot S = S$ and thus $S \cdot H = S$ also.

For each $z \in S$, let $r(z)$ be the number of representations of z of the form $z = xy^{-1}$ with $x, y \in X$. Double counting shows that $\sum_{z \in S} r(z) = |X|^2$, while by hypothesis $|S| \leq K|X|$; thus the average value of $r(z)$ is at least $|X|/K$. Since $1 < K < \phi$, $1/K > K-1$. Since $r(z) \leq |X|$ for all z, we conclude that $r(z) > (K-1)|X|$ for at least $c(K)|X|$ values of $z \in S$, for some explicitly computable $c(K) > 0$.

Suppose $z, w \in S$ is such that $r(z) > (K-1)|X|$, thus z has more than $(K-1)|X|$ representations of the form xy^{-1} with $x, y \in X$. On the other hand, the argument used to prove Proposition 4.9.1 shows that w has at least $(2-K)|X|$ representations of the form $x'(y')^{-1}$ with $x', y' \in X$. By the inclusion-exclusion formula, we can thus find representations for which $y = x'$, which implies that $zw \in S$. Since $w \in S$ was arbitrary, this implies that $z \in H$. Thus $|H| \geq c(K)|X|$. Since $S = H \cdot S$ and $|S| \leq K|X|$, this implies that S can be covered

by at most $C(K)$ right-cosets of S for some $C(K)$ depending only on K, and the claim follows. □

This result appears in [**Fr1973**], and a related argument also appears in [**Le2000**].

It looks like one should be able to get a bit more structural information on X than is given by the above conclusion, and I doubt the golden ratio is sharp either (the correct threshold should be 2, in analogy with the commutative Kneser theorem; after that, the conclusion will fail, as can be seen by taking X to be a long geometric progression). Readers here are welcome to look for improvements to these results, of course.

Notes. This article first appeared at `terrytao.wordpress.com/2009/11/10`. Thanks to Miguel Lacruz for corrections, and Ben Green and Seva Lev for references.

4.10. Nonstandard analogues of energy and density increment arguments

This article assumes some familiarity with nonstandard analysis (see e.g. Section 1.5 of *Structure and Randomness*).

Let us call a model M of a language L *weakly countably saturated*[2] if, every countable sequence $P_1(x), P_2(x), \ldots$ of formulae in L (involving countably many constants in M) which is finitely satisfiable in M (i.e. any finite collection $P_1(x), \ldots, P_n(x)$ in the sequence has a solution x in M), is automatically satisfiable in M (i.e. there is a solution x to all $P_n(x)$ simultaneously). Equivalently, a model is weakly countably saturated if the topology generated by the definable sets is *countably compact*.

Most models are not (weakly) countably saturated. Consider for instance the standard natural numbers $\mathbf{N}$ as a model for arithmetic.

[2]The stronger property of being *countably saturated* asserts that if an *arbitrary* sequence of formulae involving countably many constants is finitely satisfiable, then it is satisfiable; the relation between the two concepts is thus analogous to compactness and countable compactness. If one chooses a special type of ultrafilter, namely a "countably incomplete" ultrafilter, one can recover the full strength of countable saturation, though it is not needed for the remarks here.

Then the sequence of formulae "$x > n$" for $n = 1, 2, 3, \ldots$ is finitely satisfiable in $\mathbf{N}$, but not satisfiable.

However, if one takes a model M of L and passes to an *ultrapower* $*M$, whose elements x consist of sequences $(x_n)_{n\in\mathbf{N}}$ in M, modulo equivalence with respect to some fixed non-principal ultrafilter p, then it turns out that such models are automatically weakly countably saturated. Indeed, if $P_1(x), P_2(x), \ldots$ are finitely satisfiable in $*M$, then they are also finitely satisfiable in M (either by inspection, or by appeal to *Łos's theorem* and/or the *transfer principle* in non-standard analysis), so for each n there exists $x_n \in M$ which satisfies $P_1, \ldots, P_n$. Letting $x = (x_n)_{n\in\mathbf{N}} \in *M$ be the ultralimit of the x_n, we see that x satisfies all of the P_n at once.

In particular, non-standard models of mathematics, such as the non-standard model $*\mathbf{N}$ of the natural numbers, are automatically countably saturated. (This fact is closely related to the *idealisation axiom* in internal set theory.)

This has some cute consequences. For instance, suppose one has a non-standard metric space $*X$ (an ultralimit of standard metric spaces), and suppose one has a standard sequence $(x_n)_{n\in\mathbf{N}}$ of elements of $*X$ which are standard-Cauchy, in the sense that for any standard $\varepsilon > 0$ one has $d(x_n, x_m) < \varepsilon$ for all sufficiently large n, m. Then there exists a non-standard element $x \in *X$ such that x_n standard-converges to x in the sense that for every standard $\varepsilon > 0$ one has $d(x_n, x) < \varepsilon$ for all sufficiently large n. Indeed, from the standard-Cauchy hypothesis, one can find a standard $\varepsilon(n) > 0$ for each standard n that goes to zero (in the standard sense), such that the formulae "$d(x_n, x) < \varepsilon(n)$" are finitely satisfiable, and hence satisfiable by countable saturation. Thus we see that non-standard metric spaces are automatically "standardly complete" in some sense.

This leads to a non-standard structure theorem for Hilbert spaces, analogous to the orthogonal decomposition in Hilbert spaces:

Theorem 4.10.1 (Non-standard structure theorem for Hilbert spaces)**.** *Let $*H$ be a non-standard Hilbert space, let S be a bounded (external) subset of $*H$, and let $x \in H$. Then there exists a decomposition $x = x_S + x_{S^\perp}$, where $x_S \in *H$ is "almost standard-generated by*

S" in the sense that for every standard $\varepsilon > 0$, there exists a standard finite linear combination of elements of S which is within ε of S, and $x_{S^\perp} \in *H$ is "standard-orthogonal to S" in the sense that $\langle x_{S^\perp}, s\rangle = o(1)$ for all $s \in S$.

Proof. Let d be the infimum of all the (standard) distances from x to a standard linear combination of elements of S, then for every standard n one can find a standard linear combination x_n of elements of S which lie within $d + 1/n$ of x. From the parallelogram law we see that x_n is standard-Cauchy, and thus standard-converges to some limit $x_S \in *H$, which is then almost standard-generated by S by construction. An application of Pythagoras then shows that $x_{S^\perp} := x - x_S$ is standard-orthogonal to every element of S. □

This is the non-standard analogue of a combinatorial structure theorem for Hilbert spaces (see e.g. [**Ta2007b**, Theorem 2.6]). There is an analogous non-standard structure theorem for σ-algebras (the counterpart of [**Ta2007b**, Theorem 3.6]) which I will not discuss here, but I will give just one sample corollary:

Theorem 4.10.2 (Non-standard arithmetic regularity lemma)**.** *Let $*G$ be a non-standardly finite abelian group, and let $f : *G \to [0,1]$ be a function. Then one can split $f = f_{U^\perp} + f_U$, where $f_U : *G \to [-1,1]$ is standard-uniform in the sense that all Fourier coefficients are (uniformly) $o(1)$, and $f_{U^\perp} : *G \to [0,1]$ is standard-almost periodic in the sense that for every standard $\varepsilon > 0$, one can approximate $f_{U^\perp}$ to error ε in $L^1(*G)$ norm by a standard linear combination of characters (which is also bounded).*

This can be used for instance to give a non-standard proof of Roth's theorem (which is not much different from the "finitary ergodic" proof of Roth's theorem, given for instance in [**TaVu2006b**, Section 10.5]). There is also a non-standard version of the Szemerédi regularity lemma which can be used, among other things, to prove the hypergraph removal lemma (the proof then becomes rather close to the infinitary proof of this lemma in [**Ta2007**]). More generally, the above structure theorem can be used as a substitute for various "energy increment arguments" in the combinatorial literature, though it

does not seem that there is a significant saving in complexity in doing so unless one is performing quite a large number of these arguments.

One can also cast density increment arguments in a nonstandard framework. Here is a typical example. Call a non-standard subset X of a non-standard finite set Y *dense* if one has $|X| \geq \varepsilon|Y|$ for some standard $\varepsilon > 0$.

Theorem 4.10.3. *Suppose Szemerédi's theorem (every set of integers of positive upper density contains an arithmetic progression of length k) fails for some k. Then there exists an unbounded non-standard integer N, a dense subset A of $[N] := \{1, \ldots, N\}$ with no progressions of length k, and with the additional property that*

$$\frac{|A \cap P|}{|P|} \leq \frac{|A \cap [N]|}{N} + o(1)$$

for any subprogression P of $[N]$ of unbounded size (thus there is no sizeable density increment on any large progression).

Proof. Let $B \subset \mathbf{N}$ be a (standard) set of positive upper density which contains no progression of length k. Let $\delta := \limsup_{|P|\to\infty} |B \cap P|/|P|$ be the asymptotic maximal density of B inside a long progression, thus $\delta > 0$. For any $n > 0$, one can then find a standard integer $N_n \geq n$ and a standard subset A_n of $[N_n]$ of density at least $\delta - 1/n$ such that A_n can be embedded (after a linear transformation) inside B, so in particular A_n has no progressions of length k. Applying the saturation property, one can then find an unbounded N and a set A of $[N]$ of density at least $\delta - 1/n$ for every standard n (i.e. of density at least $\delta - o(1)$) with no progressions of length k. By construction, we also see that for any subprogression P of $[N]$ of unbounded size, A hs density at most $\delta + 1/n$ for any standard n, thus has density at most $\delta + o(1)$, and the claim follows. □

This can be used as the starting point for any density-increment based proof of Szemerédi's theorem for a fixed k, e.g. Roth's proof for $k = 3$, Gowers' proof for arbitrary k, or Szemerédi's proof for arbitrary k. (It is likely that Szemerédi's proof, in particular, simplifies a little bit when translated to the non-standard setting, though the savings are likely to be modest.)

I'm also hoping that the recent results of Hrushovski[**Hr2009**] on the noncommutative Freiman problem require only countable saturation, as this makes it more likely that they can be translated to a non-standard setting and thence to a purely finitary framework.

Notes. This article first appeared at `terrytao.wordpress.com/2009/11/10`. Balazs Szegedy noted the connection to his recent work [**Sz2009**] on higher order Fourier analysis from a nonstandard perspective.

4.11. Approximate bases, sunflowers, and nonstandard analysis

One of the most basic theorems in linear algebra is that every finite-dimensional vector space has a finite basis. Let us give a statement of this theorem in the case when the underlying field is the rationals:

Theorem 4.11.1 (Finite generation implies finite basis, infinitary version)**.** *Let V be a vector space over the rationals $\mathbf{Q}$, and let $v_1, \ldots, v_n$ be a finite collection of vectors in V. Then there exists a collection $w_1, \ldots, w_k$ of vectors in V, with $1 \leq k \leq n$, such that*

- *(w generates v) Every v_j can be expressed as a rational linear combination of the $w_1, \ldots, w_k$.*
- *(w independent) There is no non-trivial linear relation $a_1 w_1 + \ldots + a_k w_k = 0$, $a_1, \ldots, a_k \in \mathbf{Q}$ among the $w_1, \ldots, w_m$ (where non-trivial means that the a_i are not all zero).*

In fact, one can take $w_1, \ldots, w_m$ to be a subset of the $v_1, \ldots, v_n$.

Proof. We perform the following "rank reduction argument". Start with $w_1, \ldots, w_k$ initialised to $v_1, \ldots, v_n$ (so initially we have $k = n$). Clearly w generates v. If the w_i are linearly independent then we are done. Otherwise, there is a non-trivial linear relation between them; after shuffling things around, we see that one of the w_i, say w_k, is a rational linear combination of the $w_1, \ldots, w_{k-1}$. In such a case, w_k becomes redundant, and we may delete it (reducing the rank k by one). We repeat this procedure; it can only run for at most n steps and so terminates with $w_1, \ldots, w_m$ obeying both of the desired properties. □

In additive combinatorics, one often wants to use results like this in finitary settings, such as that of a cyclic group $\mathbf{Z}/p\mathbf{Z}$ where p is a large prime. Now, technically speaking, $\mathbf{Z}/p\mathbf{Z}$ is not a vector space over $\mathbf{Q}$, because one only multiply an element of $\mathbf{Z}/p\mathbf{Z}$ by a rational number if the denominator of that rational does not divide p. But for p very large, $\mathbf{Z}/p\mathbf{Z}$ "behaves" like a vector space over $\mathbf{Q}$, at least if one restricts attention to the rationals of "bounded height" - where the numerator and denominator of the rationals are bounded. Thus we shall refer to elements of $\mathbf{Z}/p\mathbf{Z}$ as "vectors" over $\mathbf{Q}$, even though strictly speaking this is not quite the case.

On the other hand, saying that one element of $\mathbf{Z}/p\mathbf{Z}$ is a rational linear combination of another set of elements is not a very interesting statement: any non-zero element of $\mathbf{Z}/p\mathbf{Z}$ already generates the entire space! However, if one again restricts attention to rational linear combinations *of bounded height*, then things become interesting again. For instance, the vector 1 can generate elements such as 37 or $\frac{p-1}{2}$ using rational linear combinations of bounded height, but will not be able to generate such elements of $\mathbf{Z}/p\mathbf{Z}$ as $\lfloor\sqrt{p}\rfloor$ without using rational numbers of unbounded height.

For similar reasons, the notion of linear independence over the rationals doesn't initially look very interesting over $\mathbf{Z}/p\mathbf{Z}$: any two non-zero elements of $\mathbf{Z}/p\mathbf{Z}$ are of course rationally dependent. But again, if one restricts attention to rational numbers of bounded height, then independence begins to emerge: for instance, 1 and $\lfloor\sqrt{p}\rfloor$ are independent in this sense.

Thus, it becomes natural to ask whether there is a "quantitative" analogue of Theorem 4.11.1, with non-trivial content in the case of "vector spaces over the bounded height rationals" such as $\mathbf{Z}/p\mathbf{Z}$, which asserts that given any bounded collection $v_1, \ldots, v_n$ of elements, one can find another set $w_1, \ldots, w_k$ which is linearly independent "over the rationals up to some height", such that the $v_1, \ldots, v_n$ can be generated by the $w_1, \ldots, w_k$ "over the rationals up to some height". Of course to make this rigorous, one needs to quantify the two heights here, the one giving the independence, and the one giving the generation. In order to be useful for applications, it turns out that one often needs the former height to be much larger

than the latter; exponentially larger, for instance, is not an uncommon request. Fortunately, one can accomplish this, at the cost of making the height somewhat large:

Theorem 4.11.2 (Finite generation implies finite basis, finitary version)**.** *Let $n \geq 1$ be an integer, and let $F : \mathbf{N} \to \mathbf{N}$ be a function. Let V be an abelian group which admits a well-defined division operation by any natural number of size at most $C(F,n)$ for some constant $C(F,n)$ depending only on F, n; for instance one can take $V = \mathbf{Z}/p\mathbf{Z}$ for p a prime larger than $C(F,n)$. Let $v_1, \ldots, v_n$ be a finite collection of "vectors" in V. Then there exists a collection $w_1, \ldots, w_k$ of vectors in V, with $1 \leq k \leq n$, as well an integer $M \geq 1$, such that*

- *(Complexity bound) $M \leq C(F,n)$ for some $C(F,n)$ depending only on F, n.*
- *(w generates v) Every v_j can be expressed as a rational linear combination of the $w_1, \ldots, w_k$ of height at most M (i.e. the numerator and denominator of the coefficients are at most M).*
- *(w independent) There is no non-trivial linear relation $a_1 w_1 + \ldots + a_k w_k = 0$ among the $w_1, \ldots, w_k$ in which the $a_1, \ldots, a_k$ are rational numbers of height at most $F(M)$.*

In fact, one can take $w_1, \ldots, w_k$ to be a subset of the $v_1, \ldots, v_n$.

Proof. We perform the same "rank reduction argument" as before, but translated to the finitary setting. Start with $w_1, \ldots, w_k$ initialised to $v_1, \ldots, v_n$ (so initially we have $k = n$), and initialise $M = 1$. Clearly w generates v at this height. If the w_i are linearly independent up to rationals of height $F(M)$ then we are done. Otherwise, there is a non-trivial linear relation between them; after shuffling things around, we see that one of the w_i, say w_k, is a rational linear combination of the $w_1, \ldots, w_{k-1}$, whose height is bounded by some function depending on $F(M)$ and k. In such a case, w_k becomes redundant, and we may delete it (reducing the rank k by one), but note that in order for the remaining $w_1, \ldots, w_{k-1}$ to generate $v_1, \ldots, v_n$ we need to raise the height upper bound for the rationals involved from M to some quantity M' depending on $M, F(M), k$. We then replace M by M' and continue the process. We repeat this procedure; it can

only run for at most n steps and so terminates with $w_1, \ldots, w_m$ and M obeying all of the desired properties. (Note that the bound on M is quite poor, being essentially an n-fold iteration of F! Thus, for instance, if F is exponential, then the bound on M is *tower-exponential* in nature.) □

Remark 4.11.3. A variant of this type of approximate basis lemma was used in [**TaVu2007**].

Looking at the statements and proofs of these two theorems it is clear that the two results are in some sense the "same" result, except that the latter has been made sufficiently quantitative that it is meaningful in such finitary settings as $\mathbf{Z}/p\mathbf{Z}$. In this note I will show how this equivalence can be made formal using the language of non-standard analysis (see Section 1.9 of *Structure and Randomness*). This is not a particularly deep (or new) observation, but it is perhaps the simplest example I know of that illustrates how nonstandard analysis can be used to transfer a quantifier-heavy finitary statement, such as Theorem 4.11.2, into a quantifier-light infinitary statement, such as Theorem 4.11.1, thus lessening the need to perform "epsilon management" duties, such as keeping track of unspecified growth functions such as F. This type of transference is discussed at length in Section 1.3 of *Structure and Randomness*.

In this particular case, the amount of effort needed to set up the nonstandard machinery in order to reduce Theorem 4.11.2 from Theorem 4.11.1 is too great for this transference to be particularly worthwhile, especially given that Theorem 4.11.2 has such a short proof. However, when performing a particularly intricate argument in additive combinatorics, in which one is performing a number of "rank reduction arguments", "energy increment arguments", "regularity lemmas", "structure theorems", and so forth, the purely finitary approach can become bogged down with all the epsilon management one needs to do to organise all the parameters that are flying around. The nonstandard approach can efficiently hide a large number of these parameters from view, and it can then become worthwhile to invest in the nonstandard framework in order to clean up the rest of a lengthy argument. Furthermore, an advantage of moving up to the infinitary

setting is that one can then deploy all the firepower of an existing well-developed infinitary theory of mathematics (in this particular case, this would be the theory of linear algebra) *out of the box*, whereas in the finitary setting one would have to painstakingly finitise each aspect of such a theory that one wished to use (imagine for instance trying to finitise the *rank-nullity theorem* for rationals of bounded height).

The nonstandard approach is very closely related to use of compactness arguments, or of the technique of taking ultralimits and ultraproducts; indeed we will use an ultrafilter in order to create the nonstandard model in the first place.

I will also discuss a two variants of both Theorem 4.11.1 and Theorem 4.11.2 which have actually shown up in my research. The first is that of the *regularity lemma* for polynomials over finite fields, which came up when studying the equidistribution of such polynomials in [**GrTa2007**]. The second comes up when is dealing not with a single finite collection $v_1, \ldots, v_n$ of vectors, but rather with a *family* $(v_{h,1}, \ldots, v_{h,n})_{h \in H}$ of such vectors, where H ranges over a large set; this gives rise to what we call the *sunflower lemma*, and came up in [**GrTaZi2009**].

This post is mostly concerned with nonstandard translations of the "rank reduction argument". Nonstandard translations of the "energy increment argument" and "density increment argument" were briefly discussed in Section 4.10.

4.11.1. Equivalence of Theorems 4.11.1 and 4.11.2. Both Theorem 4.11.1 and Theorem 4.11.2 are easy enough to prove. But we will now spend a certain amount of effort in showing that one can deduce each theorem from the other without actually going through the proof of either. This may not seem particularly worthwhile (or to be serious overkill) in the case of these two particular theorems, but the method of deduction is extremely general, and can be used to relate much more deep and difficult infinitary and finitary theorems to each other without a significant increase in effort[3].

[3]This is closely related to various *correspondence principles* between combinatorics and parts of infinitary mathematics, such as ergodic theory; see also Section 1.3 of *Structure and Randomness* for a closely related equivalence.

Let's first show why the finitary theorem, Theorem 4.11.2, implies Theorem 4.11.1. We argue by contradiction. If Theorem 4.11.1 failed, then we could find a vector space V over the rationals, and a finite collection $v_1, \ldots, v_n$ of vectors, for which *no* finite subcollection $w_1, \ldots, w_k$ of the $v_1, \ldots, v_n$ obeyed both the generation property and the linear independence property. In other words, whenever a subcollection $w_1, \ldots, w_k$ happened to generate $v_1, \ldots, v_n$ by rationals, then it must necessarily contain a linear dependence.

We use this to create a function $F : \mathbf{N} \to \mathbf{N}$ as follows. Given any natural number M, consider all the finite subcollections $w_1, \ldots, w_k$ of $v_1, \ldots, v_n$ which can generate the $v_1, \ldots, v_n$ using rationals of height at most M. By the above hypothesis, all such subcollections contain a linear dependence involving rationals of some finite height. There may be many such dependences; we pick one arbitrarily. We then choose $F(M)$ to be any natural number larger than the heights of all the rationals involved in all the linear dependencies thus chosen. (Here we implicitly use the fact that there are only finitely many subcollections of the $v_1, \ldots, v_n$ to search through.)

Having chosen this function F, we then apply Theorem 4.11.2 to the vectors $v_1, \ldots, v_n$ and this choice of function F, to obtain a subcollection $w_1, \ldots, w_k$ which generate the $v_1, \ldots, v_n$ using rationals of height at most M, and have no linear dependence involving rationals of height at most $F(M)$. But this contradicts the construction of F, and gives the claim.

Remark 4.11.4. Note how important it is here that the growth function F in Theorem 4.11.2 is not specified in advance, but is instead a parameter that can be set to be as "large" as needed. Indeed, for Theorem 4.11.2 for any fixed F (e.g. exponential, tower-exponential, Ackermann, etc.) gives a statement which is strictly "weaker" than Theorem 4.11.1 in a sense that I will not try to make precise here; it is only the union of *all* these statements for *all* conceivable F that gives the full strength of Theorem 4.11.1. A similar phenomenon occurs with the *finite convergence principle* (Section 1.3 of *Structure and Randomness*). It is this "second order" nature of infinitary statements (they quantify not just over numerical parameters such as N or ε, but also over functional parameters such as F) that make

such statements appear deeper than finitary ones, but the distinction largely disappears if one is willing to perform such second-order quantifications.

Now we turn to the more interesting deduction, which is to obtain Theorem 4.11.2 from Theorem 4.11.1. Again, one argues by contradiction. Suppose that Theorem 4.11.2 failed. Carefully negating all the quantifiers (and using the axiom of choice), we conclude that there exists a function $F : \mathbf{N} \to \mathbf{N}$ and a natural number n with the following property: given any natural number K, there exists an abelian group V_K which is divisible up to height K, and elements $v_{1,K}, \ldots, v_{n,K}$ in V_K such that there is *no* subcollection $w_{1,K}, \ldots, w_{k,K}$ of the $v_{1,K}, \ldots, v_{n,K}$, together with an integer $M \leq K$, such that $w_{1,K}, \ldots, w_{k,K}$ generate $v_{1,K}, \ldots, v_{n,K}$ using rationals of height at most M, and such that the $w_{1,K}, \ldots, w_{k,K}$ have no linear dependence using rationals of height at most $F(M)$.

We now perform an *ultralimit* as $K \to \infty$. We will not pause here to recall the machinery of ultrafilters, ultralimits, and ultraproducts, but refer the reader instead to *Section 1.5* of *Structure and Randomness* for discussion.

We pick a non-principal ultrafilter p of the natural numbers. Starting with the "standard" abelian groups V_K, we then form their *ultraproduct* $V = \prod_K V_K/p$, defined as the space of sequences $v = (v_K)_{K \in \mathbf{N}}$ with $v_K \in V_K$ for each K, modulo equivalence by p; thus two sequences $v = (v_K)_{K \in \mathbf{N}}$ and $v' = (v'_K)_{K \in \mathbf{N}}$ are considered equal if $v_K = v'_K$ for a p-large set of K (i.e. for a set of K that lies in p).

Now that non-standard objects are in play, we will need to take some care to distinguish between standard objects (e.g. standard natural numbers) and their nonstandard counterparts.

Since each of the V_K are an abelian group, V is also an abelian group (an easy special case of the *transfer principle*). Since each V_K is divisible up to height K, V is divisible up to all (standard) heights; in other words, V is actually a vector space over the (standard) rational numbers $\mathbf{Q}$. The point is that while none of the V_K are, strictly speaking, vector spaces over $\mathbf{Q}$, they increasingly behave as if they

were such spaces, and in the limit one recovers genuine vector space structure.

For each $1 \leq i \leq n$, one can take an ultralimit of the elements $v_{i,K} \in V_K$ to generate an element $v_i := (v_{i,K})_{K \in \mathbf{N}}$ of the ultraproduct V. So now we have n vectors $v_1, \ldots, v_n$ of a vector space V over $\mathbf{Q}$ - precisely the setting of Theorem 4.11.1! So we apply that theorem and obtain a subcollection $w_1, \ldots, w_k \in V$ of the $v_1, \ldots, v_n$, such that each v_i can be generated from the $w_1, \ldots, w_k$ using (standard) rationals, and such that the $w_1, \ldots, w_k$ are linearly independent over the (standard) rationals.

Since all (standard) rationals have a finite height, one can find a (standard) natural number M such that each of the v_i can be generated from the $w_1, \ldots, w_k$ using (standard) rationals of height at most M. Undoing the ultralimit, we conclude that for a p-large set of K's, all of the $v_{i,K}$ can be generated from the $w_{1,K}, \ldots, w_{k,K}$ using rationals of height at most M. But by hypothesis, this implies for all sufficiently large K in this p-large set, the $w_{1,K}, \ldots, w_{k,K}$ contain a non-trivial rational dependence of height at most $F(M)$, thus

$$\frac{a_{1,K}}{q_{1,K}} w_{1,K} + \ldots + \frac{a_{k,K}}{q_{k,K}} w_{k,K} = 0$$

for some integers $a_{i,K}, q_{i,K}$ of magnitude at most $F(M)$, with the $a_{k,K}$ not all zero.

By the pigeonhole principle (and the finiteness of $F(M)$), each of the $a_{i,K}, q_{i,K}$ is constant in K on a p-large set of K. So if we take an ultralimit again to go back to the nonstandard world, the quantities $a_i := (a_{i,K})_{K \in \mathbf{N}}$, $q_i := (q_{i,K})_{K \in \mathbf{N}}$ are standard integers (rather than merely nonstandard integers). Thus we have

$$\frac{a_1}{q_1} w_1 + \ldots + \frac{a_k}{q_k} w_k = 0$$

with the a_i not all zero, i.e. we have a linear dependence amongst the $w_1, \ldots, w_k$. But this contradicts Theorem 4.11.1.

4.11.2. Polynomials over finite fields. Let $\mathbf{F}$ a fixed finite field (e.g. the field $\mathbf{F}_2$ of two elements), and consider a high-dimensional finite vector space V over $\mathbf{F}$. A polynomial $P : \mathbf{F}^n \to \mathbf{F}$ of degree $\leq d$ can then be defined as a combination of monomials each of degree at

most d, or alternatively as a function whose $d+1^{\text{th}}$ derivative vanishes; see Section 1.12 of *Poincaré's Legacies, Vol. I* for some discussion of this equivalence.

We define the *rank* $rank_{\leq d-1}(P)$ of a degree $\leq d$ polynomial P to be the least number k of degree $\leq d-1$ polynomials $Q_1, \ldots, Q_k$, such that P is completely determined by $Q_1, \ldots, Q_k$, i.e. $P = f(Q_1, \ldots, Q_k)$ for some function $f : \mathbf{F}^k \to \mathbf{F}$. In the case when P has degree ≤ 2, this concept is very close to the familiar *rank* of a quadratic form or matrix.

A generalisation of the notion of linear independence is that of linear independence modulo low rank. Let us call a collection $P_1, \ldots, P_n$ of degree $\leq d$ polynomials *M-linearly independent* if every non-trivial linear combination $a_1 P_1 + \ldots + a_n P_n$ with $a_1, \ldots, a_n \in \mathbf{F}$ not all zero, has rank at least M:

$$rank_{\leq d-1}(a_1 P_1 + \ldots + a_n P_n) \leq M.$$

There is then the following analogue of Theorem 4.11.2:

Theorem 4.11.5 (Polynomial regularity lemma at one degree, finitary version)**.** *Let $n, d \geq 1$ be integers, let $\mathbf{F}$ be a finite field and let $F : \mathbf{N} \to \mathbf{N}$ be a function. Let V be a vector space over $\mathbf{F}$, and let $P_1, \ldots, P_n : V \to F$ be polynomials of degree $\leq d$. Then there exists a collection $Q_1, \ldots, Q_k : V \to F$ of polynomials of degree $\leq d$, with $1 \leq k \leq n$, as well an integer $M \geq 1$, such that*

- *(Complexity bound) $M \leq C(F, n, d, \mathbf{F})$ for some $C(F, n, d, \mathbf{F})$ depending only on $F, n, d, \mathbf{F}$.*
- *(Q generates P) Every P_j can be expressed as a $\mathbf{F}$-linear combination of the $Q_1, \ldots, Q_k$, plus an error E which has rank $rank_{\leq d-1}(E)$ at most M.*
- *(P independent) There is no non-trivial linear relation $a_1 Q_1 + \ldots + a_k Q_k = E$ among the $w_1, \ldots, w_m$ in which E has rank $rank_{\leq d-1}(E)$ at most $F(M)$.*

In fact, one can take $Q_1, \ldots, Q_k$ to be a subset of the $P_1, \ldots, P_n$.

This theorem can be proven in much the same way as Theorem 4.11.2, and the reader is invited to do so as an exercise. The constant

$C(F, n, d, \mathbf{F})$ can in fact be taken to be independent of d and $\mathbf{F}$, but this is not important to us here.

Roughly speaking, Theorem 4.11.5 asserts that a finite family of degree $\leq d$ polynomials can be expressed as a linear combination of degree $\leq d$ polynomials which are "linearly independent modulo low rank errors", plus some lower rank objects. One can think of this as *regularising* the degree $\leq d$ polynomials, modulo combinations of lower degree polynomials. For applications (and in particular, for understanding the equidistribution) one also needs to regularise the degree $\leq d-1$ polynomials that arise this way, and so forth for increasingly lower degrees until all polynomials are regularised. (A similar phenomenon occurs for the hypergraph regularity lemma.)

When working with theorems like this, it is helpful to think conceptually of "quotienting out" by all polynomials of low rank. Unfortunately, in the finitary setting, the polynomials of low rank do not form a group, and so the quotient is ill-defined. However, this can be rectified by passing to the infinitary setting. Indeed, once one does so, one can quotient out the low rank polynomials, and Theorem 4.11.5 follows directly from Theorem 4.11.1 (or more precisely, the analogue of that theorem in which the field of rationals $\mathbf{Q}$ is replaced by the finite field $\mathbf{F}$).

Let's see how this works. To prove Theorem 4.11.5, suppose for contradiction that the theorem failed. Then one can find $F, n, d, \mathbf{F}$, such that for every natural K, one can find a vector space V_K and polynomials $P_{1,K}, \ldots, P_{n,K} : V_K \to \mathbf{F}$ of degree $\leq d$, for which there do not exist polynomials $Q_{1,K}, \ldots, Q_{k,K}$ with $k \leq n$ and an integer $M \leq K$ such that each $P_{j,K}$ can be expressed as a linear combination of the $Q_{i,K}$ modulo an error of rank at most M, and such that there are no nontrivial linear relations amongst the $Q_{i,K}$ modulo errors of rank at most $F(M)$.

Taking an ultralimit as before, we end up with a (nonstandard) vector space V over $\mathbf{F}$ (which is likely to be infinite), and (nonstandard) polynomials $P_1, \ldots, P_n : V \to \mathbf{F}$ of degree $\leq d$ (here it is best to use the "local" definition of a polynomial of degree $\leq d$, as a (nonstandard) function whose $d+1^{\text{th}}$ derivative, but one can also view this as a (nonstandard) sum of monomials if one is careful).

The space $\mathrm{Poly}_{\leq d}(V)$ of (nonstandard) degree $\leq d$ polynomials on V is a (nonstandard) vector space over $\mathbf{F}$. Inside this vector space, one has the subspace $\mathrm{Lowrank}_{\leq d}(V)$ consisting of all polynomials $P \in \mathrm{Poly}_{\leq d}(V)$ whose rank $rank_{\leq d-1}(V)$ is a standard integer (as opposed to a nonstandard integer); call these the *bounded rank* polynomials. This is easily seen to be a subspace of $\mathrm{Poly}_{\leq d}(V)$ (although it is not a *nonstandard* or *internal* subspace, i.e. the ultralimit of subspaces of the $\mathrm{Poly}_{\leq d}(V_K)$). As such, one can rigorously form the quotient space $\mathrm{Poly}_{\leq d}(V)/\mathrm{Lowrank}_{\leq d}(V)$ of degree $\leq d$ polynomials, modulo bounded rank $\leq d$ polynomials.

The polynomials $P_1, \ldots, P_n$ then have representatives $P_1, \ldots, P_n \bmod \mathrm{Lowrank}_{\leq d}(V)$ in this quotient space. Applying Theorem 4.11.1 (for the field $\mathbf{F}$), one can then find a subcollection $Q_1, \ldots, Q_k \bmod \mathrm{Lowrank}_{\leq d}(V)$ which are linearly independent in this space, which generate $P_1, \ldots, P_n$. Undoing the quotient, we see that the $P_1, \ldots, P_n$ are linear combinations of the $Q_1, \ldots, Q_k$ plus a bounded rank error, while no nontrivial linear combination of $Q_1, \ldots, Q_k$ has bounded rank. Undoing the ultralimit as in the previous section, we obtain the desired contradiction.

We thus see that in the nonstandard world, the somewhat non-rigorous concepts of "low rank" and "high rank" can be formalised as that of "bounded rank" and "unbounded rank". Furthermore, the former space forms a subspace, so in the nonstandard world one can rigorously talk about "quotienting out by bounded rank errors". Thus we see that the algebraic machinery of quotient spaces can be applied in the nonstandard world directly, whereas in the finitary world it can only be applied heuristically. In principle, one could also start deploying more advanced tools of abstract algebra (e.g. exact sequences, cohomology, etc.) in the nonstandard setting, although this has not yet seriously begun to happen in additive combinatorics (although there are strong hints of some sort of "additive cohomology" emerging in the body of work surrounding the inverse conjecture for the Gowers norm, especially on the ergodic theory side of things).

4.11.3. Sunflowers. Now we return to vector spaces (or approximate vector spaces) V over the rationals, such as $V = \mathbf{Z}/p\mathbf{Z}$ for a large prime p. Instead of working with a single (small) tuple $v_1, \ldots, v_n$ of vectors in V, we now consider a *family* $(v_{1,h}, \ldots, v_{n,h})_{h \in H}$ of such

vectors in V, where H ranges over a large set, for instance a dense subset of the interval $X := [-N, N] = \{-N, \dots, N\}$ for some large N. This situation happens to show up in our recent work on the inverse conjecture for the Gowers norm, where the $v_{1,h}, \dots, v_{n,h}$ represent the various "frequencies" that arise in a derivative $\Delta_h f$ of a function f with respect to the shift h. (This need to consider families is an issue that also comes up in the finite field ergodic theory analogue [**BeTaZi2009**] of the inverse conjectures, due to the unbounded number of generators in that case, but interestingly can be avoided in the ergodic theory over $\mathbf{Z}$.)

In Theorem 4.11.2, the main distinction was between linear dependence and linear independence of the tuple $v_1, \dots, v_n$ (or some reduction of this tuple, such as $w_1, \dots, w_k$). We will continue to be interested in the linear dependence or independence of the tuples $v_{1,h}, \dots, v_{n,h}$ for various h. But we also wish to understand how the $v_{i,h}$ vary with h as well. At one extreme (the "structured" case), there is no dependence on h: $v_{i,h} = v_i$ for all i and all h. At the other extreme (the "pseudorandom" case), the $v_{i,h}$ are basically independent as h varies; in particular, for (almost) all of the pairs $h, h' \in H$, the tuples $v_{1,h}, \dots, v_{n,h}$ and $v_{1,h'}, \dots, v_{n,h'}$ are not just separately independent, but are *jointly* independent. One can think of $v_{1,h}, \dots, v_{n,h}$ and $v_{1,h'}, \dots, v_{n,h'}$ as being in "general position" relative to each other.

The *sunflower lemma* asserts that any family $(v_{1,h}, \dots, v_{n,h})_{h \in H}$ is basically a combination of the above scenarios, thus one can divide the family into a linearly independent *core* collection of vectors $(w_1, \dots, w_m)$ that do not depend on h, together with *petals* $(v'_{1,h}, \dots, v'_{k,h})_{h \in H'}$, which are in "general position" in the above sense, relative to the core. However, as a price one pays for this, one has to refine H to a dense subset H' of H. This lemma, which significantly generalises Theorem 4.11.2, is formalised as follows:

Theorem 4.11.6 (Sunflower lemma, finitary version)**.** *Let $n \geq 1$ be an integer, and let $F : \mathbf{N} \to \mathbf{N}$ be a function. Let V be an abelian group which admits a well-defined division operation by any natural number of size at most $C(F, n)$ for some constant $C(F, n)$ depending only on F, n. Let H be a finite set, and let $(v_{1,h}, \dots, v_{n,h})_{h \in H}$ be a collection of n-tuples of vectors in V indexed by H. Then there exists*

a subset H' of H, integers $k, m \geq 0$ with $m + k \leq n$, a collection $w_1, \ldots, w_m$ of "core" vectors in V for some m, a collection of "petal" vectors $(v'_{1,h}, \ldots, v'_{k,h})_{h \in H'}$ for each $h \in H'$, as well an integer $M \geq 1$, such that

- *(Complexity bound) $M \leq C(F, n)$ for some $C(F, n)$ depending only on F, n.*
- *(H' dense) one has $|H'| \geq c(F, n)|H|$ for some $c(F, n) > 0$ depending only on F, n.*
- *(w, v' generates v) Every $v_{j,h}$ with $1 \leq j \leq n$ and $h \in H'$ can be expressed as a rational linear combination of the $w_1, \ldots, w_m$ and $v'_{1,h}, \ldots, v'_{k,h}$ of height at most M.*
- *(w independent) There is no non-trivial rational linear relation among the $w_1, \ldots, w_m$ of height at most $F(M)$.*
- *(v' in general position relative to w) More generally, for $1 - \frac{1}{F(M)}$ of the pairs $(h, h') \in H' \times H'$, there is no non-trivial linear relation among $w_1, \ldots, w_m, v'_{1,h}, \ldots, v'_{k,h}, v'_{1,h'}, \ldots, v'_{k,h'}$ of height at most $F(M)$.*

One can take the $v'_{1,h}, \ldots, v'_{k,h}$ to be a subcollection of the $v_{1,h}, \ldots, v_{n,h}$, though this is not particularly useful in applications.

Proof. We perform a two-parameter "rank reduction argument", where the rank is indexed by the pair (k, m) (ordered lexicographically). We initially set $m = 0$, $k = n$, $H' = H$, $M = 1$, and $v'_{i,h} = v_{i,h}$ for $h \in H$.

At each stage of the iteration, w, v' will generate v (at height M), and we will have some complexity bound on M, m and some density bound on H'. So one needs to check the independence of w and the general position of v' relative to w.

If there is a linear relation of w at height $F(M)$, then one can use this to reduce the size m of the core by one, leaving the petal size k unchanged, just as in the proof of Theorem 4.11.2. So let us move on, and suppose that there is no linear relation of w at height $F(M)$, but instead there is a failure of the general position hypothesis. In other words, for at least $|H'|^2/F(M)$ pairs $(h, h') \in H' \times H'$, one can

find a relation of the form

$$a_{1,h,h'}w_1+\ldots+a_{m,h,h'}w_m+b_{1,h,h'}v'_{1,h}+\ldots+b_{k,h,h'}v'_{k,h}+c_{1,h,h'}v'_{1,h'}+\ldots+c_{k,h,h'}v'_{k,h'}=0$$

where the $a_{i,h,h'}, b_{i,h,h'}, c_{i,h,h'}$ are rationals of height at most $F(M)$, not all zero. The number of possible values for such rationals is bounded by some quantity depending on $m, k, F(M)$. Thus, by the pigeonhole principle, we can find $\gg_{F(M),m,k} |H'|^2$ pairs (i.e. at least $c(F(M),m,k)|H'|^2$ pairs for some $c(F(M),m,k) > 0$ depending only on $F(M), m, k$) such that

$$a_1w_1+\ldots+a_mw_m+b_1v'_{1,h}+\ldots+b_kv'_{k,h}+c_1v'_{1,h'}+\ldots+c_kv'_{k,h'}=0$$

for some fixed rationals a_i, b_i, c_i of height at most $F(M)$. By the pigeonhole principle again, we can then find a fixed $h_0 \in H'$ such that

$$a_1w_1+\ldots+a_mw_m+b_1v'_{1,h}+\ldots+b_kv'_{k,h}=u_{h_0}$$

for all h in some subset H'' of H' with $|H''| \gg_{F(M),m,k} |H'|$, where

$$u_{h_0} := -c_1v'_{1,h_0}-\ldots-c_kv'_{k,h_0}.$$

If the b_i and c_i all vanished then we would have a linear dependence amongst the core vectors, which we already know how to deal with. So suppose that we have at least one active petal coefficient, say b_k. Then upon rearranging, we can express $v'_{k,h}$ as some rational linear combination of the original core vectors $w_1, \ldots, w_m$, a new core vector u_{h_0}, and the other petals $v'_{1,h}, \ldots, v'_{k-1,h}$, with heights bounded by $\ll_{F(M),k,m} 1$. We may thus refine H' to H'', delete the petal vector $v'_{k,h}$, and add the vector u to the core, thus decreasing k by one and increasing m by one. One still has the generation property so long as one replaces M with a larger M' depending on $M, F(M), k, m$.

Since each iteration of this process either reduces m by one keeping k fixed, or reduces k by one increasing m, we see that after at most $2n$ steps, the process must terminate, when we have both the linear independence of the w property and the general position of the v' property. (Note here that we are basically performing a *proof by infinite descent*.) At that stage, one easily verifies that we have obtained all the required conclusions of the theorem. □

As one can see, this result is a little bit trickier to prove than Theorem 4.11.2. Let us now see how it will translate to the nonstandard setting, and see what the nonstandard analogue of Theorem 4.11.6 is. We will skip some details, and get to the point where we can motivate and prove this nonstandard analogue; this analogue does in fact imply Theorem 4.11.6 by repeating the arguments from previous sections, but we will leave this as an exercise for the interested reader.

As before, the starting point is to introduce a parameter K, so that the approximate vector space V_K now depends on K (and becomes an actual vector space in the ultralimit V), and the parameter set H_K now also depends on K. We will think of $|H_K|$ as going to infinity as $K \to \infty$, as this is the most interesting case (for bounded H_K, the result basically collapses back to Theorem 4.11.2). In that case, the ultralimit H of the H_K is a nonstandard finite set (i.e. an ultralimit of finite sets) whose (nonstandard) cardinality $|H|$ is an *unbounded* nonstandard integer: it is a nonstandard integer (indeed, it is the ultralimit of the $|H_K|$) which is larger than any standard integer. On the other hand, n and F remain standard (i.e. they do not involve K).

For each K, one starts with a family $(v_{1,h,K}, \ldots, v_{n,h,K})_{h \in H_K}$ of n-tuples of vectors in V_K. Taking ultralimits, one ends up with a family $(v_{1,h}, \ldots, v_{n,h})_{h \in H}$ of n-tuples of vectors in V. Furthermore, for each $1 \leq i \leq n$, the maps $h \mapsto v_{i,h}$ are *nonstandard* (or *internal*) functions from H to V, i.e. they are ultralimits of maps from H_K to V_K. The internal nature of these maps (which is a kind of "measurability" condition on these functions) will be important later. Of course, H and V are also internal (being ultralimits of H_K and V_K respectively).

We say that a subset H' of H is *dense* if it is an internal subset (i.e. it is the ultralimit of some subsets H'_K of H_K), and if $|H'| \geq \varepsilon |H|$ for some standard $\varepsilon > 0$ (recall that $|H'|, |H|$ are nonstandard integers). If an internal subset is not dense, we say that it is *sparse*, which in nonstandard asymptotic notation (see Section 1.3 of *Structure and Randomness*) is equivalent to $|H'| = o(|H|)$. If a statement $P(h)$ holds on all h in dense set of H, we say that it holds for *many* h; if it holds for all h outside of a sparse set, we say it holds for *almost*

all h. These are analogous to the more familiar concepts of "holding with positive probability" and "holding almost surely" in probability theory. For instance, if $P(h)$ holds for many h in H, and $Q(h)$ holds for almost all h in H, then $P(h)$ and $Q(h)$ jointly hold for many h in H. Note how all the epsilons have been neatly hidden away in this nonstandard framework.

Now we state the nonstandard analogue of Theorem 4.11.6.

Theorem 4.11.7 (Sunflower lemma, nonstandard version)**.** *Let $n \geq 1$ be a (standard) integer, let V be a (nonstandard) vector space over the standard rationals $\mathbf{Q}$, and let H be a (nonstandard) set. Let $(v_{1,h}, \ldots, v_{n,h})_{h \in H}$ be a collection of n-tuples of vectors in V indexed by H, such that all the maps $h \mapsto v_{i,h}$ for $1 \leq i \leq n$ are internal. Then there exists a dense subset H' of H, a bounded-dimensional subspace W of V, a (standard) integer $k \geq 0$ with $\dim(W) + k \leq n$, and a collection of "petal" vectors $(v'_{1,h}, \ldots, v'_{k,h})_{h \in H'}$ for each $h \in H'$, with the maps $h \mapsto v'_{i,h}$ being internal for all $1 \leq i \leq k$, such that*

- *$(W, v'$ generates $v)$ Every $v_{j,h}$ with $1 \leq j \leq n$ and $h \in H'$ lies in the span of W and the $v'_{1,h}, \ldots, v'_{k,h}$.*
- *(v' in general position relative to W) For almost all of the pairs $(h, h') \in H' \times H'$, the vectors $v'_{1,h}, \ldots, v'_{k,h}, v'_{1,h'}, \ldots, v'_{k,h'}$ are linearly independent modulo W over $\mathbf{Q}$.*

Of course, using Theorem 4.11.1 one could obtain a basis $w_1, \ldots, w_m$ for W with $m = \dim(W)$, at which point the theorem more closely resembles Theorem 4.11.6.

Proof. Define a *partial representation* of the family $(v_{1,h}, \ldots, v_{n,h})$ to be a dense subset H' of H, a bounded dimensional space W, a standard integer k with $\dim(W) + k \leq n$, and a collection of $(v'_{1,h}, \ldots, v'_{k,h})_{h \in H'}$ depending internally on h that obeys the generation property (but not necessarily the general position property). Clearly we have at least one partial representation, namely the trivial one where W is empty, $k = n$, $H' := H$, and $v'_{i,h} := v_{i,h}$. Now, among all such partial representations, let us take a representation with the minimal value of k. (Here we are of course using the *well-ordering*

property of the standard natural numbers.) We claim that this representation enjoys the general position property, which will give the claim.

Indeed, suppose this was not the case. Then, for many pairs $(h, h') \in H' \times H'$, the vectors $v'_{1,h}, \ldots, v'_{k,h}, v'_{1,h'}, \ldots, v'_{k,h'}$ have a linear dependence modulo W over $\mathbf{Q}$. (Actually, there is a technical "measurability" issue to address here, which I will return to later.) By symmetry and pigeonholing, we may assume that the $v'_{k,h}$ coefficient of (say) of this dependence is non-zero. (Again, there is a measurability issue here.) Applying the pigeonhole principle, one can find $h_0 \in H'$ such that

$$v'_{1,h}, \ldots, v'_{k,h}, v'_{1,h_0}, \ldots, v'_{k,h_0}$$

have a linear dependence over $\mathbf{Q}$ modulo W for many h. (Again, there is a measurability issue here.)

Fix h_0. The number of possible linear combinations of $v'_{1,h_0}, \ldots, v'_{k,h_0}$ is countable. Because of this (and using a "countable pigeonhole principle") that I will address below, we can find a *fixed* rational linear combination u_{h_0} of the $v'_{1,h_0}, \ldots, v'_{k,h_0}$ such that

$$v'_{1,h}, \ldots, v'_{k,h}, u_{h_0}$$

have a linear dependence over $\mathbf{Q}$ modulo W for all h in some dense subset H'' of H'. But now one can pass from H' to the dense subset H'', delete the petal $v'_{k,h}$, and add the vector u_{h_0} to the core space W, thus creating a partial representation with a smaller value of k, contradicting minimality, and we are done. □

We remark here that whereas the finitary analogue of this result was proven using the method of infinite descent, the nonstandard version could instead be proven using the (equivalent) well-ordering principle. One could easily recast the nonstandard version in descent form also, but it is somewhat more difficult to cast the finitary argument using well-ordering due to the extra parameters and quantifiers in play.

Let us now address the measurability issues. The main problem here is that the property of having a linear dependence over the standard rationals $\mathbf{Q}$ is not an internal property, because it requires

knowledge of what the standard rationals are, which is not an internal concept in the language of vector spaces. However, for each fixed choice of rational coefficients, the property of having a specific linear dependence with those selected coefficients *is* an internal concept (here we crucially rely on the hypothesis that the maps $h \mapsto v_{i,h}$ were internal), so really what we have here is a sort of "σ-internal" property (a countable union of internal properties). But this is good enough for many purposes. In particular, we have

Lemma 4.11.8 (Countable pigeonhole principle)**.** *Let H be a non-standardly finite set (i.e. the ultralimit of finite sets H_K), and for each standard natural number n, let E_n be an internal subset of H. Then one of the following holds:*

- *(Positive density) There exists a natural number n such that $h \in E_n$ for many $h \in H$ (i.e. E_n is a dense subset of H).*
- *(Zero density) For almost all $h \in H$, one has $h \notin E_n$ for all n. (In other words, the (external) set $\bigcup_{n \in \mathbf{N}} E_n$ in is contained in a sparse subset of H.)*

This lemma is sufficient to resolve all the measurability issues raised in the previous proof. It is analogous to the trivial statement in measure theory that given a countable collection of measurable subsets of a space of positive measure, either one of the measurable sets has positive measure, or else their union has measure zero (i.e. the sets fail to cover almost all of the space).

Proof. If any of the E_n are dense, we are done. So suppose this is not the case. Since E_n is a definable subset of H which is not dense, it is sparse, thus $|E_n| = o(|H|)$. Now it is convenient to undo the ultralimit and work in the finite sets H_K that H is the ultralimit of. Note that each E_n, being internal, is also an ultralimit of some finite subsets $E_{n,K}$ of H_K.

For each standard integer $M > 0$, the set $E_1 \cup \ldots \cup E_M$ is sparse in H, and in particular has density less than $1/M$. Thus, one can find a p-large set $S_M \subset \mathbf{N}$ such that

$$|E_{1,K} \cup \ldots \cup E_{M,K}| \leq |H_K|/M$$

for all $K \in S_M$. One can arrange matters so that the S_M are decreasing in M. One then sets the set E_K to equal $E_{1,K} \cup \ldots \cup E_{M,K}$, where M is the smallest integer for which $K \in S_M$ (or E_K is empty if K lies in all the S_M, or in none), and let E be the ultralimit of the E_K. Then we see that $|E| \leq |H|/M$ for every standard M, and so E is a sparse subset of H. Furthermore, E contains E_M for every standard M, and so we are in the zero density conclusion of the argument. □

Remark 4.11.9. Curiously, I don't see how to prove this lemma without unpacking the limit; it doesn't seem to follow just from, say, the *overspill principle*. Instead, it seems to be exploiting the weak countable saturation property I mentioned in Section 4.10. But perhaps I missed a simple argument.

4.11.4. Summary. Let me summarise with a brief list of pros and cons of switching to a nonstandard framework. First, the pros:

- Many "first-order" parameters such as ε or N disappear from view, as do various "negligible" errors. More importantly, "second-order" parameters, such as the function F appearing in Theorem 4.11.2, also disappear from view. (In principle, third-order and higher parameters would also disappear, though I do not yet know of an actual finitary argument in my fields of study which would have used such parameters (with the exception of Ramsey theory, where such parameters must come into play in order to generate such enormous quantities as *Graham's number*).) As such, a lot of tedious "epsilon management" disappears.
- Iterative (and often parameter-heavy) arguments can often be replaced by minimisation (or more generally, extremisation) arguments, taking advantage of such properties as the *well-ordering principle*, the *least upper bound axiom*, or compactness.
- The *transfer principle* lets one use "for free" any (first-order) statement about standard mathematics in the non-standard setting (provided that all objects involved are *internal*; see below).

- Mature and powerful theories from infinitary mathematics (e.g. linear algebra, real analysis, representation theory, topology, functional analysis, measure theory, Lie theory, ergodic theory, model theory, etc.) can be used rigorously in a nonstandard setting (as long as one is aware of the usual infinitary pitfalls, of course; see below).
- One can formally define terms that correspond to what would otherwise only be heuristic (or heavily parameterised and quantified) concepts such as "small", "large", "low rank", "independent", "uniformly distributed", etc.
- The conversion from a standard result to its nonstandard counterpart, or vice versa, is fairly quick (but see below), and generally only needs to be done only once or twice per paper.

Next, the cons:

- Often requires the axiom of choice, as well as a certain amount of set theory. (There are however weakened versions of nonstandard analysis that can avoid choice that are still suitable for many applications.)
- One needs the machinery of ultralimits and ultraproducts to set up the conversion from standard to nonstandard structures.
- The conversion usually proceeds by a proof by contradiction, which (in conjunction with the use of ultralimits) may not be particularly intuitive.
- One cannot efficiently discern what quantitative bounds emerge from a nonstandard argument (other than by painstakingly converting it back to a standard one, or by applying the tools of *proof mining*). (On the other hand, in particularly convoluted standard arguments, the quantitative bounds are already so poor - e.g. of iterated tower-exponential type - that letting go of these bounds is no great loss.)
- One has to take some care to distinguish between standard and nonstandard objects (and also between internal and external sets and functions, which are concepts somewhat

analogous to measurable and non-measurable sets and functions in measurable theory). More generally, all the usual pitfalls of infinitary analysis (e.g. interchanging limits, or the need to ensure measurability or continuity) emerge in this setting, in contrast to the finitary setting where they are usually completely trivial.

- It can be difficult at first to conceptually visualise what nonstandard objects look like (although this becomes easier once one maps nonstandard analysis concepts to heuristic concepts such as "small" and "large" as mentioned earlier, thus for instance one can think of an unbounded nonstandard natural number as being like an incredibly large standard natural number).
- It is inefficient for both nonstandard and standard arguments to coexist within a paper; this makes things a little awkward if one for instance has to cite a result from a standard mathematics paper in a nonstandard mathematics one.
- There are philosophical objections to using mathematical structures that only exist abstractly, rather than corresponding to the "real world". (Note though that similar objections were also raised in the past with regard to the use of, say, complex numbers, non-Euclidean geometries, or even negative numbers.)
- Formally, there is no increase in logical power gained by using nonstandard analysis (at least if one accepts the axiom of choice); anything which can be proven by nonstandard methods can also be proven by standard ones. In practice, though, the length and clarity of the nonstandard proof may be substantially better than the standard one.

In view of the pros and cons, I would not say that nonstandard analysis is suitable in all situations, nor is it unsuitable in all situations, but one needs to carefully evaluate the costs and benefits in a given setting; also, in some cases having both a finitary and infinitary proof side by side for the same result may be more valuable than just having one of the two proofs. My rule of thumb is that if a finitary argument is already spitting out iterated tower-exponential

type bounds or worse in an argument, this is a sign that the argument "wants" to be infinitary, and it may be simpler to move over to an infinitary setting (such as the nonstandard setting).

Notes. This article first appeared at `terrytao.wordpress.com/2009/12/13`.

4.12. The double Duhamel trick and the in/out decomposition

This is a technical post inspired by separate conversations with Jim Colliander and with Soonsik Kwon on the relationship between two techniques used to control non-radiating solutions to dispersive nonlinear equations, namely the "double Duhamel trick" and the "in/out decomposition". See for instance [**KiVi2009**] for a survey of these two techniques and other related methods in the subject. (I should caution that this article is likely to be unintelligible to anyone not already working in this area.)

For sake of discussion we shall focus on solutions to a nonlinear Schrödinger equation

$$iu_t + \Delta u = F(u)$$

and we will not concern ourselves with the specific regularity of the solution u, or the specific properties of the nonlinearity F here. We will also not address the issue of how to justify the formal computations being performed here.

Solutions to this equation enjoy the *forward Duhamel formula*

$$u(t) = e^{i(t-t_0)\Delta} u(t_0) - i \int_{t_0}^{t} e^{i(t-t')\Delta} F(u(t'))\ dt'$$

for times t to the future of t_0 in the lifespan of the solution, as well as the *backward Duhamel formula*

$$u(t) = e^{i(t-t_1)\Delta} u(t_1) + i \int_{t}^{t_1} e^{i(t-t')\Delta} F(u(t'))\ dt'$$

for all times t to the past of t_1 in the lifespan of the solution. The first formula asserts that the solution at a given time is determined by the initial state and by the immediate past, while the second formula is the time reversal of the first, asserting that the solution at a given time is determined by the final state and the immediate future. These

basic causal formulae are the foundation of the local theory of these equations, and in particular play an instrumental role in establishing local well-posedness for these equations. In this local theory, the main philosophy is to treat the homogeneous (or *linear*) term $e^{i(t-t_0)\Delta}u(t_0)$ or $e^{i(t-t_1)\Delta}u(t_1)$ as the main term, and the inhomogeneous (or *nonlinear*, or *forcing*) integral term as an error term.

The situation is reversed when one turns to the global theory, and looks at the asymptotic behaviour of a solution as one approaches a limiting time T (which can be infinite if one has global existence, or finite if one has finite time blowup). After a suitable rescaling, the linear portion of the solution often disappears from view, leaving one with an *asymptotic blowup profile* solution which is *non-radiating* in the sense that the linear components of the Duhamel formulae vanish, thus

$$u(t) = -i \int_{t_0}^{t} e^{i(t-t')\Delta} F(u(t'))\ dt' \tag{4.42}$$

and

$$u(t) = i \int_{t}^{t_1} e^{i(t-t')\Delta} F(u(t'))\ dt' \tag{4.43}$$

where t_0, t_1 are the endpoint times of existence. (This type of situation comes up for instance in the Kenig-Merle approach to critical regularity problems, by reducing to a minimal blowup solution which is almost periodic modulo symmetries, and hence non-radiating.) These types of non-radiating solutions are propelled solely by their own nonlinear self-interactions from the immediate past or immediate future; they are generalisations of "nonlinear bound states" such as solitons.

A key task is then to somehow combine the forward representation (4.42) and the backward representation (4.43) to obtain new information on $u(t)$ itself, that cannot be obtained from either representation alone; it seems that the immediate past and immediate future can collectively exert more control on the present than they each do separately. This type of problem can be abstracted as follows. Let $\|u(t)\|_{Y_+}$ be the infimal value of $\|F_+\|_N$ over all forward

representations of $u(t)$ of the form

$$u(t) = \int_{t_0}^{t} e^{i(t-t')\Delta} F_+(t')\ dt' \tag{4.44}$$

where N is some suitable spacetime norm (e.g. a Strichartz-type norm), and similarly let $\|u(t)\|_{Y_-}$ be the infimal value of $\|F_-\|_N$ over all backward representations of $u(t)$ of the form

$$u(t) = \int_{t}^{t_1} e^{i(t-t')\Delta} F_-(t')\ dt'. \tag{4.45}$$

Typically, one already has (or is willing to assume as a bootstrap hypothesis) control on $F(u)$ in the norm N, which gives control of $u(t)$ in the norms Y_+, Y_-. The task is then to use the control of both the Y_+ and Y_- norm of $u(t)$ to gain control of $u(t)$ in a more conventional Hilbert space norm X, which is typically a Sobolev space such as H^s or L^2.

One can use some classical functional analysis to clarify this situation. By the closed graph theorem, the above task is (morally, at least) equivalent to establishing an *a priori* bound of the form

$$\|u\|_X \lesssim \|u\|_{Y_+} + \|u\|_{Y_-} \tag{4.46}$$

for all reasonable u (e.g. test functions). The *double Duhamel trick* accomplishes this by establishing the stronger estimate

$$|\langle u, v\rangle_X| \lesssim \|u\|_{Y_+} \|v\|_{Y_-} \tag{4.47}$$

for all reasonable u, v; note that setting $u = v$ and applying the arithmetic-geometric inequality then gives (4.46). The point is that if u has a forward representation (4.44) and v has a backward representation (4.45), then the inner product $\langle u, v\rangle_X$ can (formally, at least) be expanded as a double integral

$$\int_{t_0}^{t} \int_{t}^{t_1} \langle e^{i(t''-t')\Delta} F_+(t'), e^{i(t''-t')\Delta} F_-(t')\rangle_X\ dt'' dt'.$$

The dispersive nature of the linear Schrödinger equation often causes $\langle e^{i(t''-t')\Delta} F_+(t'), e^{i(t''-t')\Delta} F_-(t')\rangle_X$ to decay, especially in high dimensions. In high enough dimension (typically one needs five or higher dimensions, unless one already has some spacetime control on the solution), the decay is stronger than $1/|t'-t''|^2$, so that the integrand becomes absolutely integrable and one recovers (4.47).

Unfortunately it appears that estimates of the form (4.47) fail in low dimensions (for the type of norms N that actually show up in applications); there is just too much interaction between past and future to hope for any reasonable control of this inner product. But one can try to obtain (4.46) by other means. By the Hahn-Banach theorem (and ignoring various issues related to reflexivity), (4.46) is equivalent to the assertion that every $u \in X$ can be decomposed as $u = u_+ + u_-$, where $\|u_+\|_{Y_+^*} \lesssim \|u\|_X$ and $\|u_-\|_{Y_-^*} \lesssim \|v\|_X$. Indeed once one has such a decomposition, one obtains (4.46) by computing the inner product of u with $u = u_+ + u_-$ in X in two different ways. One can also (morally at least) write $\|u_+\|_{Y_+^*}$ as $\|e^{i(\cdot-t)\Delta}u_+\|_{N^*([t_0,t])}$ and similarly write $\|u_-\|_{Y_-^*}$ as $\|e^{i(\cdot-t)\Delta}u_-\|_{N^*([t,t_1])}$

So one can dualise the task of proving (4.46) as that of obtaining a decomposition of an arbitrary initial state u into two components u_+ and u_-, where the former disperses into the past and the latter disperses into the future under the linear evolution. We do not know how to achieve this type of task efficiently in general - and doing so would likely lead to a significant advance in the subject (perhaps one of the main areas in this topic where serious harmonic analysis is likely to play a major role). But in the model case of spherically symmetric data u, one can perform such a decomposition quite easily: one uses microlocal projections to set u_+ to be the "inward" pointing component of u, which propagates towards the origin in the future and away from the origin in the past, and u_- to simimlarly be the "outward" component of u. As spherical symmetry significantly dilutes the amplitude of the solution (and hence the strength of the nonlinearity) away from the origin, this decomposition tends to work quite well for applications, and is one of the main reasons (though not the only one) why we have a global theory for low-dimensional nonlinear Schrödinger equations in the radial case, but not in general.

The in/out decomposition is a linear one, but the Hahn-Banach argument gives no reason why the decomposition needs to be linear. (Note that other well-known decompositions in analysis, such as the Fefferman-Stein decomposition of BMO, are necessarily nonlinear, a fact which is ultimately equivalent to the non-complemented nature of a certain subspace of a Banach space; see Section 1.7.) So one could

imagine a sophisticated nonlinear decomposition as a general substitute for the in/out decomposition. See for instance [**BoBr2003**] for some of the subtleties of decomposition even in very classical function spaces such as $H^{1/2}(R)$. Alternatively, there may well be a third way to obtain estimates of the form (4.46) that do not require either decomposition or the double Duhamel trick; such a method may well clarify the relative relationship between past, present, and future for critical nonlinear dispersive equations, which seems to be a key aspect of the theory that is still only partially understood. (In particular, it seems that one needs a fairly strong decoupling of the present from both the past and the future to get the sort of elliptic-like regularity results that allow us to make further progress with such equations.)

Notes. This article first appeared at `terrytao.wordpress.com/2009/12/17`. Thanks to Kareem Carr, hezhigang, and anonymous commenters for corrections.

4.13. The free nilpotent group

In a multiplicative group G, the *commutator* of two group elements g, h is defined as $[g,h] := g^{-1}h^{-1}gh$ (other conventions are also in use, though they are largely equivalent for the purposes of this discussion). A group is said to be *nilpotent of step s* (or more precisely, step $\leq s$), if all iterated commutators of order $s+1$ or higher necessarily vanish. For instance, a group is nilpotent of order 1 if and only if it is abelian, and it is nilpotent of order 2 if and only if $[[g_1,g_2],g_3] = id$ for all g_1, g_2, g_3 (i.e. all commutator elements $[g_1,g_2]$ are *central*), and so forth. A good example of an s-step nilpotent group is the group of $s+1 \times s+1$ upper-triangular *unipotent* matrices (i.e. matrices with 1s on the diagonal and zero below the diagonal), and taking values in some ring (e.g. reals, integers, complex numbers, etc.).

Another important example of nilpotent groups arise from operations on polynomials. For instance, if $V_{\leq s}$ is the vector space of real polynomials of one variable of degree at most s, then there are two natural affine actions on $V_{\leq s}$. Firstly, every polynomial Q in $V_{\leq s}$ gives rise to an "vertical" shift $P \mapsto P+Q$. Secondly, every $h \in \mathbf{R}$ gives rise to a "horizontal" shift $P \mapsto P(\cdot + h)$. The group

generated by these two shifts is a nilpotent group of step $\leq s$; this reflects the well-known fact that a polynomial of degree $\leq s$ vanishes once one differentiates more than s times. Because of this link between nilpotentcy and polynomials, one can view nilpotent algebra as a generalisation of polynomial algebra.

Suppose one has a finite number $g_1, \ldots, g_n$ of generators. Using abstract algebra, one can then construct the *free nilpotent group* $\mathcal{F}_{\leq s}(g_1, \ldots, g_n)$ of step $\leq s$, defined as the group generated by the $g_1, \ldots, g_n$ subject to the relations that all commutators of order $s+1$ involving the generators are trivial. This is the *universal object* in the category of nilpotent groups of step $\leq s$ with n marked elements $g_1, \ldots, g_n$. In other words, given any other $\leq s$-step nilpotent group G' with n marked elements $g'_1, \ldots, g'_n$, there is a unique homomorphism from the free nilpotent group to G' that maps each g_j to g'_j for $1 \leq j \leq n$. In particular, the free nilpotent group is well-defined up to isomorphism in this category.

In many applications, one wants to have a more concrete description of the free nilpotent group, so that one can perform computations more easily (and in particular, be able to tell when two words in the group are equal or not). This is easy for small values of s. For instance, when $s = 1$, $\mathcal{F}_{\leq 1}(g_1, \ldots, g_n)$ is simply the *free abelian group* generated by $g_1, \ldots, g_n$, and so every element g of $\mathcal{F}_{\leq 1}(g_1, \ldots, g_n)$ can be described uniquely as

$$g = \prod_{j=1}^{n} g_j^{m_j} := g_1^{m_1} \ldots g_n^{m_n} \tag{4.48}$$

for some integers $m_1, \ldots, m_n$, with the obvious group law. Indeed, to obtain existence of this representation, one starts with any representation of g in terms of the generators $g_1, \ldots, g_n$, and then uses the abelian property to push the g_1 factors to the far left, followed by the g_2 factors, and so forth. To show uniqueness, we observe that the group G of formal abelian products $\{g_1^{m_1} \ldots g_n^{m_n} : m_1, \ldots, m_n \in \mathbf{Z}\} \equiv \mathbf{Z}^k$ is already a ≤ 1-step nilpotent group with marked elements $g_1, \ldots, g_n$, and so there must be a homomorphism from the free group to G. Since G distinguishes all the products $g_1^{m_1} \ldots g_n^{m_n}$ from each other, the free group must also.

It is only slightly more tricky to describe the free nilpotent group $\mathcal{F}_{\leq 2}(g_1, \ldots, g_n)$ of step ≤ 2. Using the identities

$$gh = hg[g,h]; \quad gh^{-1} = ([g,h]^{-1})^{g^{-1}} h^{-1} g; \quad g^{-1}h = h[g,h]^{-1}g^{-1}; \quad g^{-1}h^{-1} := [g,h]g^{-1}h^{-1}$$

(where $g^h := h^{-1}gh$ is the conjugate of g by h) we see that whenever $1 \leq i < j \leq n$, one can push a positive or negative power of g_i past a positive or negative power of g_j, at the cost of creating a positive or negative power of $[g_i, g_j]$, or one of its conjugates. Meanwhile, in a ≤ 2-step nilpotent group, all the commutators are central, and one can pull all the commutators out of a word and collect them as in the abelian case. Doing all this, we see that every element g of $\mathcal{F}_{\leq 2}(g_1, \ldots, g_n)$ has a representation of the form

$$g = (\prod_{j=1}^{n} g_j^{m_j})(\prod_{1 \leq i < j \leq n} [g_i, g_j]^{m_{[i,j]}}) \tag{4.49}$$

for some integers m_j for $1 \leq j \leq n$ and $m_{[i,j]}$ for $1 \leq i < j \leq n$. Note that we don't need to consider commutators $[g_i, g_j]$ for $i \geq j$, since

$$[g_i, g_i] = id$$

and

$$[g_i, g_j] = [g_j, g_i]^{-1}.$$

It is possible to show also that this representation is unique, by repeating the previous argument, i.e. by showing that the set of formal products

$$G := \{(\prod_{j=1}^{k} g_j^{m_j})(\prod_{1 \leq i < j \leq n} [g_i, g_j]^{m_{[i,j]}}) : m_j, m_{[i,j]} \in \mathbf{Z}\}$$

forms a ≤ 2-step nilpotent group, after using the above rules to define the group operations. This can be done, but verifying the group axioms (particularly the associative law) for G is unpleasantly tedious.

Once one sees this, one rapidly loses an appetite for trying to obtain a similar explicit description for free nilpotent groups for higher step, especially once one starts seeing that higher commutators obey some non-obvious identities such as the *Hall-Witt identity*

$$[[g, h^{-1}], k]^h \cdot [[h, k^{-1}], g]^k \cdot [[k, g^{-1}], h]^g = 1 \tag{4.50}$$

(a nonlinear version of the *Jacobi identity* in the theory of Lie algebras), which make one less certain as to the existence or uniqueness of various proposed generalisations of the representations (4.48) or (4.49). For instance, in the free ≤ 3-step nilpotent group, it turns out that for representations of the form

$$g = (\prod_{j=1}^{n} g_j^{m_j})(\prod_{1 \leq i < j \leq n} [g_i, g_j]^{m_{[i,j]}})(\prod_{1 \leq i < j < k \leq n} [[g_i, g_j], g_k]^{n_{[[i,j],k]}})$$

one has uniqueness but not existence (e.g. even in the simplest case $n = 3$, there is no place in this representation for, say, $[[g_1, g_3], g_2]$ or $[[g_1, g_2], g_2]$), but if one tries to insert more triple commutators into the representation to make up for this, one has to be careful not to lose uniqueness due to identities such as (4.50). One can paste these in by *ad hoc* means in the $s = 3$ case, but the $s = 4$ case looks more fearsome still, especially now that the quadruple commutators split into several distinct-looking species such as $[[g_i, g_j], [g_k, g_l]]$ and $[[[g_i, g_j], g_k], g_l]$ which are nevertheless still related to each other by identities such as (4.50). While one can eventually disentangle this mess for any fixed n and s by a finite amount of combinatorial computation, it is not immediately obvious how to give an explicit description of $\mathcal{F}_{\leq s}(g_1, \ldots, g_n)$ uniformly in n and s.

Nevertheless, it turns out that one can give a reasonably tractable description of this group if one takes a *polycyclic perspective* rather than a nilpotent one - i.e. one views the free nilpotent group as a tower of group extensions of the trivial group by the cyclic group $\mathbf{Z}$. This seems to be a fairly standard observation in group theory - I found it in [**MaKaSo2004**] and [**Le2009**] - but seems not to be so widely known outside of that field, so I wanted to record it here.

4.13.1. Generalisation. The first step is to generalise the concept of a free nilpotent group to one where the generators have different "degrees". Define a *graded sequence* to be a finite ordered sequence $(g_\alpha)_{\alpha \in A}$ of formal group elements g_α, indexed by a finite, totally ordered set A, where each g_α is assigned a positive integer $\deg(g_\alpha)$, which we call the *degree* of g_α. We then define the degree of any formal iterated commutator of the g_α by declaring the degree of $[g, h]$ to be the sum of the degrees of g and h. Thus for instance $[[g_{\alpha_1}, g_{\alpha_2}], g_{\alpha_3}]$

has degree $\deg(g_{\alpha_1}) + \deg(g_{\alpha_2}) + \deg(g_{\alpha_3})$. (The ordering on A is not presently important, but will become useful for the polycyclic representation; note that such ordering has already appeared implicitly in (4.48) and (4.49).)

Define the *free $\leq s$-step nilpotent group* $\mathcal{F}_{\leq s}((g_\alpha)_{\alpha\in A})$ generated by a graded sequence $(g_\alpha)_{\alpha\in A}$ to be the group generated by the g_α, subject to the constraint that any iterated commutator of the g_α of degree greater than s is trivial. Thus the free group $\mathcal{F}_{\leq s}(g_1,\ldots,g_k)$ corresponds to the case when all the g_i are assigned a degree of 1.

Note that any element of a graded sequence of degree greater than s is automatically trivial (we view it as a 0-fold commutator of itself) and so can be automatically discarded from that sequence.

We will recursively define the free $\leq s$-step nilpotent group of some graded sequence $(g_\alpha)_{\alpha\in A}$ in terms of simpler sequences, which have fewer low-degree terms at the expense of introducing higher-degree terms, though as mentioned earlier there is no need to introduce terms of degree larger than s. Eventually this process exhausts the sequence, and at that point the free nilpotent group will be completely described.

4.13.2. Shift. It is convenient to introduce the iterated commutators $[g,mh]$ for $m = 0,1,2,\ldots$ by declaring $[g,0h] := g$ and $[g,(m+1)h] := [[g,mh],h]$, thus for instance $[g,3h] = [[[g,h],h],h]$.

Definition 4.13.1 (Shift)**.** Let $s \geq 1$ be an integer, let $(g_\alpha)_{\alpha\in A}$ be a non-empty graded sequence, and let α_0 be the minimal element of A. We define the (degree $\leq s$) *shift* $(g_\alpha)_{\alpha\in A'}$ of $(g_\alpha)_{\alpha\in A}$ by defining A' to be formed from A by removing α_0, and then adding at the end of A all commutators $[\beta,m\alpha_0]$ of degree at most s, where $\beta \in A\backslash\{\alpha_0\}$ and $m \geq 1$. For sake of concreteness we order these commutators lexicographically, so that $[\beta,m\alpha_0] \geq [\beta',m'\alpha_0]$ if $\beta > \beta'$, or if $\beta = \beta'$ and $m > m'$. (These commutators are also considered to be larger than any element of $A\backslash\{\alpha_0\}$). We give each $[\beta,m\alpha_0]$ a degree of $\deg(\beta) + m\deg(\alpha_0)$, and define the group element $g_{[\beta,m\alpha_0]}$ to be $[g_\beta, mg_{\alpha_0}]$.

Example 4.13.2. If $s \leq 3$, and the graded sequence g_a, g_b, g_c consists entirely of elements of degree 1, then the shift of this sequence is given

by

$$g_b, g_c, g_{[b,a]}, g_{[b,2a]}, g_{[c,a]}, g_{[c,2a]}$$

where $[b,a], [c,a]$ have degree 2, and $[b,2a]$, $[c,2a]$ have degree 3, and $g_{[b,a]} = [g_b, g_a]$, $g_{[b,2a]} = [g_b, 2g_a]$, etc.

The key lemma is then

Lemma 4.13.3 (Recursive description of free group)**.** *Let $s \geq 1$ be an integer, let $(g_\alpha)_{\alpha \in A}$ be a non-empty graded sequence, and let α_0 be the minimal element of A. Let $(g_\alpha)_{\alpha \in A'}$ be the shift of of $(g_\alpha)_{\alpha \in A}$. Then $\mathcal{F}_{\leq s}((g_\alpha)_{\alpha \in A})$ is generated by g_{α_0} and $\mathcal{F}_{\leq s}((g_\alpha)_{\alpha \in A'})$, and furthermore the latter group is a normal subgroup of $\mathcal{F}_{\leq s}((g_\alpha)_{\alpha \in A})$ that does not contain g_{α_0}. In other words, we have a semi-direct product representation*

$$\mathcal{F}_{\leq s}((g_\alpha)_{\alpha \in A}) = \mathbf{Z} \ltimes \mathcal{F}_{\leq s}((g_\alpha)_{\alpha \in A'})$$

with g_{α_0} being identified with $(1, id)$ and the action of $\mathbf{Z}$ being given by the conjugation action of g_{α_0}. In particular, every element g in $\mathcal{F}_{\leq s}((g_\alpha)_{\alpha \in A})$ can be uniquely expressed as $g = g_\alpha^{n_\alpha} g'$, where $g' \in \mathcal{F}_{\leq s}((g_\alpha)_{\alpha \in A'})$.

Proof. It is clear that $\mathcal{F}_{\leq s}((g_\alpha)_{\alpha \in A'})$ is a subgroup of $\mathcal{F}_{\leq s}((g_\alpha)_{\alpha \in A})$, and that it together with g_{α_0} generates $\mathcal{F}_{\leq s}((g_\alpha)_{\alpha \in A})$. To show that this subgroup is normal, it thus suffices to show that the conjugation action of g_{α_0} and $g_{\alpha_0}^{-1}$ preserve $\mathcal{F}_{\leq s}((g_\alpha)_{\alpha \in A'})$. It suffices to check this on generators. But this is clear from the identity

$$g_{\alpha_0}^{-1}[g_\beta, m g_{\alpha_0}] g_{\alpha_0} = [g_\beta, m g_{\alpha_0}][g_\beta, (m+1) g_{\alpha_0}]$$

and its inverse

$$g_{\alpha_0}[g_\beta, m g_{\alpha_0}] g_{\alpha_0}^{-1} = [g_\beta, m g_{\alpha_0}][g_\beta, (m+1) g_{\alpha_0}]^{-1}[g_\beta, (m+2) g_{\alpha_0}] \ldots$$

(note that the product terminates in finite time due to nilpotency).

Finally, we need to show that g_{α_0} is not contained in $\mathcal{F}_{\leq s}((g_\alpha)_{\alpha \in A'})$. But because the conjugation action of g_{α_0} preserves the latter group, we can form the semidirect product $G := \mathbf{Z} \ltimes \mathcal{F}_{\leq s}((g_\alpha)_{\alpha \in A'})$. By the universal nature of the free group, there must thus be a homomorphism from $\mathcal{F}_{\leq s}((g_\alpha)_{\alpha \in A})$ to G which maps g_{α_0} to $(1, id)$ and maps $\mathcal{F}_{\leq s}((g_\alpha)_{\alpha \in A'})$ to $0 \times \mathcal{F}_{\leq s}((g_\alpha)_{\alpha \in A'})$. This implies that g_{α_0} cannot lie in $\mathcal{F}_{\leq s}((g_\alpha)_{\alpha \in A'})$, and the claim follows. □

We can now iterate this. Observe that every time one shifts a non-empty graded sequence, one removes one element (the minimal element g_{α_0}) but replaces it with zero or more elements of higher degree. Iterating this process, we eventually run out of elements of degree one, then degree two, and so forth, until the sequence becomes completely empty. We glue together all the elements encountered this way and refer to the full sequence as the *completion* $(g_\alpha)_{\alpha\in\overline{A}}$ of the original sequence $(g_\alpha)_{\alpha\in A}$. As a corollary of the above lemma we thus have

Corollary 4.13.4 (Explicit description of free nilpotent group)**.** *Let $s \geq 1$ be an integer, and let $(g_\alpha)_{\alpha\in A}$ be a graded sequence. Then every element g of $\mathcal{F}_{\leq s}((g_\alpha)_{\alpha\in A})$ can be represented uniquely as*

$$\prod_{\alpha\in\overline{A}} g_\alpha^{n_\alpha}$$

where n_α is an integer, and $\overline{A}$ is the completion of A.

Example 4.13.5. We continue with the sequence g_a, g_b, g_c from Example 4.13.2, with $s = 3$. We already saw that shifting once yielded the sequence

$$g_b, g_c, g_{[b,a]}, g_{[b,2a]}, g_{[c,a]}, g_{[c,2a]}.$$

Another shift gives

$$g_c, g_{[b,a]}, g_{[b,2a]}, g_{[c,a]}, g_{[c,2a]}, g_{[c,b]}, g_{[c,2b]}, g_{[[b,a],b]}, g_{[[c,a],b]},$$

and shifting again gives

$$g_{[b,a]}, g_{[b,2a]}, g_{[c,a]}, g_{[c,2a]}, g_{[c,b]}, g_{[c,2b]}, g_{[[b,a],b]}, g_{[[c,a],b]}, g_{[[b,a],c]}, g_{[[c,a],c]}.$$

At this point, all remaining terms in the sequence have degree at least two, and further shifting simply removes the first element without adding any new elements. Thus the completion is

$$g_a, g_b, g_c, g_{[b,a]}, g_{[b,2a]}, g_{[c,a]}, g_{[c,2a]},$$

$$g_{[c,b]}, g_{[c,2b]}, g_{[[b,a],b]}, g_{[[c,a],b]}, g_{[[b,a],c]}, g_{[[c,a],c]}$$

and every element of $\mathcal{F}_{\leq 3}(g_a, g_b, g_c)$ can be uniquely expressed as

$$g_a^{n_a} g_b^{n_b} g_c^{b_c} [g_b, g_a]^{n_{[b,a]}} [g_b, 2g_a]^{n_{[b,2a]}}$$

$$[g_c, g_a]^{n_{[c,a]}} [g_c, 2g_a]^{n_{[c,2a]}} [g_c, g_b]^{n_{[c,b]}} [g_c, 2g_b]^{n_{[c,2b]}}$$

$$[[g_b, g_a], g_b]^{n_{[[b,a],b]}} [[g_c, g_a], g_b]^{n_{[[c,a],b]}} [[g_b, g_a], g_c]^{n_{[[b,a],c]}} [[g_c, g_a], g_c]^{n_{[[c,a],c]}}.$$

In [**Le2009**], a related argument was used to expand bracket polynomials (a generalisation of ordinary polynomials in which the integer part operation $x \mapsto \lfloor x \rfloor$ is introduced) of degree $\leq s$ in several variables $(x_\alpha)_{\alpha \in A}$ into a canonical basis $(x_\alpha)_{\alpha \in \overline{A}}$, where $\overline{A}$ is the same completion of A that was encountered here. This was used to show a close connection between such bracket polynomials and nilpotent groups (or more precisely, nilsequences).

Notes. This article first appeared at `terrytao.wordpress.com/2009/12/21`. Thanks to Dylan Thurston for corrections.

Bibliography

[AgKaSa2004] M. Agrawal, N. Kayal, N. Saxena, *PRIMES is in P*, Annals of Mathematics **160** (2004), no. 2, pp. 781-793.

[AjSz1974] M. Ajtai, E. Szemerédi, *Sets of lattice points that form no squares*, Stud. Sci. Math. Hungar. 9 (1974), 9–11 (1975).

[AlDuLeRoYu1994] N. Alon, R. Duke, H. Lefmann, Y. Rödl, R. Yuster, *The algorithmic aspects of the regularity lemma*, J. Algorithms **16** (1994), no. 1, 80–10.

[AlSh2008] N. Alon, A. Shapira, *Every monotone graph property is testable*, SIAM J. Comput. **38** (2008), no. 2, 505–522.

[AlSp2008] N. Alon, J. Spencer, The probabilistic method. Third edition. With an appendix on the life and work of Paul Erdos. Wiley-Interscience Series in Discrete Mathematics and Optimization. John Wiley & Sons, Inc., Hoboken, NJ, 2008.

[Au2008] T. Austin, *On exchangeable random variables and the statistics of large graphs and hypergraphs*, Probab. Surv. **5** (2008), 80–145.

[Au2009] T. Austin, *Deducing the multidimensional Szemeredi Theorem from an infinitary removal lemma*, preprint.

[Au2009b] T. Austin, *Deducing the Density Hales-Jewett Theorem from an infinitary removal lemma*, preprint.

[AuTa2010] T. Austin, T. Tao, *On the testability and repair of hereditary hypergraph properties*, preprint.

[Ax1968] J. Ax, *The elementary theory of finite fields*, Ann. of Math. **88** (1968) 239–271.

[BaGiSo1975] T. Baker, J. Gill, R. Solovay, *Relativizations of the $\mathcal{P} =? \mathcal{NP}$ question*, SIAM J. Comput. 4 (1975), no. 4, 431–442.

[Be1975] W. Beckner, *Inequalities in Fourier analysis*, Ann. of Math. **102** (1975), no. 1, 159–182.

[BeTaZi2009] V. Bergelson, T. Tao, T. Ziegler, *An inverse theorem for the uniformity seminorms associated with the action of F^ω*, preprint.

[BeLo1976] J. Bergh, J. Löfström, Interpolation spaces. An introduction. Grundlehren der Mathematischen Wissenschaften, No. 223. Springer-Verlag, Berlin-New York, 1976.

[BiRo1962] A. Białynicki-Birula, M. Rosenlicht, *Injective morphisms of real algebraic varieties*, Proc. Amer. Math. Soc. **13** (1962) 200–203.

[BoKe1996] E. Bogomolny, J. Keating, *Random matrix theory and the Riemann zeros. II. n-point correlations*, Nonlinearity **9** (1996), no. 4, 911–935.

[Bo1969] A. Borel, *Injective endomorphisms of algebraic varieties*, Arch. Math. (Basel) **20** 1969 531–537.

[Bo1999] J. Bourgain, *On the dimension of Kakeya sets and related maximal inequalities*, Geom. Funct. Anal. **9** (1999), no. 2, 256–282.

[BoBr2003] J. Bourgain, H. Brezis, *On the equation* $\operatorname{div} Y = f$ *and application to control of phases*, J. Amer. Math. Soc. **16** (2003), no. 2, 393–426.

[BudePvaR2008] G. Buskes, B. de Pagter, A. van Rooij, *The Loomis-Sikorski theorem revisited*, Algebra Universalis **58** (2008), 413–426.

[ChPa2009] T. Chen, N. Pavlovic, *The quintic NLS as the mean field limit of a boson gas with three-body interactions*, preprint.

[ClEdGuShWe1990] K. Clarkson, H. Edelsbrunner, L. Guibas, M. Sharir, E. Welzl, *Combinatorial complexity bounds for arrangements of curves and spheres*, Discrete Comput. Geom. **5** (1990), no. 2, 99–160.

[Co1989] J. B. Conrey, *More than two fifths of the zeros of the Riemann zeta function are on the critical line*, J. Reine Angew. Math. **399** (1989), 1–26.

[Dy1970] F. Dyson, *Correlations between eigenvalues of a random matrix*, Comm. Math. Phys. **19** 1970 235–250.

[ElSz2008] G. Elek, B. Szegedy, *A measure-theoretic approach to the theory of dense hypergraphs*, preprint.

[ElObTa2009] J. Ellenberg, R. Oberlin, T. Tao, *The Kakeya set and maximal conjectures for algebraic varieties over finite fields*, preprint.

[ElVeWe2009] J. Ellenberg, A. Venkatesh, C. Westerland, *Homological stability for Hurwitz spaces and the Cohen-Lenstra conjecture over function fields*, preprint.

[ErKa1940] P. Erdös, M. Kac, *The Gaussian Law of Errors in the Theory of Additive Number Theoretic Functions*, American Journal of Mathematics, volume 62, No. 1/4, (1940), pages 738742.

[EsKePoVe2008] L. Escauriaza, C. E. Kenig, G. Ponce, L. Vega, *Hardy's uncertainty principle, convexity and Schrödinger evolutions*, J. Eur. Math. Soc. (JEMS) **10** (2008), no. 4, 883–907.

[Fa2003] K. Falconer, Fractal geometry, Mathematical foundations and applications. Second edition. John Wiley & Sons, Inc., Hoboken, NJ, 2003.

[FeSt1972] C. Fefferman, E. M. Stein, *H^p spaces of several variables*, Acta Math. **129** (1972), no. 3-4, 137–193.

[FiMaSh2007] E. Fischer, A. Matsliach, A. Shapira, *Approximate Hypergraph Partitioning and Applications*, Proc. of FOCS 2007, 579–589.

[Fo2000] G. Folland, Real Analysis, Modern techniques and their applications. Second edition. Pure and Applied Mathematics (New York). A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, 1999.

[Fo1955] E. Følner, *On groups with full Banach mean value*, Math. Scand. **3** (1955), 243–254

[Fo1974] J. Fournier, *Majorants and L^p norms*, Israel J. Math. **18** (1974), 157–166.

[Fr1973] G. Freiman, *Groups and the inverse problems of additive number theory*, Number-theoretic studies in the Markov spectrum and in the structural theory of set addition, pp. 175183. Kalinin. Gos. Univ., Moscow, 1973.

[Fu1977] H. Furstenberg, *Ergodic behavior of diagonal measures and a theorem of Szemerédi on arithmetic progressions*, J. Analyse Math. **31** (1977), 204–256.

[FuKa1989] H. Furstenberg, Y. Katznelson, *A density version of the Hales-Jewett theorem for $k = 3$*, Graph theory and combinatorics (Cambridge, 1988). Discrete Math. **75** (1989), no. 1-3, 227–241.

[FuKa1991] H. Furstenberg, Y. Katznelson, *A density version of the Hales-Jewett theorem*, J. Anal. Math. **57** (1991), 64–119.

[GiTr1998] D. Gilbarg, N. Trudinger, Elliptic partial differential equations of second order. Reprint of the 1998 edition. Classics in Mathematics. Springer-Verlag, Berlin, 2001.

[GoMo1987] D. Goldston, H. Montgomery, *Pair correlation of zeros and primes in short intervals*, Analytic number theory and Diophantine problems (Stillwater, OK, 1984), 183–203, Progr. Math., 70, Birkhuser Boston, Boston, MA, 1987.

[Go1993] W. T. Gowers, B. Maurey, *The unconditional basic sequence problem*, J. Amer. Math. Soc. **6** (1993), no. 4, 851–874.

[Gr1992] A. Granville, *On elementary proofs of the prime number theorem for arithmetic progressions, without characters*, Proceedings of the

Amalfi Conference on Analytic Number Theory (Maiori, 1989), 157–194, Univ. Salerno, Salerno, 1992.

[Gr2005] A. Granville, *It is easy to determine whether a given integer is prime*, Bull. Amer. Math. Soc. (N.S.) **42** (2005), no. 1, 3–38.

[GrSo2007] A. Granville, K. Soundararajan, *Large character sums: pretentious characters and the Pólya-Vinogradov theorem*, J. Amer. Math. Soc. 20 (2007), no. 2, 357–384.

[GrTa2007] B. Green, T. Tao, *The distribution of polynomials over finite fields, with applications to the Gowers norms*, preprint.

[GrTaZi2009] B. Green, T. Tao, T. Ziegler, *An inverse theorem for the Gowers U^4 norm*, preprint.

[Gr1999] M. Gromov, *Endomorphisms of symbolic algebraic varieties*, J. Eur. Math. Soc. (JEMS) **1** (1999), no. 2, 109–197.

[Gr1966] A. Grothendieck, *Éléments de géométrie algébrique. IV. Étude locale des schémas et des morphismes de schémas. III.*, Inst. Hautes Études Sci. Publ. Math. No. 28 1966 255 pp.

[GyMaRu2008] K. Gyarmati, M. Matolcsi, I. Ruzsa, *Plünnecke's inequality for different summands*, Building bridges, 309–320, Bolyai Soc. Math. Stud., 19, Springer, Berlin, 2008

[Ho1990] L. Hörmander, The analysis of linear partial differential operators. I.-IV. Reprint of the second (1990) edition. Classics in Mathematics. Springer-Verlag, Berlin, 2003.

[HoKrPeVi2006] J. Hough, M. Krishnapur, Y. Peres, B. Virág, Determinantal processes and independence, Probab. Surv. **3** (2006), 206–229

[Hr2009] E. Hrushovski, *Stable group theory and approximate subgroups*, preprint.

[Hu1968] R. Hunt, *On the convergence of Fourier series*, 1968 Orthogonal Expansions and their Continuous Analogues (Proc. Conf., Edwardsville, Ill., 1967) pp. 235–255 Southern Illinois Univ. Press, Carbondale, Ill.

[Is2006] Y. Ishigami, *A Simple Regularization of Hypergraphs*, preprint.

[IwKo2004] H. Iwaniec, E. Kowalski, Analytic number theory. American Mathematical Society Colloquium Publications, 53. American Mathematical Society, Providence, RI, 2004.

[Jo1986] D. Joyner, *Distribution theorems of L-functions*, Pitman Research Notes in Mathematics Series, 142. Longman Scientific & Technical, Harlow; John Wiley & Sons, Inc., New York, 1986

[KaVe1983] V. Kaimanovich, A. Vershik, *Random walks on discrete groups: boundary and entropy*, Ann. Probab. 11 (1983), no. 3, 457–490.

[KiVi2009] R. Killip, M. Visan, *Nonlinear Schrödinger Equations at critical regularity*, preprint.

[KiScSt2008] K. Kirkpatrick, B. Schlein, G. Staffilani, *Derivation of the two dimensional nonlinear Schrodinger equation from many body quantum dynamics*, preprint.

[KlMa2008] S. Klainerman, M. Machedon, *On the uniqueness of solutions to the Gross-Pitaevskii hierarchy*, Comm. Math. Phys. 279 (2008), no. 1, 169–185.

[Ku1999] K. Kurdyka, *Injective endomorphisms of real algebraic sets are surjective*, Math. Ann. **313** (1999), no. 1, 69–82.

[La2001] I. Łaba, *Fuglede's conjecture for a union of two intervals*, Proc. Amer. Math. Soc. **129** (2001), no. 10, 2965–2972.

[LaTa2001] I. Łaba, T. Tao, *An x-ray transform estimate in* $\mathbf{R}^n$, Rev. Mat. Iberoamericana **17** (2001), no. 2, 375–407.

[La1996] A. Laurincikas, Limit theorems for the Riemann zeta-function. Mathematics and its Applications, 352. Kluwer Academic Publishers Group, Dordrecht, 1996.

[Le2009] A. Leibman, *A canonical form and the distribution of values of generalised polynomials,* preprint.

[Le2000] V. Lev, *Restricted Set Addition in Groups I: The Classical Setting*, Journal of the London Mathematical Society 2000 62(1):27-40

[LiLo2000] E. Lieb, E. Loss, Analysis. Second edition. Graduate Studies in Mathematics, 14. American Mathematical Society, Providence, RI, 2001.

[Li1853] J. Liouville, *Sur l'equation aux differences partielles*, J. Math. Pure et Appl. **18** (1853), 71–74.

[LiTz1971] J. Lindenstrauss, L. Tzafriri, *On the complemented subspaces problem*, Israel J. Math. **9** (1971) 263–269.

[Lo1946] L. H. Loomis, *On the representation of* σ*-complete Boolean algebras*, Bull. Amer. Math Soc. **53**, (1947). 757–760.

[LoSz2007] L. Lovász, B. Szegedy, *Szemerédi's lemma for the analyst*, Geom. Funct. Anal. **17** (2007), no. 1, 252–270

[Ly2003] R. Lyons, *Determinantal probability measures*, Publ. Math. Inst. Hautes tudes Sci. No. **98** (2003), 167–212.

[Ma2008] M. Madiman, *On the entropy of sums*, preprint.

[MaKaSo2004] W. Magnus, A. Karras, and D. Solitar, Presentations of Groups in Terms of Generators and Relations, Dover Publications, 2004.

[Ma1999] R. Matthews, *The power of one*, New Scientist, 10 July 1999, p. 26.

[Ma1995] P. Mattila, Geometry of sets and measures in Euclidean spaces. Fractals and rectifiability. Cambridge Studies in Advanced Mathematics, 44. Cambridge University Press, Cambridge, 1995.

[Ma1959] B. Mazur, *On embeddings of spheres*, Bull. Amer. Math. Soc. **65** (1959), 59–65.

[Mo2009] R. Moser, *A constructive proof of the Lovász local lemma*, Proceedings of the 41st annual ACM symposium on Theory of computing 2009, Pages 343–350.

[Na1964] I. Namioka, *Følner's conditions for amenable semi-groups*, Math. Scand. **15** (1964), 18–28.

[ON1963] R. O'Neil, *Convolution operators and $L(p, q)$ spaces*, Duke Math. J. **30** 1963 129–142.

[Po2009] D.H.J. Polymath, A new proof of the density Hales-Jewett theorem, preprint.

[PCM] The Princeton companion to mathematics. Edited by Timothy Gowers, June Barrow-Green and Imre Leader. Princeton University Press, Princeton, NJ, 2008.

[Ra1959] H. Rademacher, *On the Phragmén-Lindelöf theorem and some applications*, Math. Z **72** (1959/1960), 192–204.

[RaRu1997] A. Razborov, S. Rudich, *Natural proofs*, 26th Annual ACM Symposium on the Theory of Computing (STOC '94) (Montreal, PQ, 1994). J. Comput. System Sci. **55** (1997), no. 1, part 1, 24–35.

[Ro1982] J.-P. Rosay, *Injective holomorphic mappings*, Amer. Math. Monthly **89** (1982), no. 8, 587–588.

[Ro1953] K. Roth, *On certain sets of integers, I*, Journal of the London Mathematical Society 28 (1953), 104-109.

[Ru1962] W. Rudin, Fourier analysis on groups. Reprint of the 1962 original. Wiley Classics Library. A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, 1990. x+285 pp

[Ru1995] W. Rudin, *Injective polynomial maps are automorphisms*, Amer. Math. Monthly **102** (1995), no. 6, 540–543.

[Ru1989] I. Ruzsa, *An application of graph theory to additive number theory*, Sci. Ser. A Math. Sci. (N.S.) **3** (1989), 97–109.

[RuSz1978] I. Ruzsa, E. Szemerédi, *Triple systems with no six points carrying three triangles*, Colloq. Math. Soc. J. Bolyai, **18** (1978), 939–945.

[Sc2006] B. Schlein, *Dynamics of Bose-Einstein Condensates*, preprint.

[Se2009] J. P. Serre, *How to use finite fields for problems concerning infinite fields*, preprint.

[So2000] A. Soshnikov, *Determinantal random point fields* Uspekhi Mat. Nauk 55 (2000), no. 5(335), 107–160; translation in Russian Math. Surveys 55 (2000), no. 5, 923–975.

[St1961] E. M. Stein, *On limits of seqences of operators*, Ann. of Math. **74** (1961) 140–170.

[St1970] E. M. Stein, Singular integrals and differentiability properties of functions. Princeton Mathematical Series, No. 30 Princeton University Press, Princeton, N.J. 1970

[St1993] E. M. Stein, Harmonic analysis: real-variable methods, orthogonality, and oscillatory integrals. With the assistance of Timothy S. Murphy. Princeton Mathematical Series, 43. Monographs in Harmonic Analysis, III. Princeton University Press, Princeton, NJ, 1993.

[St1969] S. A. Stepanov, *The number of points of a hyperelliptic curve over a finite prime field*, Izv. Akad. Nauk SSSR Ser. Mat. **33** (1969) 1171–1181

[St1948] A. H. Stone, *Paracompactness and product spaces*, Bull. Amer. Math. Soc. **54**, (1948). 977–982.

[Sz2009] B. Szegedy, *Higher order Fourier analysis as an algebraic theory I*, preprint.

[SzTr1983] E. Szemerédi, W. Trotter, *Extremal problems in discrete geometry*, Combinatorica **3** (1983), no. 3-4, 381–392.

[Ta1996] M. Talagrand, *Concentration of measure and isoperimetric inequalities in product spaces*, Inst. Hautes Études Sci. Publ. Math. No. 81 (1995), 73–205

[Ta2005] M. Talagrand, The generic chaining. Upper and lower bounds of stochastic processes. Springer Monographs in Mathematics. Springer-Verlag, Berlin, 2005.

[Ta1951] A. Tarski, A decision method for elementary algebra and geometry, 2nd ed. University of California Press, Berkeley and Los Angeles, Calif., 1951.

[Ta] T. Tao, *Summability of functions*, unpublished preprint.

[Ta2006] T. Tao, Nonlinear dispersive equations. Local and global analysis. CBMS Regional Conference Series in Mathematics, 106. Published for the Conference Board of the Mathematical Sciences, Washington, DC; by the American Mathematical Society, Providence, RI, 2006.

[Ta2006b] T. Tao, *A quantitative ergodic theory proof of Szemerédi's theorem*, Electron. J. Combin. **13** (2006), no. 1.

[Ta2006c] T. Tao, *Szemerédi's regularity lemma revisited*, Contrib. Discrete Math. **1** (2006), no. 1, 8–28

[Ta2007] T. Tao, *A correspondence principle between (hyper)graph theory and probability theory, and the (hyper)graph removal lemma*, J. Anal. Math. **103** (2007), 1–45

[Ta2007b] T. Tao, *Structure and randomness in combinatorics*, Proceedings of the 48th annual symposium on Foundations of Computer Science (FOCS) 2007, 3-18.

[Ta2008] T. Tao, Structure and Randomness: pages from year one of a mathematical blog, American Mathematical Society, Providence RI, 2008.

[Ta2009] T. Tao, Poincaré's Legacies: pages from year two of a mathematical blog, Vols. I, II, American Mathematical Society, Providence RI, 2009.

[Ta2010] T. Tao, *The high exponent limit $p \to \infty$ for the one-dimensional nonlinear wave equation*, preprint.

[Ta2010b] T. Tao, *A remark on partial sums involving the Möbius function*, preprint.

[Ta2010c] T. Tao, *Sumset and inverse sumset theorems for Shannon entropy*, preprint.

[TaVu2006] T. Tao, V. Vu, *On random ± 1 matrices: singularity and determinant*, Random Structures Algorithms **28** (2006), no. 1, 1–23

[TaVu2006b] T. Tao, V. Vu, Additive combinatorics. Cambridge Studies in Advanced Mathematics, 105. Cambridge University Press, Cambridge, 2006.

[TaVu2007] T. Tao, V. Vu, On the singularity probability of random Bernoulli matrices, J. Amer. Math. Soc. **20** (2007), 603–628.

[TaWr2003] T. Tao, J. Wright, *L^p improving bounds for averages along curves*, J. Amer. Math. Soc. **16** (2003), no. 3, 605–638.

[Th1994] W. Thurston, *On proof and progress in mathematics*, Bull. Amer. Math. Soc. (N.S.) **30** (1994), no. 2, 161–177.

[To2005] C. Toth, *The Szemerédi-Trotter Theorem in the Complex Plane*, preprint.

[Uc1982] A. Uchiyama, *A constructive proof of the Fefferman-Stein decomposition of BMO* (R^n), Acta Math. **148** (1982), 215–241.

[VuWoWo2010] V. Vu, M. Wood, P. Wood, *Mapping incidences*, preprint.

[Wo1995] T. Wolff, *An improved bound for Kakeya type maximal functions*, Rev. Mat. Iberoamericana **11** (1995), no. 3, 651–674

[Wo1998] T. Wolff, A mixed norm estimate for the X-ray transform, Rev. Mat. Iberoamericana **14** (1998), no. 3, 561–600.

[Wo2003] T. Wolff, Lectures on harmonic analysis. With a foreword by Charles Fefferman and preface by Izabella Laba. Edited by Laba and

Carol Shubin. University Lecture Series, 29. American Mathematical Society, Providence, RI, 2003. x+137 pp