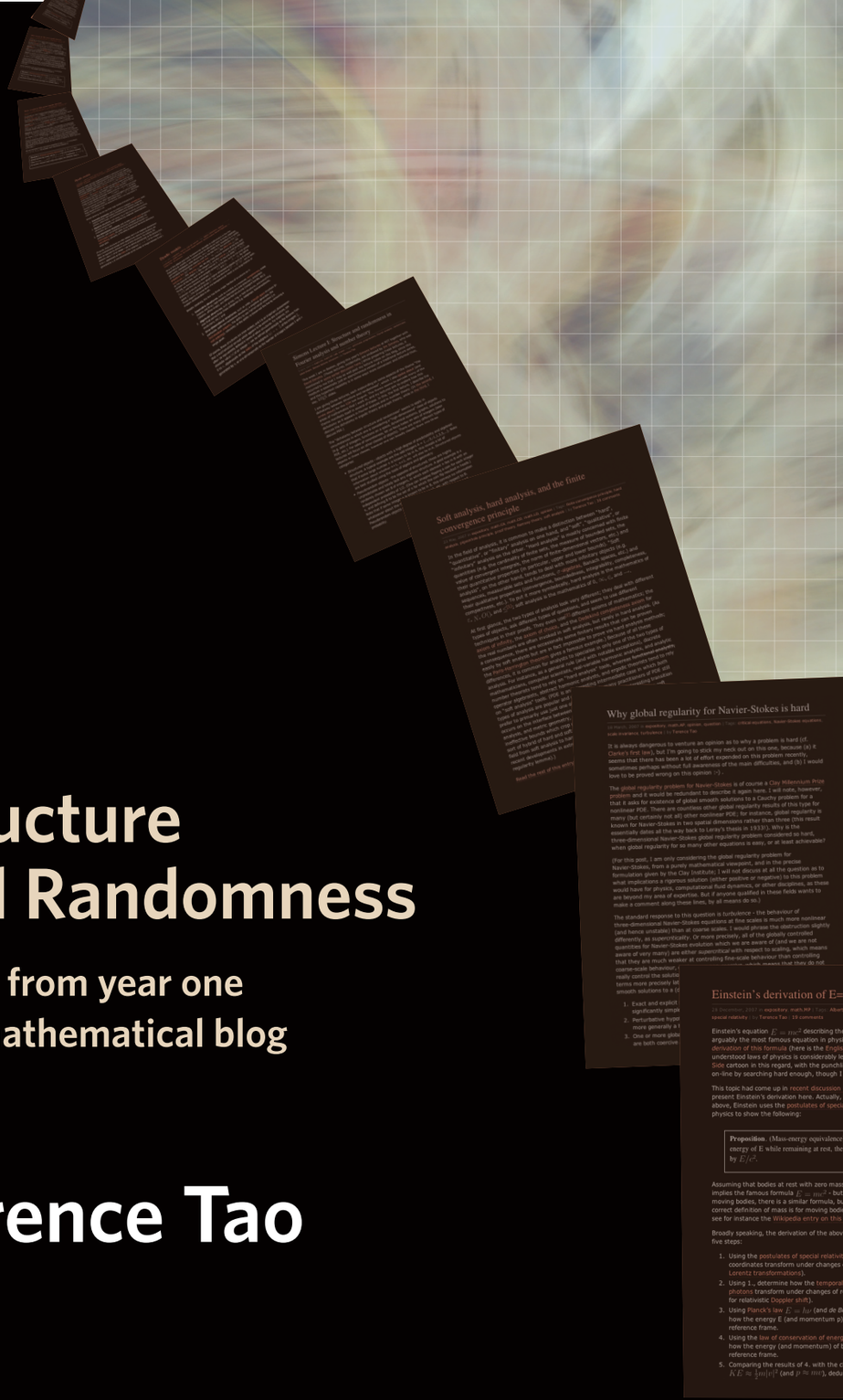


Structure and Randomness

pages from year one of a mathematical blog

Terence Tao

ISBN 978-0-8218-4441-0
9 780821 844410
MBK/50

For additional information and updates on this book, visit
www.ams.org/bookpages/mbk-50

AMS on the Web
www.ams.org

AMERICAN MATHEMATICAL SOCIETY

1

What's new - 2007: Open questions, expository articles, and lecture series from a mathematical blog

Terence Tao

April 24, 2008

[1]The author is supported by NSF grant CCF-0649473 and a grant from the MacArthur foundation.

To my advisor, Eli Stein, for showing me the importance of good exposition;
To my friends, for supporting this experiment;
And to the readers of my blog, for their feedback and contributions.

Contents

Preface

Almost nine years ago, in 1999, I began a "What's new?" page on my UCLA home page in order to keep track of various new additions to that page (e.g. papers, slides, lecture notes, expository "short stories", etc.). At first, these additions were simply listed without any commentary, but after a while I realised that this page was a good place to put a brief description and commentary on each of the mathematical articles that I was uploading to the page. (In short, I had begun blogging on my research, though I did not know this term at the time.)

Every now and then, I received an email from someone who had just read the most recent entry on my "What's new?" page and wanted to make some mathematical or bibliographic comment; this type of valuable feedback was one of the main reasons why I kept maintaining the page. But I did not think to try to encourage more of this feedback until late in 2006, when I posed a question on my "What's new?" page and got a complete solution to that problem within a matter of days. It was then that I began thinking about modernising my web page to a blog format (which a few other mathematicians had already begun doing). On 22 February 2007, I started a blog with the unimaginative name of "What's new" at `erry`ao.wordpress.com; I chose wordpress for a number of reasons, but perhaps the most decisive one was its recent decision to support LaTeX in its blog posts.

It soon became clear that the potential of this blog went beyond my original aim of merely continuing to announce my own papers and research. For instance, by far the most widely read and commented article in my blog in the first month was a non-technical article, "Quantum Mechanics and Tomb Raider" (Section 2.1), which had absolutely nothing to do with my own mathematical work. Encouraged by this, I began to experiment with other types of mathematical content on the blog; discussions of my favourite open problems, informal discussions of mathematical phenomena, principles, or tricks, guest posts by some of my colleagues, and presentations of various lectures and talks, both by myself and by others; and various bits and pieces of advice on pursuing a mathematical career and on mathematical writing. This year, I also have begun placing lecture notes for my graduate classes on my blog.

After a year of mathematical blogging, I can say that the experience has been positive, both for the readers of the blog and for myself. Firstly, the very act of writing a blog article helped me organise and clarify my thoughts on a given mathematical topic, to practice my mathematical writing and exposition skills, and also to inspect the references and other details more carefully. From insightful comments by experts in other fields of mathematics, I have learned of unexpected connections between different fields; in one or two cases, these even led to new research projects and collaborations. From the feedback from readers I obtained a substantial amount of free proofreading, while also discovering what parts of my exposition were unclear or otherwise poorly worded, helping me improve my writing style in the future. It is a truism that one of the best ways to learn a subject is to teach it; and it seems that blogging about a subject comes in as a close second.

In the last year (2007) alone, at least a dozen new active blogs in research mathematics have sprung up. I believe this is an exciting new development in mathematical exposition; research blogs seem to fill an important niche that neither traditional print media (textbooks, research articles, surveys, etc.) nor informal communications (lectures, seminars, conversations at a blackboard, etc.) adequately cover at present.

Indeed, the research blog medium is in some ways the "best of both worlds"; informal, dynamic, and interactive, as with talks and lectures, but also coming with a permanent record, a well defined author, and links to further references, as with the print media. There are bits and pieces of folklore in mathematics, such as the difference between hard and soft analysis (Section 2.3) or the use of dyadic models for non-dyadic situations (Section 2.6) which are passed down from advisor to student, or from collaborator to collaborator, but are too fuzzy and non-rigorous to be discussed in the formal literature; but they can be communicated effectively and efficiently via the semi-formal medium of research blogging.

On the other hand, blog articles still lack the permanence that print articles have, which becomes an issue when one wants to use them in citations. For this and other reasons, I have decided to convert some of my blog articles from 2007 into the book that you are currently reading. Not all of the 93 articles that I wrote in 2007 appear here; some were mostly administrative or otherwise non-mathematical in nature, some were primarily announcements of research papers or other articles which will appear elsewhere, some were contributed guest articles, and some were writeups of lectures by other mathematicians, which it seemed inappropriate to reproduce in a book such as this. Nevertheless, this still left me with 32 articles, which I have converted into print form (replacing hyperlinks with more traditional citations and footnotes, etc.). As a result, this book is not a perfect replica of the blog, but the mathematical content is largely the same. I have paraphrased some of the feedback from comments to the blog in the endnotes to each article, though for various reasons, ranging from lack of space to copyright concerns, not all comments are reproduced here.

The articles here are rather diverse in subject matter, to put it mildly, but I have nevertheless organised them into three categories. The first category concerns various open problems in mathematics that I am fond of; some are of course more difficult than others (see e.g. the article on Navier-Stokes regularity, Section 1.4), and others are rather vague and open-ended, but I find each of them interesting, not only in their own right, but because progress on them is likely to yield insights and techniques that will be useful elsewhere. The second category are the expository articles, which vary from discussions of various well-known results in maths and science (e.g. the nullstellensatz in Section 2.15, or Einstein's equation $E = mc^2$ in Section 2.17), to more philosophical explorations of mathematical ideas, tricks, tools, or principles (e.g. ultrafilters in Section 2.5, or amplification in Section 2.9), to non-technical expositions of various topics in maths and science, from quantum mechanics (Section 2.1) to single-pixel cameras (Section 2.2). Finally, I am including writeups of three lecture series I gave in 2007; my Simons lecture series at MIT on structure on randomness, my Ostrowski lecture at the University of Leiden on compressed sensing, and my Milliman lectures at the University of Washington on additive combinatorics.

In closing, I believe that this experiment with mathematical blogging has been generally successful, and I plan to continue it in the future, and perhaps generating several more books such as this one as a result. I am grateful to all the readers of my blog for supporting this experiment, for supplying invaluable feedback and corrections, and for encouraging projects such as this book conversion.

A remark on notation

One advantage of the blog format is that one can often define a technical term simply by linking to an external web page that contains the definition (e.g. a Wikipedia page). This is unfortunately not so easy to reproduce in print form, and so many standard mathematical technical terms will be used without definition in this book; this is not going to be a self-contained textbook in mathematics, but is instead a loosely connected collection of articles at various levels of technical difficulty. Of course, in the age of the internet, it is not terribly difficult to look up these definitions whenever necessary.

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.

0.0.1 Acknowledgments

Many people have contributed corrections or comments to individual blog articles, and are acknowledged in the end notes to those articles here. Thanks also to harrison, Gil Kalai, Greg Kuperberg, Phu, Jozsef Solymosi, Tom, and Y J for general corrections, reference updates, and formatting suggestions.

Chapter 1

Open problems

1.1 Best bounds for capsets

Perhaps my favourite open question is the problem on the maximal size of a *cap set* - a subset of $\mathbf{F}_3^n$ ($\mathbf{F}_3$ being the finite field of three elements) which contains no lines, or equivalently no non-trivial arithmetic progressions of length three. As an upper bound, one can easily modify the proof of Roth's theorem[Ro1953] to show that cap sets must have size $O(3^n/n)$; see [Me1995]. This of course is better than the trivial bound of 3^n once n is large. In the converse direction, the trivial example $\{0,1\}^n$ shows that cap sets can be as large as 2^n; the current world record is $(2.2174\ldots)^n$, held by Edel[Ed2004]. The gap between these two bounds is rather enormous; I would be very interested in either an improvement of the upper bound to $o(3^n/n)$, or an improvement of the lower bound to $(3-o(1))^n$. (I believe both improvements are true, though a good friend of mine disagrees about the improvement to the lower bound.)

One reason why I find this question important is that it serves as an excellent model for the analogous question of finding large sets without progressions of length three in the interval $\{1,\ldots,N\}$. Here, the best upper bound of $O(N\sqrt{\frac{\log\log N}{\log N}})$ is due to Bourgain[Bo1999] (with a recent improvement to $O(N\frac{(\log\log N)^2}{\log^{2/3} N})$ [Bo2008]), while the best lower bound of $Ne^{-C\sqrt{\log N}}$ is an ancient result of Behrend[Be1946]. Using the finite field heuristic (see Section 2.6) that $\mathbf{F}_3^n$ "behaves like" $\{1,\ldots,3^n\}$, we see that the Bourgain bound should be improvable to $O(\frac{N}{\log N})$, whereas the Edel bound should be improvable to something like $3^n e^{-C\sqrt{n}}$. However, neither argument extends easily to the other setting. Note that a conjecture of Erdős asserts that any set of positive integers whose sum of reciprocals diverges contains arbitrarily long arithmetic progressions; even for progressions of length three, this conjecture is open, and is essentially equivalent (up to log log factors) to the problem of improving the Bourgain bound to $o(\frac{N}{\log N})$.

The Roth bound of $O(3^n/n)$ appears to be the natural limit of the purely Fourier-analytic approach of Roth, and so any breakthrough would be extremely interesting, as it almost certainly would need a radically new idea. The lower bound might be improvable by some sort of algebraic geometry construction, though it is not clear at all how to achieve this.

One can interpret this problem in terms of the wonderful game "Set", in which case the problem is to find the largest number of cards one can put on the table for which nobody has a valid move. As far as I know, the best bounds on the cap set problem in small dimensions are the ones cited in [Ed2004].

There is a variant formulation of the problem which may be a little bit more tractable. Given any $0 < \delta \leq 1$, the fewest number of lines $N(\delta,n)$ in a set of $\mathbf{F}_3^n$ of density at least δ is easily shown to be $(c(\delta)+o(1))3^{2n}$ for some $0 < c(\delta) \leq 1$. (The analogue in $\mathbf{Z}/N\mathbf{Z}$ is trickier; see [Cr2008], [GrSi2008].) The reformulated question is then to get as strong a bound on $c(\delta)$ as one can. For instance, the counterexample $0,1^m \times \mathbf{F}_3^n$ shows that $c(\delta) \ll \delta^{\log_{3/2} 9/2}$, while the Roth-Meshulam argument gives $c(\delta) \gg e^{-C/\delta}$.)

1.1.1 Notes

This article was originally posted on Feb 23, 2007 at

terrytao.wordpress.com/2007/02/23

Thanks to Jordan Ellenberg for suggesting the density formulation of the problem.

Olaf Sisask points out that the result $N(\delta,n) = (c(\delta)+o(1))3^{2n}$ has an elementary proof; by considering sets in $\mathbf{F}_3^n$ of the form $A \times \mathbf{F}_3$ for some $A \subset \mathbf{F}_3^{n-1}$ one can obtain the inequality $N(\delta,n) \leq 3^2 N(\delta,n-1)$, from which the claim easily follows.

Thanks to Ben Green for corrections.

1.2 Noncommutative Freiman theorem

This is another one of my favourite open problems, falling under the heading of *inverse theorems* in arithmetic combinatorics. "Direct" theorems in arithmetic combinatorics take a finite set A in a group or ring and study things like the size of its *sum set* $A+A := \{a+b : a,b \in A\}$ or *product set* $A \cdot A := \{ab : a,b \in A\}$. For example, a typical result in this area is the *sum-product theorem*, which asserts that whenever $A \subset \mathbf{F}_p$ is a subset of a finite field of prime order with $1 \leq |A| \leq p^{1-\delta}$, then

$$\max(|A+A|, |A \cdot A|) \geq |A|^{1+\varepsilon}$$

for some $\varepsilon = \varepsilon(\delta) > 0$. This particular theorem was first proven in [BoGlKo2006] with an earlier partial result in [BoKaTa2004]; more recent and elementary proofs with civilised bounds can be found in [TaVu2006], [GlKo2008], [Ga2008], [KaSh2008]. See Section 3.3.3 for further discussion.

In contrast, inverse theorems in this subject start with a hypothesis that, say, the sum set $A+A$ of an unknown set A is small, and try to deduce structural information about A. A typical goal is to completely classify all sets A for which $A+A$ has comparable size with A. In the case of finite subsets of integers, this is Freiman's theorem[Fr1973], which roughly speaking asserts that if $|A+A| = O(|A|)$, if and only if A is a dense subset of a generalised arithmetic progression P of rank $O(1)$, where we say that A is a dense subset of B if $A \subset B$ and $|B| = O(|A|)$. (The "if and only if" has to be interpreted properly; in either the "if" or the "only if" direction, the implicit constants in the conclusion depend on the implicit constants in the hypothesis, but these dependencies are not inverses of each other.) In the case of finite subsets A of an arbitrary abelian group, we have the Freiman-Green-Ruzsa theorem [GrRu2007], which asserts that $|A+A| = O(|A|)$ if and only if A is a dense subset of a sum $P+H$ of a finite subgroup H and a generalised arithmetic progression P of rank $O(1)$.

One can view these theorems as a "robust" or "rigid" analogue of the classification of finite abelian groups. It is well known that finite abelian groups are direct sums of cyclic groups; the above results basically assert that finite sets that are "nearly groups" in that their sum set is not much larger than the set itself, are (dense subsets of) the direct sums of cyclic groups and a handful of arithmetic progressions.

The open question is to formulate an analogous conjectural classification in the non-abelian setting, thus to conjecture a reasonable classification of finite sets A in a multiplicative group G for which $|A \cdot A| = O(|A|)$. Actually for technical reasons it may be better to use $|A \cdot A \cdot A| = O(|A|)$; I refer to this condition by saying that A has *small tripling*. (Note for instance that if H is a subgroup and x is not in the normaliser of H, then $H \cup \{x\}$ has small doubling but not small tripling. On the other hand, small tripling is known to imply small quadrupling, etc., see e.g. [TaVu2006].) Note that I am not asking for a theorem here - just even stating the right conjecture would be major progress! An if and only if statement might be too ambitious initially: a first step would be to obtain a slightly looser equivalence, creating for each group G and fixed $\varepsilon > 0$ a class $\mathscr{P}$ of sets (depending on some implied constants) for which the following two statements are true:

(i) If A is a finite subset of G with small tripling, then A is a dense subset of $O(|A|^\varepsilon)$ left- or right- translates of a set P of the form $\mathscr{P}$.

(ii) If P is a set of the form $\mathscr{P}$, then there exists a dense subset A of P with small tripling (possibly with a loss of $O(|A|^\varepsilon)$ in the tripling constant).

An obvious candidate for $\mathscr{P}$ is the inverse image in $N(H)$ of a ball in a nilpotent subgroup of $N(H)/H$ of step $O(1)$, where H is a finite subgroup of G and $N(H)$ is the normaliser of H; note that property (ii) is then easy to verify. Let us call this the *standard candidate*. I do not know if this candidate fully suffices, but it seems to be a reasonably good candidate nevertheless. In this direction, some partial results are known:

- For abelian groups G, from the Freiman-Green-Ruzsa theorem, we know that the standard candidate suffices.
- For $G = SL_2(\mathbf{C})$, we know from work of Elekes and Király[ElKi2001] and Chang[Ch2008] that the standard candidate suffices.
- For $G = SL_2(\mathbf{F}_p)$, there is a partial result of Helfgott [He2008], which (roughly speaking) asserts that if A has small tripling, then either A is a dense subset of all of G, or is contained in a proper subgroup of G. It is likely that by pushing this analysis further one would obtain a candidate for $\mathscr{P}$ in this case.
- For $G = SL_3(\mathbf{Z})$, a result of Chang[Ch2008] shows that if A has small tripling, then it is contained in a nilpotent subgroup of G.
- For the lamplighter group $G = \mathbf{Z}/2\mathbf{Z} \wr \mathbf{Z}$, there is a partial result of Lindenstrauss[Li2001] which (roughly speaking) asserts that if A has small tripling, then A cannot be nearly invariant under a small number of shifts. It is also likely that by pushing the analysis further here one would get a good candidate for $\mathscr{P}$ in this case.
- For a free non-abelian group, we know (since the free group embeds into $SL_2(\mathbf{C})$) that the standard candidate suffices; a much stronger estimate in this direction was recently obtained by Razborov [Ra2008].
- For a Heisenberg group G of step 2, there is a result of myself[Ta2008], which shows that sets of small tripling also have small tripling in the abelianisation of G, and are also essentially closed under the antisymmetric form that defines G. This, in conjunction with the Freiman-Green-Ruzsa theorem, gives a characterisation, at least in principle, but it is not particularly explicit, and it may be of interest to work it out further.
- For G torsion-free, there is a partial result of Hamidoune, Lladó, and Serra[HaLlSe1998], which asserts that $|A \cdot A| \geq 2|A| - 1$, and that if $|A \cdot A| \leq 2|A|$ then A is a geometric progression with at most one element removed; in particular, the standard candidate suffices in this case.

These examples do not seem to conclusively suggest what the full classification should be. Based on analogy with the classification of finite simple groups, one might expect the full classification to be complicated, and enormously difficult to prove; on the other hand, the fact that we are in a setting where we are allowed to lose factors of $O(1)$ may mean that the problem is in fact significantly less difficult than that classification. (For instance, all the sporadic simple groups have size $O(1)$ and so even the monster group is "negligible".) Nevertheless, it seems possible to make progress on explicit groups, in particular refining the partial results already obtained for the specific groups mentioned above. An even closer analogy may be with Gromov's theorem [Gr1981] on groups of polynomial growth; in particular, the recent effective proof of this theorem by Kleiner [Kl2008] may prove to be relevant for this problem.

1.2.1 Notes

This article was originally posted on Mar 2, 2007 at

terrytao.wordpress.com/2007/03/02

Thanks to Akshay Venkatesh and Elon Lindenstrauss to pointing out the analogy with Gromov's theorem, and to Harald Helfgott for informative comments.

1.3 Mahler's conjecture for convex bodies

This question in convex geometry has been around for a while; I am fond of it because it attempts to capture the intuitively obvious fact that cubes and octahedra are the "pointiest" possible symmetric convex bodies one can create. Sadly, we still have very few tools to make this intuition rigorous (especially when compared against the assertion that the Euclidean ball is the "roundest" possible convex body, for which we have many rigorous and useful formulations).

To state the conjecture I need a little notation. Suppose we have a symmetric convex body $B \subset \mathbf{R}^d$ in a Euclidean space, thus B is open, convex, bounded, and symmetric around the origin. We can define the *polar body* $B^\circ \subset \mathbf{R}^d$ by

$$B^\circ := \{\xi \in \mathbf{R}^d : x \cdot \xi < 1 \text{ for all } x \in B\}$$

This is another symmetric convex body. One can interpret B as the unit ball of a Banach space norm on $\mathbf{R}^d$, in which case B° is simply the unit ball of the dual norm. The *Mahler volume* $v(B)$ of the body is defined as the product of the volumes of B and its polar body:

$$v(B) := \mathrm{vol}(B)\,\mathrm{vol}(B^\circ).$$

One feature of this Mahler volume is that it is an affine invariant: if $T : \mathbf{R}^d \to \mathbf{R}^d$ is any invertible linear transformation, then TB has the same Mahler volume as B. It is also clear that a body has the same Mahler volume as its polar body. Finally the Mahler volume reacts well to Cartesian products: if $B_1 \subset \mathbf{R}^{d_1}, B_2 \subset \mathbf{R}^{d_2}$ are convex bodies, one can check that

$$v(B_1 \times B_2) = v(B_1)v(B_2)/\binom{d_1+d_2}{d_1}.$$

For the unit Euclidean ball $B^d := \{x \in \mathbf{R}^d : |x| < 1\}$, the Mahler volume is given by the formula

$$v(B^d) = \frac{\Gamma(3/2)^{2d}4^d}{\Gamma(\frac{d}{2}+1)^2} = (2\pi e + o(1))^d d^{-d}$$

while for the unit cube Q^d or the unit octahedron $O_d := (Q^d)^\circ$ the Mahler volume is

$$v(Q^d) = v(O_d) = \frac{4^d}{\Gamma(d+1)} = (4e + o(1))^d d^{-d} = (\frac{4}{2\pi} + o(1))^d v(B^d).$$

One can also think of Q^d, B^d, O_d as the unit balls of the l^∞, l^2, l^1 norms respectively.

The Mahler conjecture asserts that these are the two extreme possibilities for the Mahler volume, thus for all convex bodies $B \subset \mathbf{R}^d$ we should have

$$v(Q^d) = v(O_d) \leq v(B) \leq v(B^d).$$

Intuitively, this means that the Mahler volume is capturing the "roundness" of a convex body, with balls (and affine images of balls, i.e. ellipsoids) being the roundest, and cubes and octahedra (and affine images thereof) being the pointiest.

The upper bound was established by Santaló[Sa1949] (with the three-dimensional case settled much earlier by Blaschke), using the powerful tool of *Steiner symmetrisation*, which basically is a mechanism for making a convex body rounder and rounder,

converging towards a ball. One can quickly verifies that each application of Steiner symmetrisation does not decrease the Mahler volume, and the result easily follows. As a corollary one can show that the ellipsoids are the only bodies which actually attain the maximal Mahler volume. (Several other proofs of this result, now known as the *Blaschke-Santaló inequality*, exist in the literature. It plays an important role in affine geometry, being a model example of an *affine isoperimetric inequality*.)

Somewhat amusingly, one can use Plancherel's theorem to quickly obtain a crude version of this inequality, losing a factor of $O(d)^d$; indeed, as pointed out to me by Bo'az Klartag, one can view the Mahler conjecture as a kind of "exact uncertainty principle". Unfortunately it seems that Fourier-analytic techniques are unable to solve these sorts of "sharp constant" problems (for which one cannot afford to lose unspecified absolute constants).

The lower inequality remains open. In my opinion, the main reason why this conjecture is so difficult is that unlike the upper bound, in which there is essentially only one extremiser up to affine transformations (namely the ball), there are many distinct extremisers for the lower bound - not only the cube and the octahedron (and affine images thereof), but also products of cubes and octahedra, polar bodies of products of cubes and octahedra, products of polar bodies of... well, you get the idea. (As pointed out to me by Gil Kalai, these polytopes are known as *Hanner polytopes*.) It is really difficult to conceive of any sort of flow or optimisation procedure which would converge to exactly these bodies and no others; a radically different type of argument might be needed.

The conjecture was solved for two dimensions by Mahler [Ma1939] but remains open even in three dimensions. If one is willing to lose some factors in the inequality, though, then some partial results are known. Firstly, from John's theorem[Jo1948] one trivially gets a bound of the form $v(B) \geq d^{-d/2} v(B^d)$. A significantly deeper argument of Bourgain and Milman[BoMi1987], gives a bound of the form $v(B) \geq C^{-d} v(B^d)$ for some absolute constant C; this bound is now known as the *reverse Santaló inequality*. A slightly weaker "low-tech" bound of $v(B) \geq (\log_2 d)^{-d} v(B^d)$ was given by Kuperberg[Ku1992], using only elementary methods. The best result currently known is again by Kuperberg[Ku2008], who showed that

$$v(B) \geq \frac{2^d}{\binom{2d}{d}} v(B^d) \geq (\frac{\pi}{4})^{d-1} v(Q^d)$$

using some Gauss-type linking integrals associated to a Minkowski metric in $\mathbf{R}^{d+d}$. In another direction, the Mahler conjecture has also been verified for some special classes of convex bodies, such as zonoids[Re1986] (limits of finite Minkowski sums of line segments) and 1-unconditional convex bodies[SR1981] (those which are symmetric around all coordinate hyperplanes).

There seem to be some other directions to pursue. For instance, it might be possible to show that (say) the unit cube is a *local* minimiser of Mahler volume, or at least that the Mahler volume is stationary with respect to small perturbations of the cube (whatever that means). Another possibility is to locate some reasonable measure of "pointiness" for convex bodies, which is extremised precisely at cubes, octahedra, and products or polar products thereof. Then the task would be reduced to controlling the

Mahler volume by this measure of pointiness.

1.3.1 Notes

This article was originally posted on Mar 8, 2007 at

terrytao.wordpress.com/2007/03/08

The article generated a fair amount of discussion, some of which I summarise below.

Bo'az Klartag points out that the analogous conjecture for non-symmetric bodies - namely, that the minimal Mahler volume is attained when the body is a simplex - may be easier, due to the fact that there is now only one extremiser up to affine transformations. Greg Kuperberg noted that by combining his inequalities from [Ku2008] with the Rogers-Shephard inequality[RoSh1957], that this conjecture is known up to a factor of $(\frac{\pi}{4e})^d$.

Klartag also pointed out that this asymmetric analogue has an equivalent formulation in terms of the Legendre transformation

$$L(f)(\xi) := \sup_{x \in \mathbf{R}^d} x \cdot \xi - f(x)$$

of a convex function $f : \mathbf{R}^d \to (-\infty, +\infty]$ as the pleasant-looking inequality

$$(\int_{\mathbf{R}^d} e^{-f})(\int_{\mathbf{R}^d} e^{-L(f)}) \geq e^{-d},$$

with the minimum being conjecturally attained when $f(x_1, \ldots, x_n)$ is the function which equals $-\frac{n}{2} + \sum_{i=1}^n x_i$ when $\min_i x_i > -1$, and $+\infty$ otherwise (this function models the simplex). One amusing thing about this inequality is that it is automatically implied by the apparently weaker bound

$$(\int_{\mathbf{R}^d} e^{-f})(\int_{\mathbf{R}^d} e^{-L(f)}) \geq (e - o(1))^{-d}$$

in the asymptotic limit $d \to \infty$, thanks to the "tensor power trick" (see 2.9.4). A similar observation also holds for the original Mahler conjecture.

Danny Calegari and Greg Kuperberg have suggested that for the three-dimensional problem at least, some sort of ODE flow on some moduli space of polyhedra (e.g. gradient flow of Mahler volume with respect to some affinely invariant Riemannian metric on this moduli space) may resolve the problem, although the affine-invariance of the problem does make it challenging to even produce a viable candidate for such a flow. But Greg noted that this approach does work in two dimensions, though there are known topological obstructions in four dimensions. The key difficulties with this approach seem to be selecting a metric with favourable curvature properties, and analysing the Morse structure of the Mahler volume functional.

Kuperberg also pointed out that the Blaschke-Santaló inequality has a "detropicalised" version[LuZh1997]; it may turn out that the Mahler conjecture should also be solved by attacking a detropicalised counterpart (which seems to be some sort of exotic Hausdorff-Young type inequality).

I have a set of notes on some of the more elementary aspects of the above theory at [Ta2006b].

1.4 Why global regularity for Navier-Stokes is hard

The global regularity problem for Navier-Stokes is of course a Clay Millennium Prize problem[Fe2006]. It asks for existence of global smooth solutions to a Cauchy problem for a nonlinear PDE. There are countless other global regularity results of this type for many (but certainly not all) other nonlinear PDE; for instance, global regularity is known for Navier-Stokes in two spatial dimensions rather than three (this result essentially dates all the way back to Leray's thesis[Le1933]!). Why is the three-dimensional Navier-Stokes global regularity problem considered so hard, when global regularity for so many other equations is easy, or at least achievable?

For this article, I am only considering the global regularity problem for Navier-Stokes, from a purely mathematical viewpoint, and in the precise formulation given by the Clay Institute; I will not discuss at all the question as to what implications a rigorous solution (either positive or negative) to this problem would have for physics, computational fluid dynamics, or other disciplines, as these are beyond my area of expertise.

The standard response to the above question is *turbulence* - the behaviour of three-dimensional Navier-Stokes equations at fine scales is much more nonlinear (and hence unstable) than at coarse scales. I would phrase the obstruction slightly differently, as *supercriticality*. Or more precisely, all of the globally controlled quantities for Navier-Stokes evolution which we are aware of (and we are not aware of very many) are either *supercritical* with respect to scaling, which means that they are much weaker at controlling fine-scale behaviour than controlling coarse-scale behaviour, or they are *non-coercive*, which means that they do not really control the solution at all, either at coarse scales or at fine. (I'll define these terms more precisely later.) At present, all known methods for obtaining global smooth solutions to a (deterministic) nonlinear PDE Cauchy problem require either

(I) Exact and explicit solutions (or at least an exact, explicit transformation to a significantly simpler PDE or ODE);

(II) Perturbative hypotheses (e.g. small data, data close to a special solution, or more generally a hypothesis which involves an ε somewhere); or

(III) One or more globally controlled quantities (such as the total energy) which are both coercive and either critical or subcritical.

Note that the presence of (I), (II), or (II) are currently *necessary* conditions for a global regularity result, but far from *sufficient*; otherwise, papers on the global regularity problem for various nonlinear PDE would be substantially shorter. In particular, there have been many good, deep, and highly non-trivial papers recently on global regularity for Navier-Stokes, but they all assume either (I), (II) or (II) via additional hypotheses on the data or solution. For instance, in recent years we have seen good results on global regularity assuming (II) (e.g. [KoTa2001]), as well as good results on global regularity assuming (III) (e.g. [EsSeSv2003]); a complete bibilography of recent results is unfortunately too lengthy to be given here.)

The Navier-Stokes global regularity problem for arbitrary large smooth data lacks all of these three ingredients. Reinstating (II) is impossible without changing the statement of the problem, or adding some additional hypotheses; also, in perturbative situations the Navier-Stokes equation evolves almost linearly, while in the non-perturbative setting it behaves very nonlinearly, so there is basically no chance of a reduction of the non-perturbative case to the perturbative one unless one comes up with a highly nonlinear transform to achieve this (e.g. a naive scaling argument cannot possibly work). Thus, one is left with only three possible strategies if one wants to solve the full problem:

- Solve the Navier-Stokes equation exactly and explicitly (or at least transform this equation exactly and explicitly to a simpler equation);
- Discover a new globally controlled quantity which is both coercive and either critical or subcritical; or
- Discover a new method which yields global smooth solutions even in the absence of the ingredients (I), (II), and (III) above.

For the rest of this article I refer to these strategies as "Strategy 1", "Strategy 2", and "Strategy 3" respectively.

Much effort has been expended here, especially on Strategy 3, but the supercriticality of the equation presents a truly significant obstacle which already defeats all known methods. Strategy 1 is probably hopeless; the last century of experience has shown that (with the very notable exception of completely integrable systems, of which the Navier-Stokes equations is *not* an example) most nonlinear PDE, even those arising from physics, do not enjoy explicit formulae for solutions from *arbitrary* data (although it may well be the case that there are interesting exact solutions from special (e.g. symmetric) data). Strategy 2 may have a little more hope; after all, the Poincaré conjecture became solvable (though still very far from trivial) after Perelman[Pe2002] introduced a new globally controlled quantity for Ricci flow (the *Perelman entropy*) which turned out to be both coercive and critical. (See also my exposition of this topic at [Ta2006c].) But we are still not very good at discovering new globally controlled quantities; to quote Klainerman[Kl2000], "the discovery of any new bound, stronger than that provided by the energy, for general solutions of *any* of our basic physical equations would have the significance of a major event" (emphasis mine).

I will return to Strategy 2 later, but let us now discuss Strategy 3. The first basic observation is that the Navier-Stokes equation, like many other of our basic model equations, obeys a *scale invariance*: specifically, given any scaling parameter $\lambda > 0$, and any smooth velocity field $u : [0,T) \times \mathbf{R}^3 \to \mathbf{R}^3$ solving the Navier-Stokes equations for some time T, one can form a new velocity field $u^{(\lambda)} : [0,\lambda^2 T) \times \mathbf{R}^3 \to \mathbf{R}^3$ to the Navier-Stokes equation up to time $\lambda^2 T$, by the formula

$$u^{(\lambda)}(t,x) := \frac{1}{\lambda} u(\frac{t}{\lambda^2}, \frac{x}{\lambda})$$

(Strictly speaking, this scaling invariance is only present as stated in the absence of an external force, and with the non-periodic domain $\mathbf{R}^3$ rather than the periodic domain

$\mathbf{T}^3$. One can adapt the arguments here to these other settings with some minor effort, the key point being that an approximate scale invariance can play the role of a perfect scale invariance in the considerations below. The pressure field $p(t,x)$ gets rescaled too, to $p^{(\lambda)}(t,x) := \frac{1}{\lambda^2} p(\frac{t}{\lambda^2}, \frac{x}{\lambda})$, but we will not need to study the pressure here. The viscosity ν remains unchanged.)

We shall think of the rescaling parameter λ as being large (e.g. $\lambda > 1$). One should then think of the transformation from u to $u^{(\lambda)}$ as a kind of "magnifying glass", taking fine-scale behaviour of u and matching it with an identical (but rescaled, and slowed down) coarse-scale behaviour of $u^{(\lambda)}$. The point of this magnifying glass is that it allows us to treat both fine-scale and coarse-scale behaviour on an equal footing, by identifying both types of behaviour with something that goes on at a fixed scale (e.g. the unit scale). Observe that the scaling suggests that fine-scale behaviour should play out on much smaller time scales than coarse-scale behaviour (T versus $\lambda^2 T$). Thus, for instance, if a unit-scale solution does something funny at time 1, then the rescaled fine-scale solution will exhibit something similarly funny at spatial scales $1/\lambda$ and at time $1/\lambda^2$. Blowup can occur when the solution shifts its energy into increasingly finer and finer scales, thus evolving more and more rapidly and eventually reaching a singularity in which the scale in both space and time on which the bulk of the evolution is occuring has shrunk to zero. In order to prevent blowup, therefore, we must arrest this motion of energy from coarse scales (or low frequencies) to fine scales (or high frequencies). (There are many ways in which to make these statements rigorous, for instance using Littlewood-Paley theory, which we will not discuss here, preferring instead to leave terms such as "coarse-scale" and "fine-scale" undefined.)

Now, let us take an arbitrary large-data smooth solution to Navier-Stokes, and let it evolve over a very long period of time $[0,T)$, assuming that it stays smooth except possibly at time T. At very late times of the evolution, such as those near to the final time T, there is no reason to expect the solution to resemble the initial data any more (except in perturbative regimes, but these are not available in the arbitrary large-data case). Indeed, the only control we are likely to have on the late-time stages of the solution are those provided by globally controlled quantities of the evolution. Barring a breakthrough in Strategy 2, we only have two really useful globally controlled (i.e. bounded even for very large T) quantities:

- The *maximum kinetic energy* $\sup_{0 \leq t < T} \frac{1}{2} \int_{\mathbf{R}^3} |u(t,x)|^2 \, dx$; and
- The *cumulative energy dissipation* $\frac{1}{2} \int_0^T \int_{\mathbf{R}^3} |\nabla u(t,x)|^2 \, dxdt$.

Indeed, the energy conservation law implies that these quantities are both bounded by the initial kinetic energy E, which could be large (we are assuming our data could be large) but is at least finite by hypothesis.

The above two quantities are *coercive*, in the sense that control of these quantities imply that the solution, even at very late times, stays in a bounded region of some function space. However, this is basically the only thing we know about the solution at late times (other than that it is smooth until time T, but this is a qualitative assumption and gives no bounds). So, unless there is a breakthrough in Strategy 2, we cannot rule out the worst-case scenario that the solution near time T is essentially an *arbitrary*

smooth divergence-free vector field which is bounded both in kinetic energy and in cumulative energy dissipation by E. In particular, near time T the solution could be concentrating the bulk of its energy into fine-scale behaviour, say at some spatial scale $1/\lambda$. (Of course, cumulative energy dissipation is not a function of a single time, but is an integral over all time; let me suppress this fact for the sake of the current discussion.)

Now, let us take our magnifying glass and blow up this fine-scale behaviour by λ to create a coarse-scale solution to Navier-Stokes. Given that the fine-scale solution could (in the worst-case scenario) be as bad as an arbitrary smooth vector field with kinetic energy and cumulative energy dissipation at most E, the rescaled unit-scale solution can be as bad as an arbitrary smooth vector field with kinetic energy and cumulative energy dissipation at most $E\lambda$, as a simple change-of-variables shows. Note that the control given by our two key quantities has worsened by a factor of λ; because of this worsening, we say that these quantities are *supercritical* - they become increasingly useless for controlling the solution as one moves to finer and finer scales. This should be contrasted with *critical* quantities (such as the energy for *two-dimensional* Navier-Stokes), which are invariant under scaling and thus control all scales equally well (or equally poorly), and *subcritical* quantities, control of which becomes increasingly powerful at fine scales (and increasingly useless at very coarse scales).

Now, suppose we know of examples of unit-scale solutions whose kinetic energy and cumulative energy dissipation are as large as $E\lambda$, but which can shift their energy to the next finer scale, e.g. a half-unit scale, in a bounded amount $O(1)$ of time. Given the previous discussion, we cannot rule out the possibility that our rescaled solution behaves like this example. Undoing the scaling, this means that we cannot rule out the possibility that the original solution will shift its energy from spatial scale $1/\lambda$ to spatial scale $1/2\lambda$ in time $O(1/\lambda^2)$. If this bad scenario repeats over and over again, then convergence of geometric series shows that the solution may in fact blow up in finite time. Note that the bad scenarios do not have to happen immediately after each other (the *self-similar* blowup scenario); the solution could shift from scale $1/\lambda$ to $1/2\lambda$, wait for a little bit (in rescaled time) to "mix up" the system and return to an "arbitrary" (and thus potentially "worst-case") state, and then shift to $1/4\lambda$, and so forth. While the cumulative energy dissipation bound can provide a little bit of a bound on how long the system can "wait" in such a "holding pattern", it is far too weak to prevent blowup in finite time. To put it another way, we have no rigorous, deterministic way of preventing "Maxwell's demon" from plaguing the solution at increasingly frequent (in absolute time) intervals, invoking various rescalings of the above scenario to nudge the energy of the solution into increasingly finer scales, until blowup is attained.

Thus, in order for Strategy 3 to be successful, we basically need to rule out the scenario in which unit-scale solutions with *arbitrarily large* kinetic energy and cumulative energy dissipation shift their energy to the next highest scale. But every single analytic technique we are aware of (except for those involving *exact* solutions, i.e. Strategy 1) requires at least one bound on the size of solution in order to have any chance at all. Basically, one needs at least one bound in order to control all nonlinear errors - and any strategy we know of which does not proceed via exact solutions will have at least one nonlinear error that needs to be controlled. The only thing we have here is a bound on the *scale* of the solution, which is not a bound in the sense that a norm of the solution is bounded; and so we are stuck.

To summarise, any argument which claims to yield global regularity for Navier-Stokes via Strategy 3 must inevitably (via the scale invariance) provide a radically new method for providing non-trivial control of nonlinear unit-scale solutions of arbitrary large size for unit time, which looks impossible without new breakthroughs on Strategy 1 or Strategy 2. (There are a couple of loopholes that one might try to exploit: one can instead try to refine the control on the "waiting time" or "amount of mixing" between each shift to the next finer scale, or try to exploit the fact that each such shift requires a certain amount of energy dissipation, but one can use similar scaling arguments to the preceding to show that these types of loopholes cannot be exploited without a new bound along the lines of Strategy 2, or some sort of argument which works for arbitrarily large data at unit scales.)

To rephrase in an even more jargon-heavy manner: the "energy surface" on which the dynamics is known to live in, can be quotiented by the scale invariance. After this quotienting, the solution can stray arbitrarily far from the origin even at unit scales, and so we lose all control of the solution unless we have exact control (Strategy 1) or can significantly shrink the energy surface (Strategy 2).

The above was a general critique of Strategy 3. Now I'll turn to some known specific attempts to implement Strategy 3, and discuss where the difficulty lies with these:

1. *Using weaker or approximate notions of solution* (e.g. viscosity solutions, penalised solutions, super- or sub- solutions, etc.). This type of approach dates all the way back to Leray [Le1933]. It has long been known that by weakening the nonlinear portion of Navier-Stokes (e.g. taming the nonlinearity), or strengthening the linear portion (e.g. introducing hyperdissipation), or by performing a discretisation or regularisation of spatial scales, or by relaxing the notion of a "solution", one can get global solutions to approximate Navier-Stokes equations. The hope is then to take limits and recover a smooth solution, as opposed to a mere global *weak* solution, which was already constructed by Leray for Navier-Stokes all the way back in 1933. But in order to ensure the limit is smooth, we need convergence in a strong topology. In fact, the same type of scaling arguments used before basically require that we obtain convergence in either a critical or subcritical topology. Absent a breakthrough in Strategy 2, the only type of convergences we have are in very rough - in particular, in supercritical - topologies. Attempting to upgrade such convergence to critical or subcritical topologies is the qualitative analogue of the quantitative problems discussed earlier, and ultimately faces the same problem (albeit in very different language) of trying to control unit-scale solutions of arbitrarily large size. Working in a purely qualitative setting (using limits, etc.) instead of a quantitative one (using estimates, etc.) can disguise these problems (and, unfortunately, can lead to errors if limits are manipulated carelessly), but the qualitative formalism does not magically make these problems disappear. Note that weak solutions are already known to be badly behaved for the closely related Euler equation [Sc1993]. More generally, by recasting the problem in a sufficiently abstract formalism (e.g. formal limits of near-solutions), there are a number of ways to create an abstract object which could be considered as a kind of generalised solution, but the mo-

ment one tries to establish actual control on the regularity of this generalised solution one will encounter all the supercriticality difficulties mentioned earlier.

2. *Iterative methods* (e.g. contraction mapping principle, Nash-Moser iteration, power series, etc.) *in a function space*. These methods are perturbative, and require *something* to be small: either the data has to be small, the nonlinearity has to be small, or the time of existence desired has to be small. These methods are excellent for constructing *local* solutions for large data, or global solutions for *small* data, but cannot handle global solutions for large data (running into the same problems as any other Strategy 3 approach). These approaches are also typically rather insensitive to the specific structure of the equation, which is already a major warning sign since one can easily construct (rather artificial) systems similar to Navier-Stokes for which blowup is known to occur. The optimal perturbative result is probably very close to that established by Koch-Tataru[KoTa2001], for reasons discussed in that paper.

3. *Exploiting blowup criteria*. Perturbative theory can yield some highly non-trivial blowup criteria - that certain norms of the solution must diverge if the solution is to blow up. For instance, a celebrated result of Beale-Kato-Majda[BeKaMa1984] shows that the maximal vorticity must have a divergent time integral at the blowup point. However, all such blowup criteria are subcritical or critical in nature, and thus, barring a breakthrough in Strategy 2, the known globally controlled quantities cannot be used to reach a contradiction. Scaling arguments similar to those given above show that perturbative methods cannot achieve a supercritical blowup criterion.

4. *Asymptotic analysis of the blowup point(s)*. Another proposal is to rescale the solution near a blowup point and take some sort of limit, and then continue the analysis until a contradiction ensues. This type of approach is useful in many other contexts (for instance, in understanding Ricci flow). However, in order to actually extract a useful limit (in particular, one which still solves Navier-Stokes in a strong sense, and does collapse to the trivial solution), one needs to uniformly control all rescalings of the solution - or in other words, one needs a breakthrough in Strategy 2. Another major difficulty with this approach is that blowup can occur not just at one point, but can conceivably blow up on a one-dimensional set[Sc1976]; this is another manifestation of supercriticality.

5. *Analysis of a minimal blowup solution*. This is a strategy, initiated by Bourgain [Bo1999b], which has recently been very successful (see [KeMe2006], [CoKeStTaTa2008], [RyVi2007], [Vi2007], [TaViZh2008], [KiTaVi2008]) in establishing large data global regularity for a variety of equations with a critical conserved quantity, namely to assume for contradiction that a blowup solution exists, and then extract a *minimal* blowup solution which minimises the conserved quantity. This strategy (which basically pushes the perturbative theory to its natural limit) seems set to become the standard method for dealing with large data critical equations. It has the appealing feature that there is enough compactness (or almost periodicity) in the minimal blowup solution (once one quotients

out by the scaling symmetry) that one can begin to use subcritical and supercritical conservation laws and monotonicity formulae as well (see my survey on this topic [Ta2006f]). Unfortunately, as the strategy is currently understood, it does not seem to be directly applicable to a supercritical situation (unless one simply assumes that some critical norm is globally bounded) because it is impossible, in view of the scale invariance, to minimise a non-scale-invariant quantity.

6. *Abstract approaches* (avoiding the use of properties specific to the Navier-Stokes equation). At its best, abstraction can efficiently organise and capture the key difficulties of a problem, placing the problem in a framework which allows for a direct and natural resolution of these difficulties without being distracted by irrelevant concrete details. (Kato's semigroup method[Ka1993] is a good example of this in nonlinear PDE; regrettably for this discussion, it is limited to subcritical situations.) At its worst, abstraction conceals the difficulty within some subtle notation or concept (e.g. in various types of convergence to a limit), thus incurring the risk that the difficulty is "magically" avoided by an inconspicuous error in the abstract manipulations. An abstract approach which manages to breezily ignore the supercritical nature of the problem thus looks very suspicious. More substantively, there are many equations which enjoy a coercive conservation law yet still can exhibit finite time blowup (e.g. the mass-critical focusing NLS equation); an abstract approach thus would have to exploit some subtle feature of Navier-Stokes which is not present in all the examples in which blowup is known to be possible. Such a feature is unlikely to be discovered abstractly before it is first discovered concretely; the field of PDE has proven to be the type of mathematics where progress generally starts in the concrete and then flows to the abstract, rather than vice versa.

If we abandon Strategy 1 and Strategy 3, we are thus left with Strategy 2 - discovering new bounds, stronger than those provided by the (supercritical) energy. This is not *a priori* impossible, but there is a huge gap between simply wishing for a new bound and actually discovering and then rigorously establishing one. Simply sticking the existing energy bounds into the Navier-Stokes equation and seeing what comes out will provide a few more bounds, but they will all be supercritical, as a scaling argument quickly reveals. The only other way we know of to create global non-perturbative deterministic bounds is to discover a new conserved or monotone quantity. In the past, when such quantities have been discovered, they have always been connected either to geometry (symplectic, Riemmanian, complex, etc.), to physics, or to some consistently favourable (defocusing) sign in the nonlinearity (or in various "curvatures" in the system). There appears to be very little usable geometry in the equation; on the one hand, the Euclidean structure enters the equation via the diffusive term Δ and by the divergence-free nature of the vector field, but the nonlinearity is instead describing transport by the velocity vector field, which is basically just an arbitrary volume-preserving infinitesimal diffeomorphism (and in particular does not respect the Euclidean structure at all). One can try to quotient out by this diffeomorphism (i.e. work in material coordinates) but there are very few geometric invariants left to play with when one does so. (In the case of the Euler equations, the vorticity vector field is preserved

modulo this diffeomorphism, as observed for instance in [Li2003], but this invariant is very far from coercive, being almost purely topological in nature.) The Navier-Stokes equation, being a system rather than a scalar equation, also appears to have almost no favourable sign properties, in particular ruling out the type of bounds which the maximum principle or similar comparison principles can give. This leaves physics, but apart from the energy, it is not clear if there are any physical quantities of fluids which are *deterministically* monotone. (Things look better on the stochastic level, in which the laws of thermodynamics might play a role, but the Navier-Stokes problem, as defined by the Clay institute, is deterministic, and so we have Maxwell's demon to contend with.) It would of course be fantastic to obtain a fourth source of non-perturbative controlled quantities, not arising from geometry, physics, or favourable signs, but this looks somewhat of a long shot at present. Indeed given the turbulent, unstable, and chaotic nature of Navier-Stokes, it is quite conceivable that in fact no reasonable globally controlled quantities exist beyond that which arise from the energy.

Of course, given how hard it is to show global regularity, one might try instead to establish finite time blowup instead (this also is acceptable for the Millennium prize[Fe2006]). Unfortunately, even though the Navier-Stokes equation is known to be very unstable, it is not clear at all how to pass from this to a rigorous demonstration of a blowup solution. All the rigorous finite time blowup results (as opposed to mere instability results) that I am aware of rely on one or more of the following ingredients:

(a) Exact blowup solutions (or at least an exact transformation to a significantly simpler PDE or ODE, for which blowup can be established);

(b) An ansatz for a blowup solution (or approximate solution), combined with some nonlinear stability theory for that ansatz;

(c) A comparison principle argument, dominating the solution by another object which blows up in finite time, taking the solution with it; or

(d) An indirect argument, constructing a functional of the solution which must attain an impossible value in finite time (e.g. a quantity which is manifestly non-negative for smooth solutions, but must become negative in finite time).

It may well be that there is some exotic symmetry reduction which gives (a), but no-one has located any good exactly solvable special case of Navier-Stokes (in fact, those which have been found, are known to have global smooth solutions). Method (b) is problematic for two reasons: firstly, we do not have a good ansatz for a blowup solution, but perhaps more importantly it seems hopeless to establish a stability theory for any such ansatz thus created, as this problem is essentially a more difficult version of the global regularity problem, and in particular subject to the main difficulty, namely controlling the highly nonlinear behaviour at fine scales. (One of the ironies in pursuing method (b) is that in order to establish rigorous *blowup* in some sense, one must first establish rigorous *stability* in some other (renormalised) sense.) Method (c) would require a comparison principle, which as noted before appears to be absent for the non-scalar Navier-Stokes equations. Method (d) suffers from the same problem, ultimately coming back to the "Strategy 2" problem that we have virtually no globally monotone

quantities in this system to play with (other than energy monotonicity, which clearly looks insufficient by itself). Obtaining a new type of mechanism to force blowup other than (a)-(d) above would be quite revolutionary, not just for Navier-Stokes; but I am unaware of even any proposals in these directions, though perhaps topological methods might have some effectiveness.

So, after all this negativity, do I have any positive suggestions for how to solve this problem? My opinion is that Strategy 1 is impossible, and Strategy 2 would require either some exceptionally good intuition from physics, or else an incredible stroke of luck. Which leaves Strategy 3 (and indeed, I think one of the main reasons why the Navier-Stokes problem is interesting is that it *forces* us to create a Strategy 3 technique). Given how difficult this strategy seems to be, as discussed above, I only have some extremely tentative and speculative thoughts in this direction, all of which I would classify as "blue-sky" long shots:

1. *Work with ensembles of data, rather than a single initial datum.* All of our current theory for deterministic evolution equations deals only with a single solution from a single initial datum. It may be more effective to work with parameterised familes of data and solutions, or perhaps probability measures (e.g. Gibbs measures or other invariant measures). One obvious partial result to shoot for is to try to establish global regularity for *generic* large data rather than *all* large data; in other words, acknowledge that Maxwell's demon might exist, but show that the probability of it actually intervening is very small. The problem is that we have virtually no tools for dealing with generic (average-case) data other than by treating all (worst-case) data; the enemy is that the Navier-Stokes flow itself might have some perverse entropy-reducing property which somehow makes the average case drift towards (or at least recur near) the worst case over long periods of time. This is incredibly unlikely to be the truth, but we have no tools to prevent it from happening at present.

2. *Work with a much simpler (but still supercritical) toy model.* The Navier-Stokes model is parabolic, which is nice, but is complicated in many other ways, being relatively high-dimensional and also non-scalar in nature. It may make sense to work with other, simplified models which still contain the key difficulty that the only globally controlled quantities are supercritical. Examples include the Katz-Pavlovíc dyadic model[KaPa2005] for the Euler equations (for which blowup can be demonstrated by a monotonicity argument; see [FrPa2008]), or the spherically symmetric defocusing supercritical nonlinear wave equation $-u_{tt} + \Delta u = u^7$ in three spatial dimensions.

3. *Develop non-perturbative tools to control deterministic non-integrable dynamical systems.* Throughout this post we have been discussing PDE, but actually there are similar issues arising in the nominally simpler context of finite-dimensional dynamical systems (ODE). Except in perturbative contexts (such as the neighbourhood of a fixed point or invariant torus), the long-time evolution of a dynamical system for deterministic data is still largely only controllable by the classical tools of exact solutions, conservation laws and monotonicity formulae; a discovery of a new and effective tool for this purpose would be a major

breakthrough. One natural place to start is to better understand the long-time, non-perturbative dynamics of the classical three-body problem, for which there are still fundamental unsolved questions.

4. *Establish really good bounds for critical or nearly-critical problems*. Recently, I showed [Ta2007] that having a very good bound for a critical equation essentially implies that one also has a global regularity result for a slightly supercritical equation. The idea is to use a monotonicity formula which does weaken very slightly as one passes to finer and finer scales, but such that each such passage to a finer scale costs a significant amount of monotonicity; since there is only a bounded amount of monotonicity to go around, it turns out that the latter effect just barely manages to overcome the former in my equation to recover global regularity (though by doing so, the bounds worsen from polynomial in the critical case to double exponential in my logarithmically supercritical case). I severely doubt that my method can push to non-logarithmically supercritical equations, but it does illustrate that having very strong bounds at the critical level may lead to some modest progress on the problem.

5. *Try a topological method*. This is a special case of (1). It may well be that a primarily topological argument may be used either to construct solutions, or to establish blowup; there are some precedents for this type of construction in elliptic theory. Such methods are very global by nature, and thus not restricted to perturbative or nearly-linear regimes. However, there is no obvious topology here (except possibly for that generated by the vortex filaments) and as far as I know, there is not even a "proof-of-concept" version of this idea for any evolution equation. So this is really more of a wish than any sort of concrete strategy.

6. *Understand pseudorandomness*. This is an incredibly vague statement; but part of the difficulty with this problem, which also exists in one form or another in many other famous problems (e.g. Riemann hypothesis, $P = BPP$, $P \neq NP$, twin prime and Goldbach conjectures, normality of digits of π, Collatz conjecture, etc.) is that we expect any sufficiently complex (but deterministic) dynamical system to behave "chaotically" or "pseudorandomly", but we still have very few tools for actually making this intuition precise, especially if one is considering deterministic initial data rather than generic data. Understanding pseudorandomness in other contexts, even dramatically different ones, may indirectly shed some insight on the turbulent behaviour of Navier-Stokes.

In conclusion, while it is good to occasionally have a crack at impossible problems, just to try one's luck, I would personally spend much more of my time on other, more tractable PDE problems than the Clay prize problem, though one should certainly keep that problem in mind if, in the course on working on other problems, one indeed does stumble upon something that smells like a breakthrough in Strategy 1, 2, or 3 above. (In particular, there are many other serious and interesting questions in fluid equations that are not anywhere near as difficult as global regularity for Navier-Stokes, but still highly worthwhile to resolve.)

1.4.1 Notes

This article was originally posted on Mar 18, 2007 at

terrytao.wordpress.com/2007/03/18

Nets Katz points out that a significantly simpler (but still slightly supercritical) problem would be to improve the double-exponential bound of Beale-Kato-Majda [BeKaMa1984] for the growth of vorticity for periodic solutions to the Euler equations in two dimensions.

Sarada Rajeev points out an old observation of Arnold that the Euler equations are in fact the geodesic flow in the group of volume-preserving diffeomorphisms (using the Euclidean L^2 norm of the velocity field to determine the Riemannian metric structure); such structure may well be decisive in improving our understanding of the Euler equation, and thus (indirectly) for Navier-Stokes as well.

Stephen Montgomery-Smith points out that any new conserved or monotone quantities (of the type needed to make a "Strategy 2" approach work) might distort the famous Kolmogorov $5/3$ power law for the energy spectrum. Since this law has been confirmed by many numerical experiments, this could be construed as evidence against a Strategy 2 approach working. On the other hand, Montgomery-Smith also pointed out that for two-dimensional Navier-Stokes, one has L^p bounds on vorticity which do not affect the Kraichnan 3 power law coming from the enstrophy.

After the initial posting of this article, I managed to show[Ta2008b] that the periodic global regularity problem for Navier-Stokes was equivalent to the task of obtaining a local or global H^1 bound on classical solutions, thus showing that the regularity problem is in some sense "equivalent" to that of making Strategy 2 work.

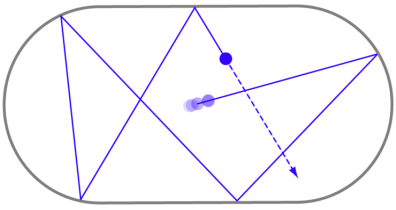


Figure 1.1: The Bunimovich stadium. (Figure from wikipedia.)

1.5 Scarring for the Bunimovich stadium

The problem of scarring for the Bunimovich stadium is well known in the area of *quantum chaos* or *quantum ergodicity* (see e.g. [BuZw2004]); I am attracted to it both for its simplicity of statement, and also because it focuses on one of the key weaknesses in our current understanding of the Laplacian, namely is that it is difficult with the tools we know to distinguish between *eigenfunctions* (exact solutions to $-\Delta u_k = \lambda_k u_k$) and *quasimodes* (approximate solutions to the same equation), unless one is willing to work with generic energy levels rather than specific energy levels.

The Bunimovich stadium Ω is the name given to any planar domain consisting of a rectangle bounded at both ends by semicircles. Thus the stadium has two flat edges (which are traditionally drawn horizontally) and two round edges: see Figure 1.1.

Despite the simple nature of this domain, the stadium enjoys some interesting classical and quantum dynamics. It was shown by Bunimovich[Bu1974] that the classical billiard ball dynamics on this stadium is *ergodic*, which means that a billiard ball with randomly chosen initial position and velocity (as depicted above) will, over time, be uniformly distributed across the billiard (as well as in the energy surface of the phase space of the billiard). On the other hand, the dynamics is not *uniquely ergodic* because there do exist some exceptional choices of initial position and velocity for which one does not have uniform distribution, namely the vertical trajectories in which the billiard reflects orthogonally off of the two flat edges indefinitely.

Rather than working with (classical) individual trajectories, one can also work with (classical) *invariant ensembles* - probability distributions in phase space which are invariant under the billiard dynamics. Ergodicity then says that (at a fixed energy) there are no invariant absolutely continuous ensemble other than the obvious one, namely the probability distribution with uniformly distributed position and velocity direction. On the other hand, unique ergodicity would say the same thing but dropping the "absolutely continuous" - but each vertical bouncing ball mode creates a singular invariant ensemble along that mode, so the stadium is not uniquely ergodic.

Now from physical considerations we expect the quantum dynamics of a system to have similar qualitative properties as the classical dynamics; this can be made precise in many cases by the mathematical theories of *semi-classical analysis* and *microlocal analysis*. The quantum analogue of the dynamics of classical ensembles is the dynam-

ics of the *Schrödinger equation*

$$i\hbar\partial_t\psi + \frac{\hbar^2}{2m}\Delta\psi = 0,$$

where we impose Dirichlet boundary conditions $\psi|_{\partial\Omega} = 0$ (one can also impose Neumann conditions if desired, the problems seem to be roughly the same). The quantum analogue of an invariant ensemble is that of a single *eigenfunction* (i.e. a solution u_k to the equation $-\Delta u_k = \lambda_k u_k$), which we normalise in the usual L^2 manner, so that $\int_\Omega |u_k|^2 = 1$. (Due to the compactness of the domain Ω, the set of eigenvalues λ_k of the Laplacian $-\Delta$ is discrete and goes to infinity, though there is some multiplicity arising from the symmetries of the stadium. These eigenvalues are the same eigenvalues that show up in the famous "can you hear the shape of a drum?" problem [Ka1966].) Roughly speaking, quantum ergodicity is then the statement that *almost all* eigenfunctions are uniformly distributed in physical space (as well as in the energy surface of phase space), whereas quantum unique ergodicity (QUE) is the statement that *all* eigenfunctions are uniformly distributed. A little more precisely:

1. If quantum ergodicity holds, then for any open subset $A \subset \Omega$ we have $\int_A |u_k|^2 \to |A|/|\Omega|$ as $\lambda_k \to \infty$, provided we exclude a set of exceptional k of density zero.

2. If quantum unique ergodicity holds, then we have the same statement as before, except that we do not need to exclude the exceptional set.

In fact, quantum ergodicity and quantum unique ergodicity say somewhat stronger things than the above two statements, but I would need tools such as pseudodifferential operators to describe these more technical statements, and so I will not do so here.

Now it turns out that for the stadium, quantum ergodicity is known to be true; this specific result was first obtained by Gérard and Leichtman[GeLi1993], although "classical ergodicity implies quantum ergodicity" results of this type go back to Schnirelman[Sn1974] (see also [Ze1990], [CdV1985]). These results are established by microlocal analysis methods, which basically proceed by aggregating all the eigenfunctions together into a single object (e.g. a heat kernel, or some other function of the Laplacian) and then analysing the resulting aggregate semiclassically. It is because of this aggregation that one only gets to control *almost all* eigenfunctions, rather than *all* eigenfunctions.

In analogy to the above theory, one generally expects classical unique ergodicity should correspond to QUE. For instance, there is the famous (and very difficult) *quantum unique ergodicity conjecture* of Rudnick and Sarnak[RuSa1994], which asserts that QUE holds for all compact manifolds without boundary with negative sectional curvature. This conjecture will not be discussed here (it would warrant an entire article in itself, and I would not be the best placed to write it). Instead, we focus on the Bunimovich stadium. The stadium is clearly not classically uniquely ergodic due to the vertical bouncing ball modes, and so one would conjecture that it is not QUE either. In fact one conjectures the slightly stronger statement:

Conjecture 1.1 (Scarring conjecture)**.** *There exists a subset $A \subset \Omega$ and a sequence u_{k_j} of eigenfunctions with $\lambda_{k_j} \to \infty$, such that $\int_A |u_{k_j}|^2$ does* not *converge to $|A|/|\Omega|$. Infor-*

mally, the eigenfunctions either concentrate (or "scar") in A, or on the complement of A.

Indeed, one expects to take A to be a union of vertical bouncing ball trajectories (indeed, from Egorov's theorem (in microlocal analysis, not the one in real analysis), this is almost the only choice). This type of failure of QUE even in the presence of quantum ergodicity has already been observed for some simpler systems, such as the Arnold cat map [FNdB2003]. Some further discussion of this conjecture can be found at [BuZw2005].

One reason this conjecture appeals to me is that there is a very plausible physical argument, due to Heller[He1991] and refined by Zelditch[Ze2004], which indicates the conjecture is almost certainly true. Roughly speaking, it runs as follows. Using the rectangular part of the stadium, it is easy to construct (high-energy) *quasimodes of order* 0 which scar (i.e. they concentrate on a proper subset A of Ω); roughly speaking, these quasimodes are solutions u to an approximate eigenfunction equation $-\Delta u = (\lambda + O(1))u$ for some λ. For instance, if the two horizontal edges of the stadium lie on the lines $y = 0$ and $y = 1$, then one can take $u(x,y) := \varphi(x)\sin(\pi n y)$ and $\lambda := \pi^2 n^2$ for some large integer n and some suitable bump function φ. Using the spectral theorem, one expects u to concentrate its energy in the band $[\pi^2 n^2 - O(1), \pi^2 n^2 + O(1)]$. On the other hand, in two dimensions the Weyl law for distribution of eigenvalues asserts that the eigenvalues have an average spacing comparable to 1. If (and this is the non-rigorous part) this average spacing also holds on a typical band $[\pi^2 n^2 - O(1), \pi^2 n^2 + O(1)]$, this shows that the above quasimode is essentially generated by only O(1) eigenfunctions. Thus, by the pigeonhole principle (or more precisely, Pythagoras' theorem), at least one of the eigenfunctions must exhibit scarring. (Actually, to make this argument fully rigorous, one needs to examine the distribution of the quasimode in momentum space as well as in physical space, and to use the full phase space definition of quantum unique ergodicity, which I have not detailed here. I thank Greg Kuperberg for pointing out this subtlety.)

The big gap in this argument is that nobody knows how to take the Weyl law (which is proven by the microlocal analysis approach, i.e. aggregate all the eigenstates together and study the combined object) and localise it to such an extremely sparse set of narrow energy bands. Using the standard error term in Weyl's law one can localise to bands of width $O(n)$ around, say, $\pi^2 n^2$, and by using the ergodicity one can squeeze this down to $o(n)$, but to even get control on a band of with width $O(n^{1-\varepsilon})$ would require a heroic effort (analogous to establishing a zero-free region $\{s : \mathrm{Re}(s) > 1 - \varepsilon\}$ for the Riemann zeta function). The enemy is somehow that around each energy level $\pi^2 n^2$, a lot of exotic eigenfunctions spontaneously appear, which manage to dissipate away the bouncing ball quasimodes into a sea of quantum chaos. This is exceedingly unlikely to happen, but we do not seem to have tools available to rule it out.

One indication that the problem is not going to be entirely trivial is that one can show (basically by unique continuation or control theory arguments) that no pure eigenfunction can be solely concentrated within the rectangular portion of the stadium (where all the vertical bouncing ball modes are); a significant portion of the energy must leak out into the two "wings" (or at least into arbitrarily small neighbourhoods of these wings). This was established by Burq and Zworski [BuZw2005].

On the other hand, the stadium is a very simple object - it is one of the simplest and most symmetric domains for which we cannot actually compute eigenfunctions or eigenvalues explicitly. It is tempting to just discard all the microlocal analysis and just try to construct eigenfunctions by brute force. But this has proven to be surprisingly difficult; indeed, despite decades of sustained study into the eigenfunctions of Laplacians (given their many applications to PDE, to number theory, to geometry, etc.) we still do not know very much about the shape and size of any *specific* eigenfunction for a general manifold, although we know plenty about the average-case behaviour (via microlocal analysis) and also know the worst-case behaviour (by Sobolev embedding or restriction theorem type tools). This conjecture is one of the simplest conjectures which would force us to develop a new tool for understanding eigenfunctions, which could then conceivably have a major impact on many areas of analysis.

One might consider modifying the stadium in order to make scarring easier to show, for instance by selecting the dimensions of the stadium appropriately (e.g. obeying a Diophantine condition), or adding a potential or magnetic term to the equation, or perhaps even changing the metric or topology. To have even a single rigorous example of a reasonable geometric operator for which scarring occurs despite the presence of quantum ergodicity would be quite remarkable, as any such result would have to involve a method that can deal with a very rare set of special eigenfunctions in a manner quite different from the generic eigenfunction.

Actually, it is already interesting to see if one can find better quasimodes than the ones listed above which exhibit scarring, i.e. to improve the $O(1)$ error in the spectral bandwidth; this specific problem has been proposed in [BuZw2004] as a possible toy version of the main problem.

1.5.1 Notes

This article was originally posted on Mar 28, 2007 at

`terrytao.wordpress.com/2007/03/28`

Greg Kuperberg and I discussed whether one could hope to obtain this conjecture by deforming continuously from the rectangle (for which the eigenfunctions are explicitly known) to the stadium. Unfortunately, since eigenvalues generically do not intersect each other under continuous deformations, the ordering of the eigenvalues does not change, and so by Weyl's law one does not expect the scarred states of the stadium to correspond to any particularly interesting states of the rectangle.

Many pictures of stadium eigenfunctions can be found online, for instance at Douglas Stone's page

`www.eng.yale.edu/stonegroup/`

or Arnd Bäcker's page

`www.physik.tu-dresden.de/\~{}baecker/`

A small number of these eigenfunctions seem to exhibit scarring, thus providing some numerical support for the above conjectures, though of course these conjectures concern the asymptotic regime in which the eigenvalue goes to infinity, and so cannot be proved or disproved solely through numerics.

1.6 Triangle and diamond densities in large dense graphs

The question in extremal graph theory I wish to discuss here originates from Luca Trevisan; it shows that we still don't know everything that we should about the "local" properties of large dense graphs.

Let $G = (V,E)$ be a large (undirected) graph, thus V is the vertex set with some large number n of vertices, and E is the collection of edges $\{x,y\}$ connecting two vertices in the graph. We can allow the graph to have loops $\{x,x\}$ if one wishes; it's not terribly important for this question (since the number of loops is so small compared to the total number of edges), so let's say there are no loops. We define three quantities of the graph G:

- The *edge density* $0 \leq \alpha \leq 1$, defined as the number of edges in G, divided by the total number of possible edges, i.e. $n(n-1)/2$;
- The *triangle density* $0 \leq \beta \leq 1$, defined as the number of triangles in G (i.e. unordered triplets $\{x,y,z\}$ such that $\{x,y\}$, $\{y,z\}$, $\{z,x\}$ all lie in G), divided by the total number of possible triangles, namely $n(n-1)(n-2)/6$;
- The *diamond density* $0 \leq \gamma \leq 1$, defined as the number of diamonds in G (i.e. unordered pairs $\{\{x,y,z\},\{x,y,w\}\}$ of triangles in G which share a common edge), divided by the total number of possible diamonds, namely $n(n-1)(n-2)(n-3)/4$.

Up to insignificant errors of $o(1)$ (i.e. anything which goes to zero as the number of vertices goes to infinity), these densities can also be interpreted probabilistically as follows: if x,y,z,w are randomly selected vertices in V, then we have

$$\begin{aligned} \mathbf{P}(\{x,y\} \in E) &= \alpha + o(1); \\ \mathbf{P}(\{x,y\},\{y,z\},\{z,x\} \in E) &= \beta + o(1); \text{ and} \\ \mathbf{P}(\{x,y\},\{y,z\},\{z,x\},\{y,w\},\{w,x\} \in E) &= \gamma + o(1). \end{aligned}$$

(The errors of $o(1)$ arise because the vertices x,y,z,w may occasionally collide with each other, though this probability becomes very small when n is large.) Thus we see that these densities are "local" qualities of the graph, as we only need to statistically sample the graph at a small number of randomly chosen vertices in order to estimate them.

A general question is to determine all the constraints relating α,β,γ in the limit $n \to \infty$. (It is known from the work of Lovász and Szegedy[LoSz2006] that the relationships between local graph densities such as these stabilise in this limit; indeed, given any error tolerance $\varepsilon > 0$ and any large graph G with densities α,β,γ, there exists a graph with "only" $O_\varepsilon(1)$ vertices whose densities α',β',γ' differ from those of G by at most ε, although the best known bounds for $O_\varepsilon(1)$ are far too poor at present to be able to get any useful information on the asymptotic constraint set by direct exhaustion by computer of small graphs.)

Let us forget about diamonds for now and only look at the edge and triangle densities α,β. Then the story is already rather non-trivial. The main concern is to figure out,

for each fixed α, what the best possible upper and lower bounds on β are (up to $o(1)$ errors); since the collection of graphs with a given edge density is "path-connected" in some sense, it is not hard to see that every value of β between the upper and lower bounds is feasible modulo $o(1)$ errors.

The best possible upper bound is easy: $\beta \leq \alpha^{3/2} + o(1)$. This can be established by either the Kruskal-Katona theorem[Kr1963, Ka1968], the Loomis-Whitney inequality[LoWh1949] (or the closely related box theorem[BoTh1995]), or just two applications of Hölder's inequality; we leave this as an exercise. The bound is sharp, as can be seen by looking at a complete subgraph on $(\alpha^{1/2} + o(1))n$ vertices. (We thank Tim Austin and Imre Leader for these observations and references, as well as those in the paragraph below.) There is some literature on refining the $o(1)$ factor; see [Ni2008] for a survey.

The lower bound is trickier. The complete bipartite graph example shows that the trivial lower bound $\beta \geq 0$ is attainable when $\alpha \leq 1/2 - o(1)$, and Turán's theorem[Tu1941] shows that this is sharp. For $\alpha \geq 1/2$, a classical theorem of Goodman[Go1959] (see also [NoSt1963]) shows that $\beta \geq \alpha(2\alpha - 1) - o(1)$. When $\alpha = 1 - 1/k$ for some integer k, this inequality is sharp, as can be seen by looking at the complete k-partite graph.

Goodman's result is thus sharp at infinitely many values of α, but it turns out that it is not quite the best bound. After several partial results, the optimal bound was obtained recently by Razborov[Ra2008b], who established for $1 - 1/k < \alpha < 1 - 1/(k+1)$ that

$$\beta \geq \frac{(k-1)\left(k - 2\sqrt{k(k-\alpha(k+1))}\right)\left(k + \sqrt{k(k-\alpha(k+1))}\right)}{k^2(k+1)^2} - o(1)$$

and that this is sharp (!) except for the $o(1)$ error (see [Fi1989] for some additional work on this error term).

Now we consider the full problem of relating edge densities, triangle densities, and diamond densities. Given that the relationships between α, β were already so complex, a full characterisation of the constraints connecting α, β, γ is probably impossible at this time (though it might be possible to prove that they can be decidable via some (impractical) computer algorithm, and it also looks feasible to determine the exact constraints between just α and γ). The question of Trevisan however focuses on a specific regime in the configuration space, in which β is exceptionally small. From the Cauchy-Schwarz inequality and the observation that a diamond is nothing more than a pair of triangles with a common edge, we obtain the inequality

$$\gamma \geq \frac{\beta^2}{\alpha} - o(1). \tag{1.1}$$

Because we understand very well when equality holds in the Cauchy-Schwartz inequality, we know that (1.1) would only be sharp when the triangles are distributed "evenly" among the edges, so that almost every edge is incident to the roughly expected number of triangles (which is roughly $\beta n/\alpha$). However, it is a remarkable fact that this type of equidistribution is known to be impossible when β is very small. Indeed, the triangle removal lemma of Ruzsa and Szemerédi[RuSz1978] asserts that if β is small, then one

can in fact make β vanish (i.e. delete all triangles) by removing at most $c(\beta)n^2$ edges, where $c(\beta) \to 0$ in the limit $\beta \to 0$. This shows that among all the roughly $\alpha n^2/2$ edges in the graph, at most $c(\beta)n^2$ of them will already be incident to all the triangles in the graph. This, and Cauchy-Schwarz, gives a bound of the form

$$\gamma \geq \frac{\beta^2}{c(\beta)} - o(1), \tag{1.2}$$

which is a better bound than (1.1) when β is small compared with α.

Trevisan's question is: can one replace $c(\beta)$ in (1.2) by any more civilised function of β? To explain what "civilised" means, let me show you the best bound that we know of today. Let $2 \uparrow\uparrow n$ be the *tower-exponential* of height n, defined recursively by

$$2 \uparrow\uparrow 1 := 2; \quad 2 \uparrow\uparrow (n+1) := 2^{2 \uparrow\uparrow n};$$

this is a very rapidly growing function, faster than exponential, double exponential, or any other finite iterated exponential. We invert this function and define the *inverse tower function* $\log_* n$ by

$$\log_* n := \inf\{m : 2 \uparrow\uparrow m \geq n\}.$$

This function goes to infinity as $n \to \infty$, but *very* slowly - slower than $\log n$ or even $\log\log n$ (which, as famously stated by Carl Pomerance, "is proven to go to infinity, but has never been observed to do so").

The best bound on $c(\beta)$ known is of the form

$$c(\beta) \ll (\log_* \frac{1}{\beta})^{-\varepsilon}$$

for some absolute constant $\varepsilon > 0$ (e.g. 1/10 would work here). This bound is so poor because the proof goes via the Szemerédi regularity lemma[Sz1978], which is known by the work of Gowers[Go1997] to necessarily have tower-type dependencies in the constants.

The open question is whether one can obtain a bound of the form (1.2) in which $1/c(\beta)$ is replaced by a quantity which grows better in β, e.g. one which grows logarithmically or double logarithmically rather than inverse-tower-exponential. Such a bound would perhaps lead the way to improving the bounds on the triangle removal lemma; we now have many proofs of this lemma, but they all rely on one form or another of the regularity lemma and so inevitably have the tower-exponential type bounds present. The triangle removal lemma is also connected to many other problems, including property testing for graphs and Szemerédi's theorem on arithmetic progressions[Sz1975] (indeed, the triangle removal lemma implies the length three special case of Szemerédi's theorem, i.e. Roth's theorem[Ro1953]), so progress on improving (1.2) may well lead to much better bounds in many other problems, as well as furnishing another tool beyond the regularity lemma with which to attack these problems. Curiously, the work of Lovász and Szegedy[LoSz2006] implies that the question can be rephrased in a purely analytic fashion, without recourse to graphs. Let

$W : [0,1]^2 \to [0,1]$ be a measurable symmetric function on the unit square, and consider the quantities

$$\beta := \int_0^1 \int_0^1 \int_0^1 W(x,y)W(y,z)W(z,x)\ dxdydz$$

and

$$\gamma := \int_0^1 \int_0^1 \int_0^1 \int_0^1 W(x,y)W(y,z)W(z,x)W(y,w)W(w,x)\ dxdydzdw.$$

Any bound connecting β and γ here is known to imply the same bound for triangle and diamond densities (with an error of $o(1)$), and vice versa. Thus, the question is now to establish the inequality $\gamma \geq \beta^2/c'(\beta)$ for some civilised value of $c;(\beta)$, which at present is only known to decay to zero as $\beta \to 0$ like an inverse tower-exponential function.

1.6.1 Notes

This article was originally posted on Apr 1, 2007 at

terrytao.wordpress.com/2007/04/01

Thanks to Vlado Nikiforov for pointing out some additional references and related questions.

Yuval Peres pointed out some similarity between this problem and a conjecture of Sidorenko[Si1994], which asserts that the number of copies of a bipartite graph (V_H, W_H, E_H) inside a larger graph (V_G, W_G, E_G) should always be at least $|V_G|^{|V_H|}|W_G|^{|W_H|}(|E_G|/|V_G$ which is asymptotically what one expects for a random graph; this conjecture is known for some simple examples of graphs H, such as cycles, paths, or stars, but is open in general.

1.7 What is a quantum honeycomb?

This problem lies in the highly interconnected interface between algebraic combinatorics (esp. the combinatorics of Young tableaux and related objects, including honeycombs and puzzles), algebraic geometry (particularly classical and quantum intersection theory and geometric invariant theory), linear algebra (additive and multiplicative, real and tropical), and the representation theory (classical, quantum, crystal, etc.) of classical groups. (Another open problem in this subject is to find a succinct and descriptive name for the field.) I myself haven't actively worked in this area for several years, but I still find it a fascinating and beautiful subject. (With respect to the dichotomy between structure and randomness (see Section 3.1), this subject lies deep within the "structure" end of the spectrum.)

As mentioned above, the problems in this area can be approached from a variety of quite diverse perspectives, but here I will focus on the linear algebra perspective, which is perhaps the most accessible. About nine years ago, Allen Knutson and I[KnTa1999] introduced a combinatorial gadget, called a *honeycomb*, which among other things controlled the relationship between the eigenvalues of two arbitrary Hermitian matrices A, B, and the eigenvalues of their sum $A+B$; this was not the first such gadget that achieved this purpose, but it was a particularly convenient one for studying this problem, in particular it was used to resolve two conjectures in the subject, the *saturation conjecture* and the *Horn conjecture*. (These conjectures have since been proven by a variety of other methods [KaLeMi2008], [DeWe2000], [Be2006], [KnTaWo2004].) There is a natural multiplicative version of these problems, which now relates the eigenvalues of two arbitrary *unitary* matrices U, V and the eigenvalues of their product UV; this led to the "quantum saturation" and "quantum Horn" conjectures, which were proven a couple years ago[Be2008]. However, the quantum analogue of a "honeycomb" remains a mystery; this is the main topic of the current article.

Let us first briefly review the additive situation. Consider three $n \times n$ Hermitian matrices A, B, C such that $A+B=C$. Being Hermitian, the matrices A,B,C are all diagonalisable with real eigenvalues. Accordingly, let us arrange the eigenvalues of A (with multiplicity) in decreasing order as

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$$

the eigenvalues of B as

$$\mu_1 \geq \mu_2 \geq \ldots \geq \mu_n$$

and the eigenvalues of C as

$$\nu_1 \geq \nu_2 \geq \ldots \geq \nu_n$$

. Thus for instance μ_2 is the second largest eigenvalue of B, etc.

An old question (essentially due to Sylvester, though this particular formulation is due to Weyl) was to determine the complete set of relationships between the λ_i, the μ_j, and the ν_k. There are a number of reasonably obvious equalities and inequalities that one can obtain here. For instance, from the obvious identity $\mathrm{tr}(A)+\mathrm{tr}(B)=\mathrm{tr}(C)$ we conclude the *trace identity*

$$\lambda_1+\ldots+\lambda_n+\mu_1+\ldots+\mu_n=\nu_1+\ldots+\nu_n,$$

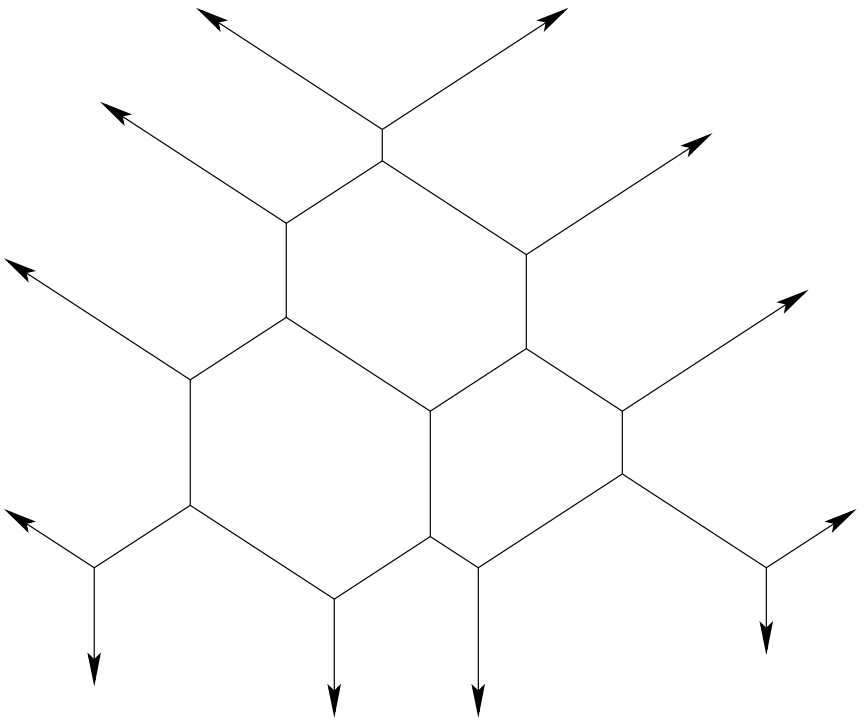


Figure 1.2: A honeycomb.

while from the minimax characterisation of the largest eigenvalue,

$$\lambda_1 = \sup_{\dim(V)=1} \operatorname{tr}(A|_V), \text{ etc.}$$

one easily obtains the triangle inequality

$$\nu_1 \leq \lambda_1 + \mu_1.$$

And so on and so forth. It turns out that the set of all possible λ_i, μ_j, ν_k form a convex cone, determined by a finite number of linear inequalities; this can be derived from symplectic geometry considerations (the Atiyah/Guillemin-Sternberg convexity theorem [At1982], [GuSt1982], or more precisely a refinement due to Kirwan[Ki1984]). A complete (in fact, overcomplete) list of such inequalities, generated by a beautifully recursive formula, was conjectured by Alfred Horn[Ho1962]. The Horn conjecture was finally settled in a combination of two papers: one by Klyachko[Kl1998], which used geometric invariant theory to reduce the problem to a simpler problem known as the *saturation conjecture*, and one by Allen Knutson and myself [KnTa1999], which established the saturation conjecture by a combinatorial argument using honeycombs.

Note that the lengths of the edges in the honeycomb are variable, but there are only three possible orientations, 120 degree angles apart. This is a honeycomb of order

4, with four half-infinite edges going in a NW direction, four in a NE direction, and four in the S direction; the coordinates of these edges are the *boundary data* of this honeycomb. [For more precise definitions, see [KnTa2001], [KnTa1999].] One can also play with honeycombs using our honeycomb applet[KnTa2001b].

Allen and I observed that honeycombs "solve" Weyl's problem in the following sense: a collection λ_i, μ_j, ν_k of putative eigenvalues can arise as the eigenvalues of a triplet $A+B=C$ of matrices if and only if they arise as the boundary values of at least one honeycomb. In fact, there is a more quantitative relationship: the "volume" of the set of all $A+B=C$ with the desired eigenvalues is equal (up to some explicit normalising factors) to the "volume" of all the honeycombs with the specified boundary data.

There is also a "discrete" or "quantised" version of the above connection. One can define an *integral* honeycomb to be one where all lengths and coordinates are integers (in particular, the boundary data will also be integers). It turns out that an integral honeycomb with boundary values λ_i, μ_j, ν_k exists if and only if the irreducible representation of $SU(n)$ with weight vector $(\nu_1, \ldots, \nu_n)$ appears at least once in the tensor product of the irreducible representations of weight vectors $(\lambda_1, \ldots, \lambda_n)$ and $(\mu_1, \ldots, \mu_n)$, and furthermore the number of integral honeycombs counts the multiplicity of this representation. This multiplicity is in fact a *Littlewood-Richardson coefficient*, and appears in a number of other contexts, such as Schubert calculus (intersection numbers for Schubert classes); see the survey [Fu2000] for details. The precise relationship between the discrete and continuous formulations of this problem is in fact the key to the Horn conjecture (and the closely related saturation conjecture), but I will not discuss this in detail here (it is covered in the above references). Honeycombs can also be linked to several other combinatorial gadgets, such as puzzles[KnTa2003], Berenstein-Zelevinsky patterns[BeZe1992], and Young tableaux: see [PaVa2005] this paper for several of these connections (and [ThYo2008] for some further recent developments).

But let us leave this additive setting for now and turn to the analogous multiplicative problem. Here, instead of three Hermitian matrices $A+B=C$, we now have three *unitary* matrices $UV=W$. It is convenient to normalise these matrices to have determinant 1, in which case we can uniquely express the eigenvalues of U as

$$e(\lambda_1), \ldots, e(\lambda_n),$$

where $\lambda_1 \geq \ldots \geq \lambda_n \geq \lambda_1 - 1$ and $\lambda_1 + \ldots + \lambda_n = 0$ and $e(x) := e^{2\pi i x}$. We can similarly express the eigenvalues of V as

$$e(\mu_1), \ldots, e(\mu_n),$$

where $\mu_1 \geq \ldots \geq \mu_n \geq \mu_1 - 1$ and $\mu_1 + \ldots + \mu_n = 0$ and the eigenvalues of W as

$$e(\nu_1), \ldots, e(\nu_n),$$

where $\nu_1 \geq \ldots \geq \nu_n \geq \nu_1 - 1$ and $\nu_1 + \ldots + \nu_n = 0$. We can now ask the multiplicative version of Weyl's problem, namely to characterise all the relationships that exist between the λ_i, the μ_j, and the ν_k. For instance, it is possible to view λ_1 as the "maximum anti-clockwise angle" that U can rotate a vector, which can eventually lead to the

inequality

$$\nu_1 \leq \lambda_1 + \mu_1.$$

One can continue creating inequalities of this type, and there will be a strong resemblance of those inequalities with those in the additive problem. This is not so surprising, since the additive problem emerges as a limiting case of the multiplicative one (if $U = \exp(\varepsilon A), V = \exp(\varepsilon B), W = \exp(\varepsilon C)$ and $UV = W$, then $A + B = C + O(\varepsilon)$ when ε is small, by the Baker-Campbell-Hausdorff formula). What is more surprising is that when the λ_i, μ_j, ν_k are sufficiently small, that the inequalities which describe the multiplicative problem are *exactly* those that describe the additive problem! In fact, it is known that the space of all possible λ_i, μ_j, ν_k for the multiplicative problem is a convex polytope contained within the convex cone for the additive problem, and in fact a quantum version of the Horn conjecture (i.e. an explicit recursive description of the faces of this polytope) was proven by Belkale [Be2008] (building upon earlier work in [AgWo1998], [Be2001]). For instance, while for the additive problem there is the constraint

$$\nu_{i+j-1} \leq \lambda_i + \mu_j$$

whenever $i + j - 1 \leq n$ (the Weyl inequalities), in the multiplicative problem one also has the additional constraint

$$\lambda_i + \mu_j \leq \nu_{i+j} + 1.$$

As with the additive problem, the complete set of all inequalities of this form turns out to be rather messy to describe and I will not do so here.

Just as the additive Weyl problem turned out to be linked to Schubert calculus (the intersection numbers of Schubert classes), the multiplicative problem turned out to be linked to quantum Schubert calculus (the Gromov-Witten numbers of the same classes), and making this link precise turned out to be the key to the proof of the quantum Horn conjecture.

This solves the "qualitative" version of the multiplicative Weyl problem, namely whether there exists any triple $UV = W$ with the specified eigenvalues. However, one can still ask "quantitative" versions, namely to compute the volume of the space of all such triples. There is also the discretised quantitative version, which concerns either the Gromov-Witten numbers for Schubert classes, or else the multiplicities of fusion products in the Verlinde algebra of $SU(n)$; these are rather technical and we refer to [AgWo1998] for details. There should exist some concept of "quantum honeycomb" which computes all of these numbers, in much the same way that the usual honeycombs computes the volume of the space of solutions to $A + B = C$ with specified eigenvalues, intersection numbers for Schubert classes, or multiplicities of tensor products of $SU(n)$ irreducible representations. Vaguely speaking it seems that one wants to construct an analogue of the planar honeycomb which lives instead on something like a two-dimensional torus, but it is not entirely clear (even when $n = 2$) what the precise definition of this object should be.

It may seem like one needs to learn a fearsome amount of machinery to attack this problem, but actually I think one can at least *guess* what the quantum honeycomb should be just by experimentation with small cases $n = 1, 2, 3$ and by using various sanity checks (this is how Allen and I discovered the additive honeycombs). For instance,

the equation $UV = W$ has the cyclic symmetry $(U,V,W) \mapsto (V,W^{-1},U^{-1})$ and so the quantum honeycomb should enjoy a similar symmetry. There is also the translation symmetry $(U,V,W) \mapsto (e(\alpha)U, e(\beta)V, e(\gamma)W)$ whenever $\alpha+\beta+\gamma = 0$, so quantum honeycombs should be translation invariant. When the honeycomb is small (all vertices close to the origin) there should be a bijective correspondence between the quantum honeycomb and the regular honeycomb. The constraints between all the boundary values are already known due to the resolution of the quantum Horn conjecture. There are some other extreme cases which are also understood quite well, for instance when one of the matrices is very close to the identity but the other two are not.

My guess is that once a reasonable candidate for a quantum honeycomb is found which passes all the obvious sanity checks, actually verifying that it computes everything that it should will be a relatively routine matter (we have many different combinatorial ways of establishing things like this). This will give a combinatorial tool for computing a number of interesting quantities, and will probably shed some light also as to why these honeycombs appear in the subject in the first place. (It seems to be somehow related to the Dynkin diagram A_n for the underlying group $SU(n)$; it has proven a little tricky to try to find analogues of these objects for the other Dynkin diagrams.) Certainly they seem to be computing *something* rather non-trivial; for instance the Littlewood-Richardson numbers that are computed by additive honeycombs have even been proposed to play a role in lower bounds in complexity theory, and specifically the $P \neq NP$ problem[Mu2007]!

1.7.1 Notes

This article was originally posted on Apr 19, 2007 at

terrytao.wordpress.com/2007/04/19

A java applet demonstrating honeycoms in action can be found at [KnTa2001b].

Thanks to Allen Knutson for suggestions and encouragement.

1.8 Boundedness of the trilinear Hilbert transform

This is a well-known problem in multilinear harmonic analysis; it is fascinating to me because it lies barely beyond the reach of the best technology we have for these problems (namely, multiscale time-frequency analysis), and because the most recent developments in quadratic Fourier analysis seem likely to shed some light on this problem.

Recall that the Hilbert transform is defined on test functions $f \in \mathscr{S}(\mathbf{R})$ (up to irrelevant constants) as

$$Hf(x) := p.v. \int_{\mathbf{R}} f(x+t)\frac{dt}{t},$$

where the integral is evaluated in the principal value sense (removing the region $|t| < \varepsilon$ to ensure integrability, and then taking the limit as $\varepsilon \to 0$.)

One of the basic results in (linear) harmonic analysis is that the Hilbert transform is bounded on $L^p(\mathbf{R})$ for every $1 < p < \infty$, thus for each such p there exists a finite constant C_p such that

$$\|Hf\|_{L^p(\mathbf{R})} \leq C_p \|f\|_{L^p(\mathbf{R})}.$$

One can view boundedness result (which is of importance in complex analysis and one-dimensional Fourier analysis, while also providing a model case of the more general Calderón-Zygmund theory of singular integral operators) as an assertion that the Hilbert transform is "not much larger than" the identity operator. And indeed the two operators are very similar; both are invariant under translations and dilations, and on the Fourier side, the Hilbert transform barely changes the magnitude of the Fourier transform at all:

$$\hat{Hf}(\xi) = \pi i \operatorname{sgn}(\xi)\hat{f}(\xi).$$

In fact, one can show (see e.g. [St1970]) that the only reasonable (e.g. L^2-bounded) operators which are invariant under translations and dilations are just the linear combinations of the Hilbert transform and the identity operator. (A useful heuristic in this area is to view the singular kernel $p.v.1/t$ as being of similar "strength" to the Dirac delta function $\delta(t)$ - for instance, they have same scale-invariance properties.)

Note that the Hilbert transform is formally a convolution of f with the kernel $1/t$. This kernel is almost, but not quite, absolutely integrable - the integral of $1/|t|$ diverges logarithmically both at zero and at infinity. If the kernel was absolutely integrable, then the above L^p boundedness result would be a simple consequence of Young's inequality (or Minkowski's inequality); the difficulty is thus "just" one of avoiding a logarithmic divergence. To put it another way, if one dyadically decomposes the Hilbert transform into pieces localised at different scales (e.g. restricting to an "annulus" $|t| \sim 2^n$), then it is a triviality to establish boundedness of each component; the difficulty is ensuring that there is enough cancellation or orthogonality that one can sum over the (logarithmically infinite number of) scales and still recover boundedness.

There are a number of ways to establish boundedness of the Hilbert transform. One way is to decompose all functions involved into wavelets - functions which are localised in space and scale, and whose frequencies stay at a fixed distance from the origin (relative to the scale). By using standard estimates concerning how a function can be decomposed into wavelets, how the Hilbert transform acts on wavelets, and how

wavelets can be used to reconstitute functions, one can establish the desired boundedness. The use of wavelets to mediate the action of the Hilbert transform fits well with the two symmetries of the Hilbert transform (translation and scaling), because the collection of wavelets also obeys (discrete versions of) these symmetries. One can view the theory of such wavelets as a dyadic framework for Calderón-Zygmund theory.

Just as the Hilbert transform behaves like the identity, it was conjectured by Calderón (motivated by the study of the Cauchy integral on Lipschitz curves) that the *bilinear Hilbert transform*

$$B(f,g)(x) := p.v. \int_{\mathbf{R}} f(x+t)g(x+2t)\frac{dt}{t}$$

would behave like the pointwise product operator $f,g \mapsto fg$ (exhibiting again the analogy between $p.v.1/t$ and $\delta(t)$), in particular one should have the Hölder-type inequality

$$\|B(f,g)\|_{L^r(\mathbf{R})} \leq C_{p,q}\|f\|_{L^p(\mathbf{R})}\|g\|_{L^q(\mathbf{R})} \tag{1.3}$$

whenever $1 < p,q < \infty$ and $\frac{1}{r} = \frac{1}{p} + \frac{1}{q}$. (There is nothing special about the "2" in the definition of the bilinear Hilbert transform; one can replace this constant by any other constant α except for 0, 1, or ∞, though it is a delicate issue to maintain good control on the constant $C_{p,q}$ when α approaches one of these exceptional values. Note that by setting$g = 1$ and looking at the limiting case $q = \infty$ we recover the linear Hilbert transform theory from the bilinear one; thus we expect the bilinear theory to be harder.) Again, this claim is trivial when localising to a single scale $|t| \sim 2^n$, as it can then be quickly deduced from Hölder's inequality. The difficulty is then to combine all the scales together.

It took some time to realise that Calderón-Zygmund theory, despite being incredibly effective in the linear setting, was not quite the right tool for the bilinear problem. One way to see the problem is to observe that the bilinear Hilbert transform B (or more precisely, the estimate (1.3)) enjoys one additional symmetry beyond the scaling and translation symmetries that the Hilbert transform H obeyed. Namely, one has the *modulation invariance*

$$B(e_{-2\xi}f, e_\xi g) = e_{-\xi}B(f,g)$$

for any frequency ξ, where $e_\xi(x) := e^{2\pi i\xi x}$ is the linear plane wave of frequency ξ, which leads to a modulation symmetry for the estimate (1.3). This symmetry - which has no non-trivial analogue in the linear Hilbert transform - is a consequence of the algebraic identity

$$\xi x - 2\xi(x+t) + \xi(x+2t) = 0$$

which can in turn be viewed as an assertion that linear functions have a vanishing second derivative.

It is a general principle that if one wants to establish a delicate estimate which is invariant under some non-compact group of symmetries, then the *proof* of that estimate should also be largely invariant under that symmetry (or, if it does eventually decide to break the symmetry (e.g. by performing a normalisation), it should do so in a way that will yield some tangible profit). Calderón-Zygmund theory gives the frequency origin $\xi = 0$ a preferred role (for instance, all wavelets have mean zero, i.e. their Fourier

transforms vanish at the frequency origin), and so is not the appropriate tool for any modulation-invariant problem.

The conjecture of Calderón was finally verified in a breakthrough pair of papers by Lacey and Thiele [LaTh1997, LaTh1999], first in the "easy" region $2 < p,q,r' < \infty$ (in which all functions are locally in L^2 and so local Fourier analytic methods are particularly tractable) and then in the significantly larger region where $r > 2/3$. (Extending the latter result to $r = 2/3$ or beyond remains open, and can be viewed as a toy version of the trilinear Hilbert transform question discussed below.) The key idea (dating back to [Fe1973]) was to replace the wavelet decomposition by a more general *wave packet* decomposition - wave packets being functions which are well localised in position, scale, and frequency, but are more general than wavelets in that their frequencies do not need to hover near the origin; in particular, the wave packet framework enjoys the same symmetries as the estimate that one is seeking to prove. (As such, wave packets are a highly overdetermined basis, in contrast to the exact bases that wavelets offers, but this turns out to not be a problem, provided that one focuses more on decomposing the *operator B* rather than the individual functions f, g.) Once the wave packets are used to mediate the action of the bilinear Hilbert transform B, Lacey and Thiele then used a carefully chosen combinatorial algorithm to organise these packets into "trees" concentrated in mostly disjoint regions of phase space, applying (modulated) Calderón-Zygmund theory to each tree, and then using orthogonality methods to sum the contributions of the trees together. (The same method also leads to the simplest proof known[LaTh2000] of Carleson's celebrated theorem[Ca1966] on convergence of Fourier series.)

Since the Lacey-Thiele breakthrough, there has been a flurry of other papers (including some that I was involved in) extending the time-frequency method to many other types of operators; all of these had the characteristic that these operators were invariant (or "morally" invariant) under translation, dilation, and some sort of modulation; this includes a number of operators of interest to ergodic theory and to nonlinear scattering theory. However, in this post I want to instead discuss an operator which does not lie in this class, namely the *trilinear Hilbert transform*

$$T(f,g,h)(x) := p.v. \int_{\mathbf{R}} f(x+t)g(x+2t)h(x+3t)\frac{dt}{t}.$$

Again, since we expect $p.v.1/t$ to behave like $\delta(t)$, we expect the trilinear Hilbert transform to obey a Hölder-type inequality

$$\|T(f,g,h)\|_{L^s(\mathbf{R})} \leq C_{p,q,r}\|f\|_{L^p(\mathbf{R})}\|g\|_{L^q(\mathbf{R})}\|h\|_{L^r(\mathbf{R})} \tag{1.4}$$

whenever $1 < p,q,r < \infty$ and $\frac{1}{s} = \frac{1}{p} + \frac{1}{q} + \frac{1}{r}$. This conjecture is currently unknown for any exponents p,q,r - even the case $p = q = r = 4$, which is the "easiest" case by symmetry, duality and interpolation arguments. The main new difficulty is that in addition to the three existing invariances of translation, scaling, and modulation (actually, modulation is now a two-parameter invariance), one now also has a *quadratic modulation invariance*

$$T(q_{-3\xi}f, q_{3\xi}g, q_{-\xi}h) = q_{-\xi}T(f,g,h)$$

for any "quadratic frequency" ξ, where $q_\xi(x) := e^{2\pi i \xi x^2}$ is the quadratic plane wave of frequency ξ, which leads to a quadratic modulation symmetry for the estimate (1.4). This symmetry is a consequence of the algebraic identity

$$\xi x^2 - 3\xi(x+t)^2 + 3\xi(x+2t)^2 - \xi(x+3t)^2 = 0$$

which can in turn be viewed as an assertion that quadratic functions have a vanishing third derivative.

It is because of this symmetry that time-frequency methods based on Fefferman-Lacey-Thiele style wave packets seem to be ineffective (though the failure is very slight; one can control entire "forests" of trees of wave packets, but when summing up all the relevant forests in the problem one unfortunately encounters a logarithmic divergence; also, it is known that if one ignores the sign of the wave packet coefficients and only concentrates on the magnitude - which one can get away with for the bilinear Hilbert transform - then the associated trilinear expression is in fact divergent). Indeed, wave packets are certainly not invariant under quadratic modulations. One can then hope to work with the next obvious generalisation of wave packets, namely the "chirps" - quadratically modulated wave packets - but the combinatorics of organising these chirps into anything resembling trees or forests seems to be very difficult. Also, recent work in the additive combinatorial approach to Szemerédi's theorem[Sz1975] (as well as in the ergodic theory approaches) suggests that these quadratic modulations might not be the only obstruction, that other "2-step nilpotent" modulations may also need to be somehow catered for. Indeed I suspect that some of the modern theory of Szemerédi's theorem for progressions of length 4 will have to be invoked in order to solve the trilinear problem. (Again based on analogy with the literature on Szemerédi's theorem, the problem of quartilinear and higher Hilbert transforms is likely to be significantly more difficult still, and thus not worth studying at this stage.)

This problem may be too difficult to attack directly, and one might look at some easier model problems first. One that was already briefly mentioned above was to return to the bilinear Hilbert transform and try to establish an endpoint result at $r = 2/3$. At this point there is again a logarithmic failure of the time-frequency method, and so one is forced to hunt for a different approach. Another is to look at the bilinear maximal operator

$$M(f,g)(x) := sup_{r>0} \frac{1}{2r} \int_{-r}^{r} f(x+t)g(x+2t)dt$$

which is a bilinear variant of the Hardy-Littlewood maximal operator, in much the same way that the bilinear Hilbert transform is a variant of the linear Hilbert transform. It was shown by Lacey[La2000] that this operator obeys most of the bounds that the bilinear Hilbert transform does, but the argument is rather complicated, combining the time-frequency analysis with some Fourier-analytic maximal inequalities of Bourgain[Bo1990]. In particular, despite the "positive" (non-oscillatory) nature of the maximal operator, the only known proof of the boundedness of this operator is oscillatory. It is thus natural to seek a "positive" proof that does not require as much use of oscillatory tools such as the Fourier transform, in particular it is tempting to try an additive combinatorial approach. Such an approach has had some success with a slightly easier operator in a similar spirit, in an unpublished paper of Demeter, Thiele,

and myself[DeTaTh2007]. There is also a paper of Christ[Ch2001] in which a different type of additive combinatorics (coming, in fact, from work on the Kakeya problem [KaTa1999]) was used to establish a non-trivial estimate for single-scale model of various multilinear Hilbert transform or maximal operators. If these operators are understood better, then perhaps additive combinatorics can be used to attack the trilinear maximal operator, and thence to the trilinear Hilbert transform. (This trilinear maximal operator, incidentally, has some applications to pointwise convergence of multiple averages in ergodic theory.) In the opposite direction, the recent paper [De2008] uses additive combinatorics methods to establish some ranges of exponents for which the trilinear Hilbert transform is unbounded.

Another, rather different, approach would be to work in the "finite field model" in which the underlying field $\mathbf{R}$ is replaced by a Cantor ring $\overline{F(t)}$ of formal Laurent series over a finite field F; in such dyadic models (see Section 2.6) the analysis is known to be somewhat simpler (in large part because in this non-Archimedean setting it now becomes possible to create wave packets which are localised in both space and frequency). Nazarov has an unpublished proof of the boundedness of the bilinear Hilbert transform in characteristic 3 settings based on a Bellman function approach; it may be that one could achieve something similar over the field of 4 elements for (a suitably defined version of) the trilinear Hilbert transform. This would at least give supporting evidence for the analogous conjecture in $\mathbf{R}$, although it looks unlikely that a positive result in the dyadic setting would have a *direct* impact on the continuous one.

A related question is to find a fast way to compute the quadrilinear form $\int T(f,g,h)k$ if we discretise f,g,h,k to live on a cyclic group $\mathbf{Z}/N\mathbf{Z}$ rather than on $\mathbf{R}$. A naive expansion of this transform requries about $O(N^2)$ operations; this is in contrast to the bilinear Hilbert transform, in which wave packet expansions allow one to compute $\int B(f,g)h$ in $O(N\log N)$ operations. It may be that a fast algorithm for the trilinear Hilbert transform may suggest a good decomposition in order to prove the above conjecture.

1.8.1 Notes

This article was originally posted on May 10, 2007 at

`terrytao.wordpress.com/2007/05/10`

Thanks to Gil Kalai for helpful comments.

1.9 Effective Skolem-Mahler-Lech theorem

The Skolem-Mahler-Lech theorem[Sk1933, Ma1935, Ma1956, Le1953] in algebraic number theory is a significant generalisation of the obvious statement that a polynomial either has finitely many zeroes (in particular, the set of zeroes is bounded), or it vanishes identically. It appeals to me (despite not really being within my areas of expertise) because it is one of the simplest (non-artificial) results I know of which (currently) comes with an *ineffective* bound - a bound which is provably finite, but which cannot be computed! It appears that to obtain an effective result, one may need a rather different proof.

Ineffective bounds seem to arise particularly often in number theory. I am aware of at least three ways in which they come in:

1. By using methods from soft (infinitary) analysis.
2. By using the fact that any finite set in a metric space is bounded (i.e. is contained in a ball of finite radius centred at a designated origin).
3. By using the fact that any set of finite diameter in a metric space is bounded.

Regarding # 1, there are often ways to make these arguments quantitative and effective, as discussed in Section 2.3. But # 2 and # 3 seem to be irreducibly ineffective: if you know that a set A has finite cardinality or finite diameter, you know it has finite distance to the origin, but an upper bound on the cardinality or diameter does not translate to an effective bound on the radius of the ball centred at the origin needed to contain the set. [In the spirit of Section 2.3, one can conclude an effective "meta-bound" on such a set, establishing a large annulus $\{x : N \leq |x| \leq N + F(N)\}$ in which the set has no presence, but this is not particularly satisfactory.] The problem with the Skolem-Mahler-Lech theorem is that all the known proofs use # 2 at some point.

So, what *is* the Skolem-Mahler-Lech theorem? There are many ways to phrase it, but let us use the formulation using linear recurrence sequences, and in particular restrict attention to *integer¿* linear recurrence sequences for simplicity (which was the scope of the original result of Skolem; Mahler and Lech handled algebraic numbers and elements of fields of characteristic zero respectively. The situation in positive characteristic is more subtle, as this recent paper of Derksen[De2007] shows). By definition, an integer linear recurrence sequence is a sequence $x_0, x_1, x_2, \ldots$ of integers which obeys a linear recurrence relation

$$x_n = a_1 x_{n-1} + a_2 x_{n-2} + \ldots + a_d x_{n-d}$$

for some integer $d \geq 1$ (the *degree* of the linear recurrence sequence), some integer coefficients $a_1, \ldots, a_d$, and all $n \geq d$. This data, together with the first d values $x_0, \ldots, x_{d-1}$ of the sequence, clearly determine the entire sequence. The most famous example of a linear recurrence sequence is of course the Fibonacci sequence $0, 1, 1, 2, 3, 5, \ldots$ given by

$$x_n = x_{n-1} + x_{n-2}; \quad x_0 = 0, x_1 = 1.$$

It is also a nice exercise to show that any polynomial sequence (e.g. the squares $0, 1, 4, 9, \ldots$) is a linear recurrence sequence, or more generally that the component-wise sum or product of two linear recurrence sequences is another linear recurrence

sequence. (Hint: this is related to the fact that the sum or product of algebraic integers is again an algebraic integer.)

The Skolem-Mahler-Lech theorem concerns the set of zeroes $Z := \{n \in \mathbf{N} : x_n = 0\}$ of a given integer linear recurrence sequence. In the case of the Fibonacci sequence, the set of zeroes is pretty boring; it is just $\{0\}$. To give a slightly less trivial example, the linear recurrence sequence

$$x_n = x_{n-2}; \quad x_0 = 0, x_1 = 1$$

has a zero set which is the even numbers $\{0,2,4,\ldots\}$. Similarly, the linear recurrence sequence

$$x_n = x_{n-4} + x_{n-2}; \quad x_0 = x_1 = x_3 = 0, x_2 = 1$$

has a zero set $\{0,1,3,5,\ldots\}$, i.e. the odd numbers together with 0. One can ask whether more interesting zero sets are possible; for instance, can one design a linear recurrence system which only vanishes at the square numbers $\{0,1,4,9,\ldots\}$? The Skolem-Mahler-Lech theorem says no:

Theorem 1.2 (Skolem-Mahler-Lech theorem)**.** *The zero set of a linear recurrence set is eventually periodic, i.e. it agrees with a periodic set for sufficiently large n. In fact, a slightly stronger statement is true: the zero set is the union of a finite set and a finite number of residue classes* $\{n \in \mathbf{N} : n = r \bmod m\}$.

Interestingly, all known proofs of this theorem require that one introduce the p-adic integers $\mathbf{Z}_p$ (or a thinly disguised version thereof). Let me quickly sketch a proof as follows (loosely based on the proof of Hansel[Ha1985]). Firstly it is not hard to reduce to the case where the final coefficient a_d is non-zero. Then, by elementary linear algebra, one can get a closed form for the linear recurrence sequence as

$$x_n = \langle A^n v, w \rangle$$

where A is an invertible $d \times d$ matrix with integer coefficients, and v, w are d-dimensional vectors with integer coefficients (one can write A, v, w explicitly in terms of $a_1,\ldots,a_d$ and $x_0,\ldots,x_{d-1}$, but it is not necessary to do so for this argument). Since A is invertible, we can find a large prime p such that A is also invertible modulo p (any prime not dividing $\det(A)$ will do). Let us now fix this p. The invertible matrices $A^n \bmod p$ over the finite field $\mathbf{F}_p$ take on only finitely many values, thus by the pigeonhole principle there must exist a finite $m \geq 1$ such that $A^m = I \bmod p$, where I is the identity matrix. This m is going to be the eventual period of the zero set; more precisely, for every $r = 0,\ldots,m-1$, I claim that the modified zero set $\{n \in \mathbf{N} : x_{mn+r} = 0\}$ is either finite, or equal to all of $\mathbf{N}$ (this will clearly imply the Skolem-Mahler-Lech theorem). To see this claim, suppose that this modified zero set is infinite for some r, thus

$$\langle A^{mn} A^r v, w \rangle = 0$$

for infinitely any n. By construction of m, we can write $A^m = I + pB$ for some integer-valued matrix B, thus $P(n) = 0$ for infinitely many integers n, where

$$P(n) := \langle (1+pB)^n A^r v, w \rangle.$$

This identity makes sense in the (rational) integers $\mathbf{Z}$, and hence also in the larger ring of p-adic integers $\mathbf{Z}_p$. On the other hand, observe from binomial expansion that $P(n)$ can be expressed as a formal power series in p with coefficients polynomial in n:

$$P(n) = \sum_{j=0}^{\infty} p^j P_j(n).$$

Because of this, the function $P : \mathbf{Z} \to \mathbf{Z}$ extends continuously to a function $P : \mathbf{Z}_p \to \mathbf{Z}_p$, such that $n \mapsto P(n) \bmod p^j$ is a polynomial in n for each j. In other words, P is a uniform limit of polynomials in the p-adic topology. [Note that the Stone-Weierstrass theorem is not applicable here, because we are using p-adic valued functions rather than real or complex-valued ones.] What we need to show is that if P has infinitely many zeroes, then it vanishes everywhere (which is of course what we expect polynomial-like objects to do). Now if $P(n_0) = 0$ for some n_0 (either an integer or a p-adic integer, it doesn't matter), one can then formally divide $P(n) = P(n) - P(n_0)$ by $n - n_0$ (as in the factor theorem from high school algebra) to conclude that $P(n) = (n - n_0)Q(n)$ for some continuous function Q which, like P, is a polynomial modulo p^j for each j, and whose "constant coefficient" $Q_0(n)$ either vanishes, or has degree strictly less than the corresponding coefficient $P_0(n)$ of P. Iterating this fact, we eventually see that if P has infinitely many zeroes, then it contains a factor with vanishing constant coefficient, or in other words it is divisible by p. We iterate this fact again and conclude that if P has infinitely many zeroes then it must be divisible by arbitrarily many powers of p, and thus must vanish everywhere. This concludes the proof.

Now, the above proof clearly gives a quite effective and computable bound on the eventual period m of the zero set. By working somewhat harder, Evertse, Schlickewei, and Schmidt[EvScSc2002] obtained an effective bound on how many exceptional zeroes there are - zeroes of one of these almost polynomials P which do not cause P to vanish entirely. But there appears to be no effective bound known as to how *large* these zeroes are! In particular, one does not even know how to decide whether this set is non-empty, thus we have the following open problem:

Problem 1.3. Given an integer linear recurrence sequence (i.e. given the data $d, a_1, \ldots, a_d, x_0, \ldots, x_d$ as integers), is the truth of the statement "$x_n \neq 0$ for all n" decidable in finite time?

(Note that I am only asking here for decidability, and not even asking for effective bounds.) It is faintly outrageous that this problem is still open; it is saying that we do not know how to decide the halting problem even for "linear" automata!

The basic problem seems to boil down to one of determining whether an "almost polynomial" $P : \mathbf{Z}_p \to \mathbf{Z}_p$ (i.e. a uniform limit of polynomials) has an *integer* zero or not. It is not too hard to find the *p-adic* zeroes of P to any specified accuracy (by using the p-adic version of Newton's method, i.e. Hensel's lemma), but it seems that one needs to know the zeroes to *infinite* accuracy in order to decide whether they are integers or not. It may be that some techniques from Diophantine approximation (e.g. some sort of p-adic analogue of the Thue-Siegel-Roth theorem[Ro1955]) are relevant. Alternatively, one might want to find a completely different proof of the Skolem-Mahler-Lech theorem, which does not use p-adics at all.

1.9.1 Notes

This article was originally posted on May 25, 2007 at

terrytao.wordpress.com/2007/05/25

I thank Kousha Etessami and Tom Lenagan for drawing this problem to my attention. Thanks to Johan Richter and Maurizio for corrections.

Akshay Venkatesh, Jordan Ellenberg, and Felipe Voloch observed that there were several deeper theorems in arithmetic geometry than the Skolem-Mahler-Lech theorem which were similarly ineffective, such as Chabauty's theorem and Falting's theorem; indeed one can view these three theorems as counting rational points on linear tori, abelian varieties, and higher genus varieties respectively.

Kousha Etessami also pointed out the work of Blondel and Portier[BlPo2002] showing the NP-hardness of determining whether an integer linear recurrence contained a zero, as well as the survey [HaHaHiKa2005].

1.10 The parity problem in sieve theory

The parity problem is a notorious problem in sieve theory: this theory was invented[1] in order to count prime patterns of various types (e.g. twin primes), but despite superb success in obtaining upper bounds on the number of such patterns, it has proven to be somewhat disappointing in obtaining lower bounds. Even the task of reproving Euclid's theorem - that there are infinitely many primes - seems to be extremely difficult to do by sieve theoretic means, unless one of course injects into the theory an estimate at least as strong as Euclid's theorem (such as the prime number theorem). The main obstruction is the *parity problem*: even assuming such strong hypotheses as the Elliott-Halberstam conjecture[ElHa1969] (a sort of "super-generalised Riemann Hypothesis" for sieves), sieve theory is largely (but not completely) unable to distinguish numbers with an odd number of prime factors from numbers with an even number of prime factors. This "parity barrier" has been broken for some select patterns of primes by injecting some powerful non-sieve theory methods into the subject, but remains a formidable obstacle in general.

I'll discuss the parity problem in more detail later in this article, but I want to first discuss how sieves work (drawing in part on some excellent unpublished lecture notes of Iwaniec); the basic ideas are elementary and conceptually simple, but there are many details and technicalities involved in actually executing these ideas, and which I will try to suppress for sake of exposition.

1.10.1 A brief history of sieve theory

Let's consider a basic question in prime number theory, namely how to count the number of primes in a given range, say between N and $2N$ for some large integer N. (This problem is more or less equivalent to that of counting primes between 1 and N, thanks to dyadic decomposition, but by keeping the magnitude of all numbers comparable to N we can simplify some (very minor) technicalities.) Of course, we know that this particular question can be settled fairly satisfactorily (the answer is $(1+o(1))\frac{N}{\log N}$) using known facts about the Riemann zeta function, but let us pretend for now that we do not know about this function. (Once one moves to slightly more complicated additive questions about the primes, such as counting twin primes, the theory of the zeta function and its relatives becomes much less powerful, even assuming such things as the Riemann hypothesis; the problem is that these functions are measuring the *multiplicative* structure of the primes rather than the additive structure.)

The set of primes does not appear to have enough usable structure in order to perform such counts quickly. However, one can count other sets of numbers between N and $2N$ with much more ease. For instance, the set of integers between N and $2N$ can be easily counted with small error:

$$|\{n \in [N,2N] : n \text{ integer}\}| = N + O(1);$$

the error term $O(1)$ in this case is in fact just 1. (Here we use $|X|$ to denote the cardinality of a finite set X.) Similarly, we can count, say, the number of odd numbers

[1] Sieves can also be used to study many other things than primes, of course, but we shall focus only on primes in this article.

between N and $2N$,

$$|\{n \in [N,2N] : n \text{ odd}\}| = \frac{1}{2}N + O(1),$$

simply because the set of odd numbers has density $\frac{1}{2}$ and is periodic of period 2. The error term $O(1)$ now depends on the parity of N. More generally, we can count any given residue class in $[N,2N]$ to a reasonable accuracy:

$$|\{n \in [N,2N] : n = a \bmod q\}| = \frac{1}{q}N + O(1),$$

where the error term is now more complicated, and depends on what N is doing modulo q. This estimate is quite good as long as q is small compared with N, but once q is very large, the error term $O(1)$ can begin to overwhelm the main term (especially if the main term is going to appear in a delicate summation with lots of cancellation). In general, any summation involving the main term N/q will be relatively easy to manipulate (because it is essentially multiplicative in q, and thus amenable to all the methods of multiplicative number theory, in particular Euler products and zeta functions); it is the error term $O(1)$ which causes all the difficulty.

Once we have figured out how to count these basic sets, we can also count some combinations of these sets, as long as these combinations are simple enough. For instance, suppose we want to count

$$|\{n \in [N,2N] : n \text{ coprime to } 2,3\}| \tag{1.5}$$

Well, we know that the total number of integers in $[N,2N]$ is $N + O(1)$. Of this set, we know that $\frac{1}{2}N + O(1)$ of the elements are not coprime to 2 (i.e. they are divisible by 2), and that $\frac{1}{3}N + O(1)$ are not coprime to 3. So we should subtract those two sets from the original set, leaving $\frac{1}{6}N + O(1)$. But the numbers which are divisible by both 2 and 3 (i.e. divisible by 6) have been subtracted twice, so we have to put them back in; this adds in another $\frac{1}{6}N + O(1)$, giving a final count of $\frac{1}{3}N + O(1)$ for the quantity (1.5); this is of course a simple instance of the inclusion-exclusion principle in action. An alternative way to estimate (1.5) is to use the Chinese remainder theorem to rewrite (1.5) as

$$|\{n \in [N,2N] : n = 1,5 \bmod 6\}|$$

and use our ability to count residue classes modulo 6 to get the same final count of $\frac{1}{3}N + O(1)$ (though the precise bound on the error term will be slightly different). For very small moduli such as 2 and 3, the Chinese remainder theorem is quite efficient, but it is somewhat rigid, and for higher moduli (e.g. for moduli much larger than $\log N$) it turns out that the more flexible inclusion-exclusion principle gives much better results (after applying some tricks to optimise the efficiency of that principle).

We can of course continue the example of (1.5), counting the numbers in $[N,2N]$ which are coprime to 2, 3, 5, 7, etc., which by the sieve of Eratosthenes will eventually give us a count for the primes in $[N,2N]$, but let us pause for a moment to look at the larger picture. We have seen that some sets in $[N,2N]$ are fairly easy to count accurately (e.g. residue classes with small modulus), and others are not (e.g. primes, twin primes).

What is the defining characteristic of the former types of sets? One reasonable answer is that the sets that are easy to count are *low-complexity*, but this is a rather vaguely defined term. I would like to propose instead that sets (or more generally, weight functions - see below) are easy to count (or at least estimate) whenever they are *smooth* in a certain sense to be made more precise shortly. This terminology comes from harmonic analysis rather than from number theory (though number theory does have the related concept of a *smooth number*), so I will now digress a little bit to talk about smoothness, as it seems to me that this concept implicitly underlies the basic strategy of sieve theory.

Instead of talking about the problem of (approximately) counting a given set in $[N, 2N]$, let us consider instead the analogous problem of (approximately) computing the area of a given region E (e.g. a solid ellipse) in the unit square $[0,1]^2$. As we are taught in high school, one way to do this is to subdivide the square into smaller squares, e.g. squares of length 10^{-n} for some n, and count how many of these small squares lie completely or partially in the set E, and multiply by the area of each square; this is of course the prelude to the Riemann integral. It works well as long as the set E is "smooth" in the sense that most of the small squares are either completely inside or completely outside the set E, with few borderline cases; this notion of smoothness can viewed as a quantitative version of Riemann integrability. Another way of saying this is that if one wants to determine whether a given point (x,y) lies in E, it is usually enough just to compute x and y to the first n significant digits in the decimal expansion.

Now we return to counting sets in $[N, 2N]$. One can also define the notion of a "smooth set" here by again using the most significant digits of the numbers n in the interval $[N, 2N]$; for instance, the set $[1.1N, 1.2N]$ would be quite smooth, as one would be fairly confident whether n would lie in this set or not after looking at just the top two or three significant digits. However, with this "Euclidean" or "Archimedean" notion of smoothness, sets such as the primes or the odd numbers are certainly not smooth. However, things look a lot better if we change the metric, or (more informally) if we redefine what "most significant digit" is. For instance, if we view the *last* digit in the base 10 expansion of a number n (i.e. the value of n mod 10) as the most significant one, rather than the first - or more precisely, if we use the 10-adic metric intead of the Euclidean one, thus embedding the integers into $\mathbf{Z}_{10}$ rather than into $\mathbf{R}$ - then the odd numbers become quite smooth (the most significant digit completely determines membership in this set). The primes in $[N, 2N]$ are not fully smooth, but they do exhibit some partial smoothness; indeed, if the most significant digit is 0, 2, 4, 5, 6, or 8, this fully determines membership in the set, though if the most significant digit is 1, 3, 7, or 9 then one only has partial information on membership in the set.

Now, the 10-adic metric is not fully satisfactory for characterising the elusive concept of number-theoretic "smoothness". For instance, the multiples of 3 should be a smooth set, but this is not the case in the 10-adic metric (one really needs *all* the digits before one can be sure whether a number is a multiple of 3!). Also, we have the problem that the set $[N/2, N]$ itself is now no longer smooth. This can be fixed by working not with just the Euclidean metric or a single n-adic metric, but with the product of *all* the n-adic metrics and the Euclidean metric at once. Actually, thanks to the Chinese remainder theorem, it is enough to work with the product of the p-adic metrics for primes p and the Euclidean metric, thus embedding the integers in the integer

adele ring $\mathbf{R} \times \prod_p \mathbf{Z}_p$. For some strange reason, this adele ring is not explicitly used in most treatments of sieve theory, despite its obvious relevance (and despite the amply demonstrated usefulness of this ring in algebraic number theory or in the theory of L-functions, as exhibited for instance by Tate's thesis[Ta1950]). At any rate, we are only using the notion of "smoothness" in a very informal sense, and so we will not need the full formalism of the adeles here. Suffice to say that a set of integers in $[N, 2N]$ is "smooth" if membership in that set can be largely determined by its most significant digits in the Euclidean sense, and also in the p-adic senses for all small p; roughly speaking, this means that this set is approximately the pullback of some "low complexity" set in the adele ring - a set which can be efficiently fashioned out of a few of basic sets which generate the topology and σ-algebra of that ring. (Actually, in many applications of sieve theory, we only need to deal with moduli q which are square-free, which means that we can replace the p-adics $\mathbf{Z}_p$ with the cyclic group $\mathbf{Z}/p\mathbf{Z}$, and so it is now just the residues mod p for small p, together with the Euclidean most significant digits, which should control what smooth sets are; thus the adele ring has been replaced by the product $\mathbf{R} \times \prod_p (\mathbf{Z}/p\mathbf{Z})$.)

[A little bit of trivia: the idea of using $\mathbf{R} \times \prod_p (\mathbf{Z}/p\mathbf{Z})$ as a proxy for the integers seems to go all the way back to Sun Tzu, who introduced the Chinese Remainder Theorem in order to efficiently count the number of soldiers in an army, by making them line up in columns of (say) 7, 11, and 13 and count the three remainders, thus determining this number up to a multiple of $7 \times 11 \times 13 = 1001$; doing a crude calculation to compute the most significant digits in $\mathbf{R}$ of size of the army would then finish the job.]

Let us now return back to sieve theory, and the task of counting "rough" sets such as the primes in $[N, 2N]$. Since we know how to accurately count "smooth" sets such as $\{n \in [N, 2N] : n = a \bmod q\}$ with q small, one can try to describe the rough set of primes as some sort of combination of smooth sets. The most direct implementation of this idea is the sieve of Eratosthenes; if one then tries to compute the number of primes using the inclusion-exclusion principle, one obtains the *Legendre sieve* (we implicitly used this idea previously when counting the quantity (1.5)). However, the number of terms in the inclusion-exclusion formula is very large; if one runs the sieve of Eratosthenes for k steps (i.e. sieving out multiples of the first k primes), there are basically 2^k terms in the inclusion-exclusion formula, leading to an error term which in the worst case could be of size $O(2^k)$. A related issue is that the modulus q in many of the terms in the Legendre sieve become quite large - as large as the product of the first k primes (which turns out to be roughly e^k in size). Since the set one is trying to count is only of size N, we thus see that the Legendre sieve becomes useless after just $\log N$ or so steps of the Eratosthenes sieve, which is well short of what one needs to accurately count primes (which requires that one uses $N^{1/2}/\log N$ or so steps). More generally, "exact" sieves such as the Legendre sieve are useful for any situation involving only a logarithmically small number of moduli, but are unsuitable for sieving with much larger numbers of moduli.

One can view the early development of sieve theory as a concerted effort to rectify the drawbacks of the Legendre sieve. The first main idea here is to not try to compute the size of the rough set exactly - as this is too "expensive" in terms of the number of smooth sets required to fully describe the rough set - but instead to just settle for upper

or lower bounds on the size of this set, which use fewer smooth sets. There is thus a tradeoff between how well the bounds approximate the original set, and how well one can compute the bounds themselves; by selecting various parameters appropriately one can optimise this tradeoff and obtain a final bound which is non-trivial but not completely exact. For instance, in using the Legendre sieve to try to count primes between N and $2N$, one can instead use that sieve to count the much larger set of numbers between N and $2N$ which are coprime to the first k primes, thus giving an upper bound for the primes between N and $2N$. It turns out that the optimal value of k here is roughly $\log N$ or so (after this, the error terms in the Legendre sieve get out of hand), and give an upper bound of $O(N/\log\log N)$ for the number of primes between N and $2N$ - somewhat far from the truth (which is $\sim N/\log N$), but still non-trivial.

In a similar spirit, one can work with various truncated and approximate versions of the inclusion-exclusion formula which involve fewer terms. For instance, to estimate the cardinality $|\bigcup_{j=1}^k A_j|$ of the union of k sets, one can replace the inclusion-exclusion formula

$$\begin{aligned} |\bigcup_{j=1}^k A_j| = \sum_{j=1}^k |A_j| - \sum_{1\leq j_1<j_2\leq k} |A_{j_1}\cap A_{j_2}| \\ + \sum_{1\leq j_1<j_2<j_3\leq k} |A_{j_1}\cap A_{j_2}\cap A_{j_3}| \ldots \end{aligned} \tag{1.6}$$

by the obvious upper bound

$$|\bigcup_{j=1}^k A_j| \leq \sum_{j=1}^k |A_j|$$

(also known as the *union bound*), or by the slightly less obvious lower bound

$$|\bigcup_{j=1}^k A_j| \geq \sum_{j=1}^k |A_j| - \sum_{1\leq j_1<j_2\leq k} |A_{j_1}\cap A_{j_2}|.$$

More generally, if one takes the first n terms on the right-hand side of (1.6), this will be an upper bound for the left-hand side for odd n and a lower bound for even n. These inequalities, known as the *Bonferroni inequalities*, are a nice exercise to prove: they are equivalent to the observation that in the binomial identity

$$0 = (1-1)^m = \binom{m}{0} - \binom{m}{1} + \binom{m}{2} - \binom{m}{3} + \ldots + (-1)^m \binom{m}{m}$$

for any $m \geq 1$, the partial sums on the right-hand side alternate in sign between non-negative and non-positive. If one inserts these inequalities into the Legendre sieve and optimises the parameter, one can improve the upper bound for the number of primes in $[N,2N]$ to $O(N\log\log N/\log N)$, which is significantly closer to the truth. Unfortunately, this method does not provide any lower bound other than the trivial bound of 0; either the main term is negative, or the error term swamps the main term. A similar argument was used by Brun[HaRi1974] to show that the number of twin primes in $[N,2N]$ was $O(N(\log\log N/\log N)^2)$ (again, the truth is conjectured to be $\sim N/\log^2 N$),

which implied his famous theorem that the sum of reciprocals of the twin primes is convergent.

The full inclusion-exclusion expansion is a sum over 2^k terms, which one can view as binary strings of 0s and 1s of length k. In the Bonferroni inequalities, one only sums over a smaller collection of strings, namely the *Hamming ball* of strings which only involve n or fewer 1s. There are other collections of strings one can use which lead to upper or lower bounds; one can imagine revealing such a string one digit at a time and then deciding whether to keep or toss out this string once some threshold rule is reached. There are various ways to select these thresholding rules, leading to the family of *combinatorial* sieves. One particularly efficient such rule is similar to that given by the Bonferroni inequalities, but instead of using the number of 1s in a string to determine membership in the summation, one uses a *weighted* number of 1s (giving large primes more weight than small primes, because they tend to increase the modulus too quickly and thus should be removed from the sum sooner than the small primes). This leads to the *beta sieve*, which for instance gives the correct order of magnitude of $O(N/\log N)$ for the number of primes in $[N,2N]$ or $O(N/\log^2 N)$ for the number of twin primes in $[N,2N]$. This sieve is also powerful enough to give lower bounds, but only if one stops the sieve somewhat early, thus enlarging the set of primes to a set of *almost primes* (numbers which are coprime to all numbers less than a certain threshold, and thus have a bounded number of prime factors). For instance, this sieve can show that there are an infinite number of twins $n, n+2$, each of which has at most nine prime factors (the number nine is not optimal, but to get better results requires much more work).

There seems however to be a limit as to what can be accomplished by purely combinatorial sieves. The problem stems from the "binary" viewpoint of such sieves: any given term in the inclusion-exclusion expansion is either included or excluded from the sieve upper or lower bound, and there is no middle ground. This leads to the next main idea in modern sieve theory, which is to work not with the cardinalities of sets in $[N,2N]$, but rather with more flexible notion of sums of *weight functions* (real-valued functions on [N,2N]). The starting point is the obvious formula

$$|A| = \sum_{n \in [N,2N]} 1_A(n)$$

for the cardinality of a set A in $[N,2N]$, where 1_A is the indicator function of the set A. Applying this to smooth sets such as $\{n \in [N,2N] : n = a \bmod q\}$, we obtain

$$\sum_{n \in [N,2N]} 1_{n=a \bmod q}(n) = \frac{N}{q} + O(1);$$

in particular, specialising to the 0 residue class 0 mod q (which is the residue class of importance for counting primes) we have

$$\sum_{n \in [N,2N]} 1_{d|n}(n) = \frac{N}{d} + O(1)$$

for any d. Thus if we can obtain a pointwise upper bound on 1_A by a divisor sum

(which is a number-theoretic analogue of a *smooth function*), thus

$$1_A(n) \leq \sum_d c_d 1_{d|n}(n) \tag{1.7}$$

for all n and some real constants c_d (which could be positive or negative), then on summing we obtain the upper bound

$$|A| \leq N \sum_d \frac{c_d}{d} + O(\sum_d |\lambda_d|). \tag{1.8}$$

One can also hope to obtain lower bounds on $|A|$ by a similar procedure (though in practice, lower bounds for primes have proven to be much more difficult to obtain, due to the parity problem which we will discuss below). These strategies are suited for the task of bounding the number of primes in $[N, 2N]$; if one wants to do something fancier such as counting twin primes $n, n+2$, one has to either involve more residue classes (e.g. the class $-2 \bmod q$ will play a role in the twin prime problem) or else insert additional weights in the summation (e.g. weighting all summations in n by an additional factor of $\Lambda(n+2)$, where Λ is the von Mangoldt function). To simplify the exposition, though, we shall just stick with the plainer problem of counting primes.

The above strategies generalise the combinatorial sieve strategy, which is a special case in which the constants c_d are restricted to be $+1$, 0, or -1. In practice, the sum $\sum_d \frac{c_d}{d}$ in (1.8) is relatively easy to sum by multiplicative number theory techniques; the coefficients c_d, in applications, usually involve the *Möbius function* $\mu(d)$ (which is unsurprising, since they are encoding some sort of inclusion-exclusion principle), and are often related to the coefficients of a Hasse-Weil zeta function, as they basically count solutions modulo d to some set of algebraic equations. The main task is thus to ensure that the error term in (1.8) does not swamp the main term. To do this, one basically needs the weights c_d to be concentrated on those d which are relatively small compared with N, for instance they might be restricted to some range $d \leq R$ where the *sieve level* $R = N^\theta$ is some small power of N. Thus for instance, starting with the identity

$$\Lambda(n) = \sum_{d|n} \mu(d) \log \frac{n}{d} = -\sum_{d|n} \mu(d) \log(d), \tag{1.9}$$

which corresponds to the zeta-function identity

$$-\frac{\zeta'(s)}{\zeta(s)} = \zeta(s) \frac{d}{ds} \frac{1}{\zeta(s)},$$

where Λ is the von Mangoldt function and μ is the Möbius function, we obtain the upper bound

$$1_A(n) \leq -\sum_{d \leq 2N} \mu(d) \frac{\log d}{\log N} 1_{d|n}$$

where A denotes the primes from N to $2N$. This is already enough (together with the elementary asymptotic $\sum_{d \leq 2N} \frac{\mu(d)}{d} \log d = O(1)$) to obtain the weak prime number theorem $|A| = O(N/\log N)$, but unfortunately this method does not give a nontrivial lower bound for $|A|$. However, a variant of the method does give a nice asymptotic for

P_2 almost primes - products of at most two (large) primes (e.g. primes larger than N^ε for some fixed $\varepsilon > 0$). Indeed, if one introduces the second von Mangoldt function

$$\Lambda_2(n) := \sum_{d|n} \mu(d) \log^2(\frac{n}{d}) = \Lambda(n)\log n + \sum_{d|n} \Lambda(d)\Lambda(\frac{n}{d}) \tag{1.10}$$

which is mostly supported on P_2 almost primes (indeed, $\Lambda_2(p) = \log^2 p$ and $\Lambda_2(pq) = 2\log p \log q$ for distinct primes p, q, and Λ_2 is mostly zero otherwise), and uses the elementary asymptotic

$$\sum_{d \leq 2N} \frac{\mu(d)}{d} \log^2 d = 2\log N + O(1)$$

then one obtains the *Selberg symmetry formula*

$$\sum_{n \leq N} \Lambda_2(N) = 2N \log N + O(N).$$

This formula (together with the weak prime number theorem mentioned earlier) easily implies an "P_2 almost prime number theorem", namely that the number of P_2 almost primes less than N is $(2+o(1))\frac{N}{\log N}$. [This fact is much easier to prove than the prime number theorem itself. In terms of zeta functions, the reason why the prime number theorem is difficult is that the simple pole of $\frac{\zeta'(s)}{\zeta(s)}$ at $s=1$ could conceivably be counteracted by other simple poles on the line $\mathrm{Re}(s) = 1$. On the other hand, the P_2 almost prime number theorem is much easier because the effect of the *double* pole of $\frac{\zeta''(s)}{\zeta(s)}$ at $s=1$ cannot be counteracted by the other poles on the line $\mathrm{Re}(s) = 1$, which are at most simple.]

The P_2 almost prime number theorem establishes the prime number theorem "up to a factor of 2". It is surprisingly difficult to improve upon this factor of 2 by elementary methods, though once one can replace 2 by $2-\varepsilon$ for some $\varepsilon > 0$ (a fact which is roughly equivalent to the absence of zeroes of $\zeta(s)$ on the line $\mathrm{Re}(s) = 1$), one can iterate the Selberg symmetry formula (together with the tautological fact that an P_2 almost prime is either a prime or the product of two primes) to get the prime number theorem; this is essentially the Erdős-Selberg [Er1949, Se1949] elementary proof of that theorem.

One can obtain other divisor bounds of the form (1.7) by various tricks, for instance by modifying the weights in the above formulae (1.9), (1.10). A surprisingly useful upper bound for the primes between N and $2N$ is obtained by the simple observation that

$$1_A(n) \leq (\sum_{d<N} \lambda_d 1_{d|n}(n))^2$$

whenever λ_d are *arbitrary* real numbers with $\lambda_1 = 1$, basically because the square of any real number is non-negative. This leads to the *Selberg sieve*, which suffices for many applications; for instance, it can prove the *Brun-Titchmarsh inequality*[Ti1930], which asserts that the number of primes between N and $N+M$ is at most $(2+o(1))M/\log M$, which is again off by a factor of 2 from the truth when N and M are reasonably comparable. (The $o(1)$ error can even be essentially deleted by working harder; see

[MoVa1973].) There are also some useful lower bounds for the indicator function of the almost primes of divisor sum type, which can be used for instance to derive Chen's theorem[Ch1973] that there are infinitely many primes p such that $p+2$ is a P_2 almost prime, or the theorem that there are infinitely many P_2 almost primes of the form n^2+1.

1.10.2 The parity problem

To summarise the above discussion, sieve theory methods can provide good upper bounds, lower bounds, and even asymptotics for almost primes, which lead to upper bounds for primes which tend to be off by a constant factor such as 2. Rather frustratingly, though, sieve methods have proven largely unable to count or even lower bound the primes themselves, thus leaving the twin prime conjecture (or similar conjectures, such as the conjecture that there are infinitely many primes of the form n^2+1) still out of reach. The reason for this - the *parity problem* - was first clarified by Selberg[Gr2001]. Roughly speaking, it asserts:

Problem 1.4 (Parity problem)**.** If A is a set whose elements are all products of an odd number of primes (or are all products of an even number of primes), then (without injecting additional ingredients), sieve theory is unable to provide non-trivial lower bounds on the size of A. Also, any upper bounds must be off from the truth by a factor of 2 or more.

Thus we can hope to count P_2 almost primes (because they can have either an odd or an even number of factors), or to count numbers which are the product of 6 or 7 primes (which can for instance be done by a sieve of Bombieri[Bo1977]), but we cannot hope to use plain sieve theory to just count primes, or just count semiprimes (the product of exactly two primes).

To explain this problem, we introduce the Liouville function $\lambda(n)$ (a close relative of the Möbius function), which is equal to $+1$ when n is the product of an even number of primes and -1 otherwise. Thus the parity problem applies whenever λ is identically $+1$ or identically -1 on the set A of interest.

The Liouville function oscillates quite randomly between $+1$ and -1. Indeed, the prime number theorem turns out to be equivalent to the assertion that λ is asymptotically of mean zero,

$$\sum_{n\leq N} \lambda(n) = o(N)$$

(a fact first observed by Landau), and if the Riemann hypothesis is true then we have a much better estimate

$$\sum_{n\leq N} \lambda(n) = O_\varepsilon(N^{1/2+\varepsilon}) \text{ for all } \varepsilon > 0.$$

Assuming the generalised Riemann hypothesis, we have a similar claim for residue classes:

$$\sum_{n\leq N} 1_{n=a \bmod q}\lambda(n) = O_\varepsilon(N^{1/2+\varepsilon}) \text{ for all } \varepsilon > 0.$$

What this basically means is that the Liouville function is essentially orthogonal to all smooth sets, or all smooth functions. Since sieve theory attempts to estimate everything

in terms of smooth sets and functions, it thus cannot eliminate an inherent ambiguity coming from the Liouville function. More concretely, let A be a set where λ is constant (e.g. λ is identically -1, which would be the case if A consisted of primes) and suppose we attempt to establish a lower bound for the size of a set A in, say, $[N,2N]$ by setting up a divisor sum lower bound

$$1_A(n) \geq \sum_d c_d 1_{d|n}(n), \tag{1.11}$$

where the divisors d are concentrated in $d \leq R$ for some reasonably small sieve level R. If we sum in n we obtain a lower bound of the form

$$|A| \geq \sum_d c_d \frac{N}{d} + \dots \tag{1.12}$$

and we can hope that the main term $\sum_d c_d \frac{N}{d}$ will be strictly positive and the error term is of lesser order, thus giving a non-trivial lower bound on $|A|$. Unfortunately, if we multiply both sides of (1.11) by the non-negative weight $1+\lambda(n)$ and sum in n, we obtain

$$0 \geq \sum_d c_d 1_{d|n}(n)(1+\lambda(n))$$

since we are assuming λ to equal -1 on A. If we sum this in n, and use the fact that λ is essentially orthogonal to divisor sums, we obtain

$$0 \geq \sum_d c_d \frac{N}{d} + \dots$$

which basically means that the bound (1.12) cannot improve upon the trivial bound $|A| \geq 0$. A similar argument using the weight $1-\lambda(n)$ also shows that any upper bound on $|A|$ obtained via sieve theory has to essentially be at least as large as $2|A|$.

Despite this parity problem, there are a few results in which sieve theory, in conjunction with other methods, can be used to count primes. The first of these is the elementary proof of the prime number theorem alluded to earlier, using the multiplicative structure of the primes inside the almost primes. This method unfortunately does not seem to generalise well to non-multiplicative prime counting problems; for instance, the product of twin primes is not a twin almost prime, and so these methods do not seem to have much hope of resolving the twin prime conjecture. Other examples arise if one starts counting certain special *two-parameter* families of primes; for instance, Friedlander and Iwaniec[FrIw1998] showed that there are infinitely many primes of the form a^2+b^4 by a lengthy argument which started with Vaughan's identity, which is sort of like an exact sieve, but with a (non-smooth) error term which has the form of a bilinear sum, which captures correlation with the Liouville function. The main difficulty is to control this bilinear error term, which after a number of (non-trivial) arithmetic manipulations (in particular, factorising a^2+b^4 over the Gaussian integers) reduces to understanding some correlations between the Möbius function and the Jacobi symbol, which is then achieved by a variety of number-theoretic tools. The method was then modified by Heath-Brown[HB2001] to also show infinitely many primes of the form

a^3+2b^3. Related results for other cubic forms using similar methods have since been obtained in [HBMo2002], [He2006] (analogous claims for quadratic forms date back to [Iw1974]). These methods all seem to require that the form be representable as a norm over some number field and so it does not seem as yet to yield a general procedure to resolve the parity problem.

The parity problem can also be sometimes be overcome when there is an exceptional Siegel zero, which basically means that there is a quadratic character $\chi(n) = \left(\frac{n}{q}\right)$ which correlates very strongly with the primes. Morally speaking, this means that the primes can be largely recovered from the P_2 almost primes as being those almost primes which are quadratic non-residues modulo the conductor q of χ, and this additional information seems (in principle, at least) to overcome the parity problem obstacle (related to this is the fact that Siegel zeroes, if they exist, disprove the generalised Riemann hypothesis, and so the Liouville function is no longer as uniformly distributed on smooth sets as Selberg's analysis assumed). For instance, Heath-Brown[HB1983] showed that if a Siegel zero existed, then there are infinitely many prime twins. Of course, assuming GRH then there are no Siegel zeroes, in which case these results would be technically vacuous; however, they do suggest that to break the parity barrier, we may assume without loss of generality that there are no Siegel zeroes.

Another known way to partially get around the parity problem is to combine precise asymptotics on almost primes (or of weight functions concentrated near the almost primes) with a lower bound on the number of primes, and then use combinatorial tools to parlay the lower bound on primes into lower bounds on prime patterns. For instance, suppose you knew could count the set

$$A := \{n \in [N,2N] : n, n+2, n+6 \in P_2\}$$

accurately (where P_2 is the set of P_2-almost primes), and also obtain sufficiently good lower bounds on the sets

$$A_1 := \{n \in A : n \text{ prime}\}$$

$$A_2 := \{n \in A : n+2 \text{ prime}\}$$

$$A_3 := \{n \in A : n+6 \text{ prime}\},$$

and more precisely that one obtains

$$|A_1| + |A_2| + |A_3| > |A|.$$

(For comparison, the parity problem predicts that one cannot hope to do any better than showing that $|A_1|, |A_2|, |A_3| \geq |A|/2$, so the above inequality is not ruled out by the parity problem obstruction.)

Then, just from the pigeonhole principle, one deduces the existence of $n \in [N,2N]$ such that at least two of n, $n+2$, $n+6$ are prime, thus yielding a pair of primes whose gap is at most 6. This naive approach does not quite work directly, but by carefully optimising the argument (for instance, replacing the condition $n, n+2, n+6 \in P_2$ with something more like $n(n+2)(n+6) \in P_6$), Goldston, Yildirim, and Pintz[GoYiPi2008] were recently able to show unconditionally that prime gaps in $[N,2N]$ could be as small

as $o(\log N)$, and could in fact be as small as 16 infinitely often if one assumes the Elliot-Halberstam conjecture[ElHa1969].

In a somewhat similar spirit, my result with Ben Green[GrTa2008] establishing that the primes contain arbitrarily long progressions proceeds by first using sieve theory methods to show that the almost primes (or more precisely, a suitable weight function ν concentrated near the almost primes) are very pseudorandomly distributed, in the sense that several self-correlations of ν can be computed and agree closely with what one would have predicted if the almost primes were distributed randomly (after accounting for some irregularities caused by small moduli). Because of the parity problem, the primes themselves are not known to be as pseudorandomly distributed as the almost primes; however, the prime number theorem does at least tell us that the primes have a positive relative density in the almost primes. The main task is then to show that any set of positive relative density in a sufficiently pseudorandom set contains arithmetic progressions of any specified length; this combinatorial result (a "relative Szemerédi theorem") plays roughly the same role that the pigeonhole principle did in the work of Goldston-Yildirim-Pintz. (On the other hand, the relative Szemerédi theorem works even for arbitrarily low density, whereas the pigeonhole principle does not; because of this, our sieve theory analysis is far less delicate than that in Goldston-Yildirim-Pintz.)

It is probably premature, with our current understanding, to try to find a systematic way to get around the parity problem in general, but it seems likely that we will be able to find some further ways to get around the parity problem in special cases, and perhaps once we have assembled enough of these special cases, it will become clearer what to do in general.

1.10.3 Notes

This article was originally posted on Jun 5, 2007 at

`terrytao.wordpress.com/2007/06/05`

Emmanuel Kowalski pointed out the need to distinguish between almost primes n which had only $O(1)$ factors, and almost primes n which were coprime to all numbers between 2 and n^c for some $0 < c < 1$. The latter type of almost prime (which is sparser) are the ones which are of importance in sieve theory; their density is similar to that of the primes (i.e. comparable to $1/\log n$) whereas the former type of almost prime has density $(\log\log n)^{O(1)}/\log n$ instead.

Felipe Voloch noted that by using Galois theory techniques one can sometimes convert upper bounds in prime number estimates to lower bounds, though this method does not seem to combine well with sieve theory methods.

Emmanuel Kowalski, Jordan Ellenberg, and Keith Conrad had some interesting discussions on the role (or lack thereof) of adeles in sieve theory, and on how to define the correct analogue of a "box" to sieve over in other number fields.

Ben Green pointed out the relationship between elementary sieving methods and the "W-trick" used in our papers on arithmetic progressions of primes.

1.11 Deterministic RIP matrices

This problem in compressed sensing is an example of a *derandomisation problem*: take an object which, currently, can only be constructed efficiently by a probabilistic method, and figure out a deterministic construction of comparable strength and practicality. (For a general comparison of probabilistic and deterministic algorithms, see [Wi2008].)

I will define exactly what RIP matrices (the RIP stands for *restricted isometry property*) are later in this post. For now, let us just say that they are a generalisation of (rectangular) orthogonal matrices, in which the columns are locally almost orthogonal rather than globally perfectly orthogonal. Because of this, it turns out that one can pack significantly more columns into a RIP matrix than an orthogonal matrix, while still capturing many of the desirable features of orthogonal matrices, such as stable and computable invertibility (as long as one restricts attention to *sparse* or *compressible* vectors). Thus RIP matrices can "squash" sparse vectors from high-dimensional space into a low-dimensional while still being able to reconstruct those vectors; this property underlies many of the recent results on compressed sensing today.

There are several constructions of RIP matrices known today (e.g. random normalised Gaussian matrices, random normalised Bernoulli matrices, or random normalised minors of a discrete Fourier transform matrix) but (if one wants the sparsity parameter to be large) they are all *probabilistic* in nature; in particular, these constructions are not 100% guaranteed to actually produce a RIP matrix, although in many cases the failure rate can be proven to be exponentially small in the size of the matrix. Furthermore, there is no fast (e.g. sub-exponential time) algorithm known to test whether any given matrix is RIP or not. The failure rate is small enough that this is not a problem for most applications (especially since many compressed sensing applications are for environments which are already expected to be noisy in many other ways), but is slightly dissatisfying from a theoretical point of view. One is thus interested in finding a deterministic construction which can locate RIP matrices in a reasonably rapid manner. (One could of course simply search through all matrices in a given class and test each one for the RIP property, but this is an exponential-time algorithm, and thus totally impractical for applications.) In analogy with error-correcting codes, it may be that algebraic or number-theoretic constructions may hold the most promise for such deterministic RIP matrices (possibly assuming some unproven conjectures on exponential sums); this has already been accomplished by de Vore[dV2008] for RIP matrices with small sparsity parameter.

Before we define RIP matrices explicitly, let us first recall what an (rectangular) orthogonal matrix is. We will view an $m \times n$ matrix (m rows and n columns) as a collection $v_1, \ldots, v_n$ of column vectors in the (complex) vector space $\mathbf{C}^m$, or equivalently as a means of linearly transformed an n-dimensional vector $(a_1, \ldots, a_n)$ as an m-dimensional vector $\sum_{j=1}^n a_j v_j$. I will call such a matrix orthogonal if these column vectors are orthonormal, i.e. they all have unit length $\|v_j\| = 1$ and are orthogonal to each other: $\langle v_j, v_k \rangle = 0$ whenever $j \neq k$.

Orthonormal vectors have several pleasant properties. One of them is Pythagoras'

theorem

$$\|\sum_{j=1}^{n} a_j v_j\|^2 = \sum_{j=1}^{n} |a_j|^2 \tag{1.13}$$

valid for all complex numbers $a_1,\ldots,a_n$. In other words, the linear encoding $(a_1,\ldots,a_n) \mapsto \sum_{j=1}^n a_j v_j$ is an isometry. This implies that such an encoding can be inverted in a stable manner: given the encoded vector $w = \sum_{j=1}^n a_j v_j$ one can uniquely recover the original coefficients $a_1,\ldots,a_n$, and furthermore that small changes in w will not cause large fluctuations in $a_1,\ldots,a_n$. Indeed, one can reconstruct the coefficients a_i quickly and explicitly by the formula

$$a_j = \langle w, v_j \rangle. \tag{1.14}$$

One would like to make n as large as possible, and m as small as possible, so that one can transform as high-dimensional vectors as possible using only as low-dimensional space as possible to store the transformed vectors. There is however a basic obstruction to this, which is that an orthogonal matrix can only exist when $n \leq m$; for if n is larger than m, then there are too many vectors $v_1,\ldots,v_n$ to remain linearly independent in $\mathbf{C}^m$, and one must have a non-trivial linear independence

$$a_1 v_1 + \ldots + a_n v_n = 0$$

for some $(a_1,\ldots,a_n) \neq (0,\ldots,0)$, which is inconsistent with (1.13).

One can try to circumvent this restriction by weakening the condition (1.13) to (say)

$$0.9 \sum_{j=1}^{n} |a_j|^2 \leq \|\sum_{j=1}^{n} a_j v_j\|^2 \leq 1.1 \sum_{j=1}^{n} |a_j|^2 \tag{1.15}$$

for all complex numbers $a_1,\ldots,a_n$. (The constants 0.9 and 1.1 are not terribly important for this discussion.) Thus we only require that Pythagoras' theorem hold *approximately* rather than exactly; this is equivalent to requiring that the transpose of this matrix forms a frame. (In harmonic analysis, one would say that the vectors $v_1,\ldots,v_n$ are *almost orthogonal* rather than perfectly orthogonal.) This enlarges the class of matrices that one can consider, but unfortunately does not remove the condition $n \leq m$, since the linear dependence argument which showed that $n > m$ was incompatible with (1.13), also shows that $n > m$ is incompatible with (1.15).

It turns out, though, that one can pack more than m vectors into $\mathbf{C}^m$ if one *localises* the almost orthogonality condition (1.15) so that it only holds for *sparse* sets of coefficients $a_1,\ldots,a_n$. Specifically, we fix a parameter S (less than m), and say that the matrix $(v_1,\ldots,v_n)$ obeys the *RIP with sparsity S* if one has the almost orthogonality condition (1.15) for any set of coefficients $(a_1,\ldots,a_n)$, such that at most S of the a_j are non-zero. [The RIP is also known as the *Uniform Uncertainty Principle* (UUP) in the literature, particularly with regard to Fourier-type vectors; see Section 3.2.] In other words, we only assume that any S of the n vectors $v_1,\ldots,v_n$ are almost orthogonal at one time. (It is important here that we require almost orthogonality rather than perfect orthogonality, since as soon as a set of vectors are pairwise perfectly orthogonal, they are of course jointly perfectly orthogonal. In contrast, the constants 0.9 and 1.1 in the RIP condition will deteriorate as S increases, so that local almost orthogonality does not imply global almost orthogonality.) The RIP property is more powerful

(and hence more useful) when S is large; in particular one would like to approach the "information-theoretic limit" when S is comparable in magnitude to m.

Roughly speaking, a set of vectors $(v_1,\ldots,v_n)$ which obey the RIP are "just as good" as an orthonormal set of vectors, so long as one doesn't look at more than S of these vectors at a time. For instance, one can easily show that the map $(a_1,\ldots,a_n) \mapsto \sum_j a_j v_j$ is still injective as long as one restricts attention to input vectors which are $S/2$-sparse or better (i.e. at most $S/2$ of the coefficients are allowed to be non-zero). This still leaves the question of how to efficiently recover the sparse coefficients $(a_1,\ldots,a_n)$ from the transformed vector $w = \sum_j a_j v_j$. The algorithm (1.14) is no longer accurate; however if the coefficients are just a little bit sparser than $S/2$ (e.g. $S/3$ will do) then one can instead use the algorithm of *basis pursuit* to recover the coefficients $(a_1,\ldots,a_n)$ perfectly. Namely, it turns out[CaTa2005] that among all the possible representations $w = \sum_j b_j v_j$ of w, the one which minimises the l^1 norm $\sum_j |b_j|$ will be the one which matches the $S/3$-sparse representation $\sum_j a_j v_j$ exactly. (This has an interesting geometric interpretation: if we normalise all the vectors v_j to have unit length, then this result says that the simplest (sparsest) way to get from 0 to w by moving in the directions $v_1,\ldots,v_n$ is also the shortest way to get there.) There are also some related results regarding coefficients $(a_1,\ldots,a_n)$ which are merely compressible instead of sparse, but these are a bit more technical; see my paper with Emmanuel Candes[CaTa2007] for details.

It turns out that RIP matrices can have many more columns than rows; indeed, as shown in [Do2006], [CaTa2006], n can be as large as $m\exp(cm/S)$ for some absolute constant c ¿ 0. (Subsequent proofs also appeared in [CaRuTaVe2005], [BaDadVWa2008].) The construction is in fact very easy; one simply selects the vectors $v_1,\ldots,v_n$ *randomly*, either as random unit vectors or as random normalised Gaussian vectors (so all coefficients of each v_i are independent Gaussians with mean zero and variance $1/m$). The point is that in a high-dimensional space such as $\mathbf{C}^m$, any two randomly selected vectors are very likely to be almost orthogonal to each other; for instance, it is an easy computation that the dot product between two random normalised Gaussian vectors has a variance of only $O(1/m)$, even though the vectors themselves have a magnitude very close to 1. Note though that control of these dot products is really only enough to obtain the RIP for relatively small S, e.g. $S = O(\sqrt{(}m))$. For large S, one needs slightly more advanced tools, such as large deviation bounds on the singular values of rectangular Gaussian matrices (which are closely related to the Johnson-Lindenstrauss lemma[JoLi1984]).

The results for small sparsity S are relatively easy to duplicate by deterministic means. In particular, the paper of de Vore[dV2008] mentioned earlier uses a polynomial construction to obtain RIP matrices with S close to $\sqrt{m}$, and n equal to an arbitrarily large power of m, essentially by ensuring that all the column vectors have a low inner product with each other (of magnitude roughly $\sqrt{1/m}$ or so, matching what the random construction gives, and almost certainly best possible). But to get to larger values of S (and in particular, to situations in which S is comparable to m) may require a more refined calculation (possibly involving higher moments of the Gramian matrix, as was done in [CaRoTa2006] in the random case). Alternatively, one may rely on conjecture rather than rigorous results; for instance, it could well be that the matrices of de Vore satisfy the RIP for far larger sparsities S than are rigorously proven in that

paper.

An alternate approach, and one of interest in its own right, is to work on improving the time it takes to verify that a given matrix (possibly one of a special form) obeys the RIP. The brute-force approach of checking the singular values of every set of S column vectors requires a run time comparable to $\binom{n}{S}$ or worse, which is quite poor. (A variant approach has recently been proposed by Sharon, Wright, and Ma[ShWrMa2008] but has similar run time costs.) But perhaps there exist specially structured matrices for which the RIP is easier to verify, and for which it is still likely that the RIP holds. This would give a probablistic algorithm for producing rigorously certified RIP matrices with a reasonable average-case run time.

1.11.1 Notes

This article was originally posted on Jul 2, 2007 at

`terrytao.wordpress.com/2007/07/02`

Thanks to Ajay Bangla for corrections.

Igor Carron asked what happened if one relaxed the RIP condition so that one had restricted isometry for *most* collections of sparse columns rather than *all*. It may be easier to construct matrices with this weaker property, though these matrices seem to be somewhat less useful for applications and for rigorous theoretical results.

1.12 The nonlinear Carleson conjecture

In this article I will describe the "nonlinear Carleson theorem" conjecture, which is still one of my favourite open problems, being an excellent benchmark for measuring progress in the (still nascent) field of "nonlinear Fourier analysis", while also being of interest in its own right in scattering and spectral theory.

My starting point will be the one-dimensional time-independent Schrödinger equation

$$-u_{xx}(k,x) + V(x)u(k,x) = k^2 u(k,x) \tag{1.16}$$

where $V : \mathbf{R} \to \mathbf{R}$ is a given potential function, $k \in \mathbf{R}$ is a frequency parameter, and $u : \mathbf{R} \times \mathbf{R} \to \mathbf{C}$ is the wave function. This equation (after reinstating constants such as Planck's constant $\hbar$, which we have normalised away) describes the instantaneous state of a quantum particle with energy k^2 in the presence of the potential V. To avoid technicalities let us assume that V is smooth and compactly supported (say in the interval $[-R,R]$) for now, though the eventual conjecture will concern potentials V that are merely square-integrable.

For each fixed frequency k, the equation (1.16) is a linear homogeneous second order ODE, and so has a two-dimensional space of solutions. In the free case $V = 0$, the solution space is given by

$$u(k,x) = \alpha(k)e^{ikx} + \beta(k)e^{-ikx} \tag{1.17}$$

where $\alpha(k)$ and $\beta(k)$ are arbitrary complex numbers; physically, these numbers represent the amplitudes of the rightward and leftward propagating components of the solution respectively.

Now suppose that V is non-zero, but is still compactly supported on an interval $[-R,+R]$. Then for a fixed frequency k, a solution to (1.16) will still behave like (1.17) in the regions $x > R$ and $x < R$, where the potential vanishes; however, the amplitudes on either side of the potential may be different. Thus we would have

$$u(k,x) = \alpha_+(k)e^{ikx} + \beta_+(k)e^{-ikx}$$

for $x > R$ and

$$u(k,x) = \alpha_-(k)e^{ikx} + \beta_-(k)e^{-ikx}$$

for $x < -R$. Since there is only a two-dimensional linear space of solutions, the four complex numbers $\alpha_-(k), \beta_-(k), \alpha_+(k), \beta_+(k)$ must be related to each other by a linear relationship of the form

$$\begin{pmatrix} \alpha_+(k) \\ \beta_+(k) \end{pmatrix} = \widehat{V}(k) \begin{pmatrix} \alpha_-(k) \\ \beta_-(k) \end{pmatrix}$$

where $\widehat{V}(k)$ is a 2×2 matrix depending on V and k, known as the *scattering matrix* of V at frequency k. (We choose this notation to deliberately invoke a resemblance to the Fourier transform $\hat{V}(k) := \int_{-\infty}^{\infty} V(x)e^{-2ikx}\,dx$ of V; more on this later.) Physically, this matrix determines how much of an incoming wave at frequency k gets reflected by the potential, and how much gets transmitted.

What can we say about the matrix $\widehat{V}(k)$? By using the Wronskian of two solutions to (1.16) (or by viewing (1.16) as a Hamiltonian flow in phase space) we can show that $\widehat{V}(k)$ must have determinant 1. Also, by using the observation that the solution space to (1.16) is closed under complex conjugation $u(k,x) \mapsto \overline{u(k,x)}$, one sees that each coefficient of the matrix $\widehat{V}(k)$ is the complex conjugate of the diagonally opposite coefficient. Combining the two, we see that $\widehat{V}(k)$ takes values in the Lie group

$$SU(1,1) := \{\begin{pmatrix} a & \overline{b} \\ b & \overline{a} \end{pmatrix} : a,b \in \mathbf{C}, |a|^2 - |b|^2 = 1\}$$

(which, incidentally, is isomorphic to $SL_2(\mathbf{R})$), thus we have

$$\widehat{V}(k) = \begin{pmatrix} a(k) & \overline{b(k)} \\ b(k) & \overline{a(k)} \end{pmatrix}$$

for some functions $a : \mathbf{R} \to \mathbf{C}$ and $b : \mathbf{R} \to \mathbf{C}$ obeying the constraint $|a(k)|^2 - |b(k)|^2 = 1$. (The functions $\frac{1}{a(k)}$ and $\frac{b(k)}{a(k)}$ are sometimes known as the *transmission coefficient* and *reflection coefficient* respectively; note that they square-sum to 1, a fact related to the law of conservation of energy.) These coefficients evolve in a beautifully simple manner if V evolves via the Korteweg-de Vries (KdV) equation $V_t + V_{xxx} = 6VV_x$ (indeed, one has $\partial_t a = 0$ and $\partial_t b = 8ik^3 b$), being part of the fascinating subject of completely integrable systems, but that is a long story which we will not discuss here. This connection does however provide one important source of motivation for studying the *scattering transform* $V \mapsto \widehat{V}$ and its inverse.

What are the values of the coefficients $a(k), b(k)$? In the free case $V = 0$, one has $a(k) = 1$ and $b(k) = 0$. When V is non-zero but very small, one can linearise in V (discarding all terms of order $O(V^2)$ or higher), and obtain the approximation

$$a(k) \approx 1 - \frac{i}{2k}\int_{-\infty}^{\infty} V; \quad b(k) \approx \frac{-i}{2k}\hat{V}(k)$$

known as the *Born approximation*; this helps explain why we think of $\widehat{V}(k)$ as a nonlinear variant of the Fourier transform. A slightly more precise approximation, known as the *WKB approximation*, is

$$a(k) \approx e^{-\frac{i}{2k}\int_{-\infty}^{\infty} V}; \quad b(k) \approx \frac{-i}{2k} e^{-\frac{i}{2k}\int_{-\infty}^{\infty} V} \int_{-\infty}^{\infty} V(x) e^{-2ikx + \frac{i}{k}\int_{-\infty}^{x} V}\, dx.$$

(One can avoid the additional technicalities caused by the WKB phase correction by working with the Dirac equation instead of the Schrödinger; this formulation is in fact cleaner in many respects, but we shall stick with the more traditional Schrödinger formulation here. More generally, one can consider analogous scattering transforms for AKNS systems.) One can in fact expand $a(k)$ and $b(k)$ as a formal power series of multilinear integrals in V (distorted slightly by the WKB phase correction $e^{\frac{i}{k}\int_{-\infty}^{x} V}$). It is relatively easy to show that this multilinear series is absolutely convergent for every

k when the potential V is absolutely integrable (this is the nonlinear analogue to the obvious fact that the Fourier integral $\hat{V}(k) = \int_{-\infty}^{\infty} V(k)e^{-2ikx}$ is absolutely convergent when V is absolutely integrable; it can also be deduced without recourse to multilinear series by using Levinson's theorem.) If V is not absolutely integrable, but instead lies in $L^p(\mathbf{R})$ for some $p > 1$, then the series can diverge for some k; this fact is closely related to a classic result of Wigner and von Neumann that the Schrdinger operator can contain embedded pure point spectrum. However, Christ and Kiselev[ChKi2001] showed that the series is absolutely convergent for almost every k in the case $1 < p < 2$ (this is a non-linear version of the Hausdorff-Young inequality). In fact they proved a stronger statement, namely that for almost every k, the eigenfunctions $x \mapsto u(k,x)$ are bounded (and converge asymptotically to plane waves $\alpha_{\pm}(k)e^{ikx} + \beta_{\pm}(k)e^{-ikx}$ as $x \to \infty$). There is an analogue of the Born and WKB approximations for these eigenfunctions, which shows that the Christ-Kiselev result is the nonlinear analogue of a classical result of Menshov, Paley and Zygmund showing the conditional convergence of the Fourier integral $\int_{-\infty}^{\infty} V(x)e^{-2ikx}\,dx$ for almost every k when $V \in L^p(\mathbf{R})$ for some $1 < p < 2$.

The analogue of the Menshov-Paley-Zygmund theorem at the endpoint $p = 2$ is the celebrated theorem of Carleson[Ca1966] on almost everywhere convergence of Fourier series of L^2 functions. (The claim fails for $p > 2$, as can be seen by investigating random Fourier series, though I don't recall the reference for this fact.) The nonlinear version of this would assert that for square-integrable potentials V, the eigenfunctions $x \mapsto u(k,x)$ are bounded for almost every k. This is the nonlinear Carleson theorem conjecture. Unfortunately, it cannot be established by multilinear series, because of a divergence in the trilinear term of the expansion[MuTaTh2003]; but other methods may succeed instead. For instance, the weaker statement that the coefficients $a(k)$ and $b(k)$ (defined by density) are well defined and finite almost everywhere for square-integrable V (which is a nonlinear analogue of Plancherel's theorem that the Fourier transform can be defined by density on $L^2(\mathbf{R})$) was essentially established by Deift and Killip [DeKi1999], using a trace formula (a nonlinear analogue to Plancherel's formula). Also, the "dyadic" or "function field" model (cf. Section 2.6) of the conjecture is known[MuTaTh2003b], by a modification of Carleson's original argument. But the general case still seems to require more tools; for instance, we still do not have a good nonlinear Littlewood-Paley theory (except in the dyadic case), which is preventing time-frequency type arguments from being extended directly to the nonlinear setting.

1.12.1 Notes

This article was originally posted on Dec 17, 2007 at

terrytao.wordpress.com/2007/12/17

Chapter 2

Expository articles

2.1 Quantum mechanics and Tomb Raider

Quantum mechanics has a number of weird consequences, but in this article I will focus on three (inter-related) ones:

- Objects can behave both like particles (with definite position and a continuum of states) and waves (with indefinite position and (in confined situations) quantised states);
- The equations that govern quantum mechanics are deterministic, but the standard interpretation of the solutions (the *Copenhagen interpretation*) of these equations is probabilistic; and
- If instead one applies the laws of quantum mechanics literally at the macroscopic scale (via the *relative state interpretation*, more popularly known as the *many worlds intepretation*), then the universe itself must split into the superposition of many distinct "worlds".

What I will attempt to do here is to use the familiar concept of a computer game as a classical conceptual model with which to capture these non-classical phenomena. The exact choice of game is not terribly important, but let us pick *Tomb Raider* - a popular game from about ten years ago, in which the heroine, Lara Croft, explores various tombs and dungeons, solving puzzles and dodging traps, in order to achieve some objective. It is quite common for Lara to die in the game, for instance by failing to evade one of the traps. (I should warn that this analogy will be rather violent on certain computer-generated characters.)

The thing about such games is that there is an "internal universe", in which Lara interacts with other game elements, and occasionally is killed by them, and an "external universe", where the computer or console running the game, together with the human who is playing the game, resides. While the game is running, these two universes run more or less in parallel; but there are certain operations, notably the "save game" and "restore game" features, which disrupt this relationship. These operations are utterly mundane to people like us who reside in the external universe, but it is an interesting thought experiment to view them from the perspective of someone like Lara, in the internal universe. (I will eventually try to connect this with quantum mechanics, but please be patient for now.) Of course, for this we will need to presume that the Tomb Raider game is so advanced that Lara has levels of self-awareness and artificial intelligence which are comparable to our own. In particular, we will imagine that Lara is independent enough to play the game without direct intervention from the player, whose role shall be largely confined to that of saving, loading, and observing the game.

Imagine first that Lara is about to navigate a tricky rolling boulder puzzle, when she hears a distant rumbling sound - the sound of her player saving her game to disk. From the perspective of the player, we suppose that what happens next is the following: Lara navigates the boulder puzzle but fails, being killed in the process; then the player restores the game from the save point and then Lara successfully makes it through the boulder puzzle.

Now, how does the situation look from Lara's point of view? At the save point, Lara's reality diverges into a superposition of two non-interacting paths, one in which she dies in the boulder puzzle, and one in which she lives. (Yes, just like that cat.) Her future becomes indeterministic. If she had consulted with an infinitely prescient oracle before reaching the save point as to whether she would survive the boulder puzzle, the only truthful answer this oracle could give is "50% yes, and 50% no".

This simple example shows that the *internal* game universe can become indeterministic, even though the *external* one might be utterly deterministic. However, this example does not fully capture the weirdness of quantum mechanics, because in each one of the two alternate states Lara could find herself in (surviving the puzzle or being killed by it), she does not experience any effects from the other state at all, and could reasonably assume that she lives in a classical, deterministic universe.

So, let's make the game a bit more interesting. Let us assume that every time Lara dies, she leaves behind a corpse in that location for future incarnations of Lara to encounter. Then Lara will start noticing the following phenomenon (assuming she survives at all): whenever she navigates any particularly tricky puzzle, she usually encounters a number of corpses which look uncannily like herself. This disturbing phenomenon is difficult to explain to Lara using a purely classical deterministic model of reality; the simplest (and truest) explanation that one can give her is a "many-worlds" interpretation of reality, and that the various possible states of Lara's existence have some partial interaction with each other. Another valid (and largely equivalent) explanation would be that every time Lara passes a save point to navigate some tricky puzzle, Lara's "particle-like" existence splits into a "wave-like" superposition of Lara-states, which then evolves in a complicated way until the puzzle is resolved one way or the other, at which point Lara's wave function "collapses" in a non-deterministic fashion back to a particle-like state (which is either entirely alive or entirely dead).

Now, in the real world, it is only microscopic objects such as electrons which seem to exhibit this quantum behaviour; macroscopic objects, such as you and I, do not directly experience the kind of phenomena that Lara does, and we cannot interview individual electrons to find out their stories either. Nevertheless, by studying the statistical behaviour of large numbers of microscopic objects we can indirectly infer their quantum nature via experiment and theoretical reasoning. Let us again use the Tomb Raider analogy to illustrate this. Suppose now that Tomb Raider does not only have Lara as the main heroine, but in fact has a large number of playable characters, who explore a large number deadly tombs, often with fatal effect (and thus leading to multiple game restores). Let us suppose that inside this game universe there is also a scientist (let's call her Jacqueline) who studies the behaviour of these adventurers going through the tombs. However, Jacqueline does not experience the tombs directly, nor does she actually communicate with any of these adventurers. Each tomb is explored by only one adventurer; regardless of whether she lives or dies, the tomb is considered "used up".

Jacqueline observes several types of trapped tombs in her world, and gathers data as to how likely an adventurer is to survive any given type of tomb. She learns that each type of tomb has a fixed survival rate - e.g. she may observe that a tomb of type A has a 20% survival rate, whilst a tomb of type B has a 50% survival rate - but that it seems impossible to predict with any certainty whether any given adventurer will

survive any given type of tomb. So far, this is something which could be explained classically; each tomb may have a certain number of lethal traps in them, and whether an adventurer survives these traps or not may entirely be due to random chance or other "hidden variables".

But then Jacqueline encounters a mysterious *quantisation* phenomenon: the survival rate for various tombs are always one of the numbers $100\%, 50\%, 33.3\ldots\%, 25\%, 20\%, \ldots$; in other words, the "frequency" of success for a tomb is always of the form $1/n$ for some integer n. This phenomenon would be difficult to explain in a classical universe, since the effects of random chance should be able to produce a continuum of survival probabilities.

Here's what is going on. In order for Lara (or any other adventurer) to survive a tomb of a given type, she needs to stack together a certain number of corpses together to reach a certain switch; if she cannot attain that level of "constructive interference" to reach that switch, she dies. The type of tomb determines exactly how many corpses are needed; for instance, a tomb of type A might requires four corpses to be stacked together. Then the player who is playing Lara will have to let her die four times before she can successfully get through the tomb; and so from her perspective, Lara's chances of survival are only 20%. In each possible state of the game universe, there is only one Lara which goes into the tomb, who either lives or dies; but her survival rate here is what it is because of her interaction with other states of Lara (which Jacqueline cannot see directly, as she does not actually enter the tomb).

In our own reality, a familiar example of this type of quantum effect is the fact that each atom (e.g. sodium or neon) can only emit certain wavelengths of light (which end up being quantised somewhat analogously to the survival probabilities above); for instance, sodium only emits yellow light, neon emits blue, and so forth. The electrons in such atoms, in order to emit such light, are in some sense clambering over skeletons of themselves to do so; the more commonly given explanation is that the electron is behaving like a wave within the confines of an atom, and thus can only oscillate at certain frequencies (similarly to how a plucked string of a musical instrument can only exhibit a certain set of wavelengths, which coincidentally are also proportional to $1/n$ for integer n). Mathematically, this "quantisation" of frequency can be computed using the bound states of a Schrödinger operator with potential. [I will not attempt to stretch the Tomb Raider analogy so far as to try to model the Schrödinger equation! In particular, the complex phase of the wave function - which is a fundamental feature of quantum mechanics - is not easy at all to motivate in a classical setting.]

Now let's use the Tomb Raider analogy to explain why microscopic objects (such as electrons) experience quantum effects, but macroscopic ones (or even mesoscopic ones, such as large molecues) seemingly do not. Let's assume that Tomb Raider is now a two-player co-operative game, with two players playing two characters (let's call them Lara and Indiana) as they simultaneously explore different parts of their world. The players can choose to save the entire game, and then restore back to that point; this resets both Lara and Indiana back to the state they were in at that save point.

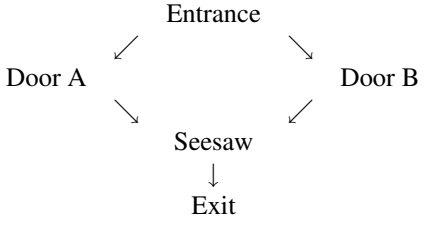
Now, this game still has the strange feature of corpses of Lara and Indiana from previous games appearing in later ones. However, we assume that Lara and Indiana are *entangled* in the following way: if Lara is in tomb A and Indiana is in tomb B, then Lara and Indiana can each encounter corpses of their respective former selves, but

only if *both* Lara *and* Indiana died in tombs A and B respectively in a single previous game. If in a previous game, Lara died in tomb A and Indiana died in tomb C, then this time round, Lara will not see any corpse (and of course, neither will Indiana). (This entanglement can be described a bit better by using *tensor products*; rather than saying that Lara died in A and Indiana died in B, one should instead think of Lara$\otimes$Indiana dying in $|A\rangle \otimes |B\rangle$, which is a state which is orthogonal to $|A\rangle \otimes |C\rangle$.) With this type of entanglement, one can see that there is going to be significantly less "quantum weirdness" going on; Lara and Indiana, adventuring separately but simultaneously, are going to encounter far fewer corpses of themselves than Lara adventuring alone would. And if there were many many adventurers entangled together exploring simultaneously, the quantum effects drop to virtually nothing, and things now look classical unless the adventurers are somehow organised to "resonate" in a special way (much as *Bose-Einstein condensates* operate in our own world).

The Tomb Raider analogy is admittedly not a perfect model for quantum mechanics. In the latter, the various possible basis states of a system interfere with each other via linear superposition of their complex phases, whereas in the former, the basis states interfere in an ordered nonlinear fashion, with the states associated to earlier games influencing the states of later games, but not vice versa. Another very important feature of quantum mechanics - namely, the ability to change the set of basis states used to decompose the full state of the system - does not have a counterpart in the Tomb Raider model. Nevertheless, this model is still sufficiently non-classical (when viewed from the internal universe) to construct some partial analogues of well-known quantum phenomena. We illustrate this with two more examples.

2.1.1 A two-slit experiment

The famous *two-slit experiment* involves a particle, such as an electron, being sent through a barrier with two slits. It can turn out that the particle can reach a certain destination beyond the barrier if one of the slits is covered up, but that this destination becomes inaccessible if both slits are opened up. A somewhat similar phenomenon can be simulated in the Tomb Raider universe described above, using the following kind of tomb:



Suppose that Door A and Door B are one-way; at the entrance to the tomb, Lara has to choose one of the two doors, and on doing so, is stuck on one end of the seesaw. Suppose that the seesaw is lethally trapped in such a way that one has to keep the seesaw balanced for, say, five minutes, otherwise the trap is set off, killing anyone on either side of the seesaw. Classically, it would be impossible for Lara to reach the exit, as she can only be on one side of the seesaw and so cannot maintain that seesaw's balance. But if she goes through once, say on side A, and then dies, then when the

game is restored, she can go in on side B and balance herself against the corpse from the previous game to defuse the trap. So she in fact has up to a 50% chance of survival here. (Actually, if she chooses a door randomly each time, and the player restores the game until she makes it through, the net chance of survival is only $2\ln 2 - 1 = 38.6\ldots\%$ - why?) On the other hand, if either of the doors is locked in advance, then her survival rate drops to 0%.

This does not have an easy classical explanation within the game universe, even with hidden variables, at least if you make the locality assumption that Lara can only go through one of the two one-way doors, and if you assume that the locks have no effect other than to stop Lara from choosing one of the doors.

2.1.2 Bell's inequality violation

Before we begin this example, let us recall some inequalities from classical probability. If A and B are two events, then we have the inclusion-exclusion identity

$$\mathbf{P}(A \vee B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \wedge B)$$

where $A \vee B$ is the event that at least one of A and B occur, and $A \wedge B$ is the event that A and B both occur. Since $\mathbf{P}(A \vee B)$ clearly cannot exceed 1, we conclude that

$$\mathbf{P}(A \wedge B) \geq \mathbf{P}(A) + \mathbf{P}(B) - 1. \tag{2.1}$$

Note that this inequality holds regardless of whether A and B are independent or not.

Iterating (2.1), we conclude that for any three events A, B, C, we have

$$\mathbf{P}(A \wedge B \wedge C) \geq \mathbf{P}(A) + \mathbf{P}(B) + \mathbf{P}(C) - 2. \tag{2.2}$$

Now let $l_1, l_2, i_1, i_2 \in \{0,1\}$ be four random variables (possibly dependent). Observe that if the event occurs that $l_1 = i_1$, $l_1 = i_2$, and $l_2 = i_1$, then we necessarily have $l_2 = i_2$. We conclude that

$$\mathbf{P}(l_2 = i_2) \geq \mathbf{P}(l_1 = i_1) + \mathbf{P}(l_1 = i_2) + \mathbf{P}(l_2 = i_1) - 2. \tag{2.3}$$

Again, we emphasise that this inequality must hold regardless of whether l_1, l_2, i_1, i_2 are independent or not. This inequality is a variant of the famous *Bell inequality*, and is known as the CHSH inequality.

We will now create a Tomb Raider experiment that shows that the internal game reality cannot be modeled by classical probability, at least if one insists that only one instance of the game universe exists. We will need two game characters, Lara and Indiana, who are exploring this map:

$$\begin{array}{ccccc} & & \text{Start} & & \\ & & \downarrow & & \\ \text{Gate } L & \leftarrow & \text{Tomb } A & \rightarrow & \text{Gate } I \end{array}$$

Gate L and Gate I both have two up-down switches which either character can manipulate into any of the four positions before trying to open the gate: up-up, up-down,

down-up, or down-down. However, the gates are trapped: only two of the positions allow the gate to be opened safely; the other two positions will ensure that the gate electrocutes whoever is trying to open it. Lara and Indiana know that the gates are anti-symmetric: if one flips both switches then that toggles whether the gate is safe or not (e.g. if down-up is safe, then up-down electrocutes). But they do not know exactly which combinations are safe.

Lara and Indiana (starting in the position "Start") desperately need to open both gates before a certain time limit, but do not know which of the combinations are safe. They have just enough time for Lara to go to Gate L through Tomb A, and for Indiana to go to Gate I through Tomb A, but there is not enough time for Lara to communicate to Indiana what she sees at Gate L, or conversely.

They believe (inaccurately, as it turns out) that inside Tomb A, there is inscribed a combination (of one of the four positions) which will safely open both gates. Their plan is to jointly go to Tomb A, find the combination, write that combination down on two pieces of paper (one for Lara, one for Indiana), and then Lara and Indiana will travel separately to Gate L and Gate I to try that combination to unlock both gates. At this point, the player saves the game and play continues repeatedly from this restore point. We re-emphasise that the player actually has no control over Lara and Indianas actions; they are independent AIs, following the plan described above.

Unfortunately for Lara and Indiana, the combination in Tomb A is simply a random combination - up-up, up-down, down-up, and down-down are each 25% likely to be found in Tomb A. In truth, the combinations to Gate L and Gate I have been set by Jacqueline. Jacqueline has set Gate L to one of the following two settings:

- Setting L_1: Gate L will open safely if the switches are up-up or up-down, but electrocutes if the switches are down-up or down-down
- Setting L_2: Gate L will open safely if the switches are up-up or down-up, but electrocutes if the switches are up-down or down-down.

Similarly, Jacqueline has set Gate I to one of the following two settings:

- Setting I_1: Gate I will open safely if the switches are up-up or up-down, but electrocutes if the switches are down-up or down-down.
- Setting I_2: Gate I will open safely if the switches are up-down or down-down, but electrocutes if the switches are down-up or up-up.

Note that these settings obey the anti-symmetry property mentioned earlier.

Jacqueline sets Gate L to setting L_a for some $a = 1,2$, and Gate I to setting I_b for some $b = 1,2$, and measures the probability p_{ab} of the event that Lara and Indiana both survive, or both die, thus computing four numbers $p_{11}, p_{12}, p_{21}, p_{22}$. (To do this, one would have to assume that the experiment can be repeated a large number of times, for instance by assuming that a large number of copies of these tombs and gates exist across the game universe, with a different pair of adventurers exploring each such copy.)

Jacqueline does not know the contents (or "hidden variables") of Tomb A, and does not know what Lara and Indiana's strategy is to open the gates (in particular, the strategy could be randomly chosen rather than deterministic). However, if she assumes that

communication between Lara and Indiana is local (thus Lara cannot transmit information about Gate L to Indiana at Gate I, or vice versa), and that the universe is classical (in particular, that no multiple copies of the universe exist), then she can deduce a certain theoretical inequality connecting the four numbers $p_{11}, p_{12}, p_{21}, p_{22}$. Indeed, she can write $p_{ab} = \mathbf{P}(l_a = i_b)$, where l_a is the random variable that equals 1 when Lara sets the switches of gate L to a position which is safe for L_a and 0 otherwise, and similarly i_b is the random variable that equals 1 when Indiana sets the switches of gate I to a position which is safe for I_b and 0 otherwise. Applying (2.3), we conclude that

$$p_{22} \geq p_{11} + p_{12} + p_{21} - 2 \tag{2.4}$$

regardless of what goes on in Tomb A, and regardless of what strategy Indiana and Lara execute.

We now show that in the actual Tomb Raider universe, the inequality (2.4) is violated - which proves to Jacqueline that her universe must either be non-local (with instantaneous information transmission) or non-classical (with the true state of the game universe being described as a superposition of more than one classical state).

First suppose that Gate L and Gate I are both set to setting 1, thus they open on up-* settings (i.e. up-up or up-down) and electrocute on down-*. If Lara and Indiana find an up-* pattern in Tomb A then they both survive. In some cases they may both be electrocuted, but only if they both hold down-* codes. If Lara and Indiana later encounter corpses of themselves clutching a down-* code, they are intelligent enough to apply the opposite of that code (overriding whatever false clue they got from Tomb A) and pass through safely. As the situation is totally symmetric we see in this case that $p = p_{11} = 1$.

Now suppose that Gate L and Gate I are both set to setting 2, thus Gate L is only safe for *-up and gate I is only safe for *-down. Then what happens every time the game is played is that exactly one of Lara or Indiana dies. Note that due to the entangled nature of the corpse mechanic, this means that Lara and Indiana never see any useful corpses which could save their lives. So in this case $p = p_{22} = 0$.

Now suppose that Gate L is in setting 1 and Gate I is in setting 2, or vice versa. Then what happens, if Indiana and Lara see no corpses, is that they have an independent 50% chance of survival, and thus a 50% chance of meeting the same fate. On the other hand, if Indiana and Lara see corpses (and the way the mechanic works, if one of them sees a corpse, the other does also), then they will use the more intelligent negation strategy to open both gates. Thus in these cases p_{12} or p_{21} is *strictly* greater than $1/2$.

Putting all these estimates together, we violate the inequality (2.4).

2.1.3 Notes

This article was originally posted on Feb 26, 2007 at

`terrytao.wordpress.com/2007/02/26`

It was derived from an interesting conversation I had several years ago with my friend Jason Newquist, on trying to find some intuitive analogies for the non-classical nature of quantum mechanics.

2.2 Compressed sensing and single-pixel cameras

I've had a number of people ask me exactly what "compressed sensing" means, and how a single pixel camera[BaKe2008] could possibly work (and how it might be advantageous over traditional cameras in certain circumstances). There is a large literature on the subject[BaKe2008b], but as the field is relatively recent, there does not yet appear to be a good non-technical introduction to the subject. So here's my stab at the topic, which should hopefully be accessible to a non-mathematical audience.

For sake of concreteness I'll primarily discuss the camera application, although compressed sensing is a more general measurement paradigm which is applicable to other contexts than imaging (e.g. astronomy, MRI, statistical selection, etc.), as I'll briefly remark upon at the end of this post.

The purpose of a camera is, of course, to record images. To simplify the discussion, let us think of an image as a rectangular array, e.g. a 1024×2048 array of pixels (thus there are 2 megapixels in all). To ignore the (minor) issue of colour, let us assume that we are just taking a black-and-white picture, so that each pixel is measured in grayscale as an integer (e.g. an 8-bit integer from 0 to 255, or a 16-bit integer from 0 to 65535) which signifies the intensity of each pixel.

Now, to oversimplify things quite a bit, a traditional digital camera would take one measurement of intensity for each of its pixels (so, about 2 million measurements in the above example), resulting in a relatively large image file (2MB if one uses 8-bit grayscale, or 4MB if one uses 16-bit grayscale). Mathematically, this file can be represented by a very high-dimensional vector of numbers (in this example, the dimension is about 2 million).

2.2.1 Traditional compression

Before I get to the new story of "compressed sensing", I have to first quickly review the somewhat older story of plain old "compression". (Those who already know how image compression works can skip forward to the next section.)

The 2-megapixel images described above can take up a lot of disk space on the camera (or on some computer where the images are later uploaded), and also take a non-trivial amount of time (and energy) to transfer from one medium to another. So, it is common practice to get the camera to *compress* the image, from an initial large size (e.g. 2MB) to a much smaller size (e.g. 200KB, which is 10% of the size). The thing is that while the space of *all* images has 2MB worth of "degrees of freedom" or "entropy", the space of all *interesting* images is much smaller, and can be stored using much less space, especially if one is willing to throw away some of the quality of the image. (Indeed, if one generates an image at random, one will almost certainly not get an interesting image; instead, one will just get random noise looking much like the static one can get on TV screens.)

How can one compress an image? There are many ways, some of which are rather technical, but let me try to give a non-technical (and slightly inaccurate) sketch of how it is done. It is quite typical for an image to have a large featureless component - for instance, in a landscape, up to half of the picture might be taken up by a monochromatic sky background. Suppose for instance that we locate a large square, say 100×100

pixels, which are all exactly the same colour - e.g. all white. Without compression, this square would take $10,000$ bytes to store (using 8-bit grayscale); however, instead, one can simply record the dimensions and location of the square, and note a single colour with which to paint the entire square; this will require only four or five bytes in all to record, leading to a massive space saving. Now in practice, we don't get such an impressive gain in compression, because even apparently featureless regions have some small colour variation between them. So, given a featureless square, what one can do is record the *average* colour of that square, and then subtract that average off from the image, leaving a small residual error. One can then locate more squares where the average colour is significant, and subtract those off as well. If one does this a couple times, eventually the only stuff left will be very small in magnitude (intensity), and not noticeable to the human eye. So we can throw away the rest of the image and record only the size, location, and intensity of the "significant" squares of the image. We can then reverse this process later and reconstruct a slightly lower-quality replica of the original image, which uses much less space.

Now, the above algorithm is not all that effective in practice, as it does not cope well with sharp transitions from one colour to another. It turns out to be better to work not with average colours in squares, but rather with average colour *imbalances* in squares - the extent to which the intensity on (say) the right half of the square is higher on average than the intensity on the left. One can formalise this by using the (two-dimensional) *Haar wavelet system*. It then turns out that one can work with "smoother" wavelet systems which are less susceptible to artefacts, but this is a technicality which we will not discuss here. But all of these systems lead to similar schemes: one represents the original image as a linear superposition of various "wavelets" (the analogues of the coloured squares in the preceding paragraph), stores all the significant (large magnitude) wavelet coefficients, and throws away (or "thresholds") all the rest. This type of "hard wavelet coefficient thresholding" compression algorithm is not nearly as sophisticated as the ones actually used in practice (for instance in the JPEG 2000 standard) but it is somewhat illustrative of the general principles in compression.

To summarise (and to oversimplify somewhat), the original 1024×2048 image may have two million degrees of freedom, and in particular if one wants to express this image in terms of wavelets then one would need thus need two million different wavelets in order to reconstruct all images perfectly. However, the typical *interesting* image is very *sparse* or *compressible* in the wavelet basis: perhaps only a hundred thousand of the wavelets already capture all the notable features of the image, with the remaining 1.9 million wavelets only contributing a very small amount of "random noise" which is largely invisible to most observers. (This is not always the case: heavily *textured* images - e.g. images containing hair, fur, etc. - are not particularly compressible in the wavelet basis, and pose a challenge for image compression algorithms. But that is another story.)

Now, if we (or the camera) knew in advance *which* hundred thousand of the 2 million wavelet coefficients are going to be the important ones, then the camera could just measure those coefficients and not even bother trying to measure the rest. (It is possible to measure a single coefficient by applying a suitable "filter" or "mask" to the image, and making a single intensity measurement to what comes out.) However, the camera does not know which of the coefficients are going to be the key ones, so it

must instead measure all 2 million pixels, convert the image to a wavelet basis, locate the hundred thousand dominant wavelet coefficients to keep, and throw away the rest. (This is of course only a caricature of how the image compression algorithm really works, but we will use it for sake of discussion.)

Now, of course, modern digital cameras work pretty well, and why should we try to improve on something which isn't obviously broken? Indeed, the above algorithm, in which one collects an enormous amount of data but only saves a fraction of it, works just fine for consumer photography. Furthermore, with data storage becoming quite cheap, it is now often feasible to use modern cameras to take many images with no compression whatsoever. Also, the computing power required to perform the compression is manageable, even if it does contribute to the notoriously battery-draining energy consumption level of these cameras. However, there are non-consumer imaging applications in which this type of data collection paradigm is infeasible, most notably in sensor networks. If one wants to collect data using thousands of sensors, which each need to stay *in situ* for long periods of time such as months, then it becomes necessary to make the sensors as cheap and as low-power as possible - which in particular rules out the use of devices which require heavy computer processing power at the sensor end (although - and this is important - we are still allowed the luxury of all the computer power that modern technology affords us at the *receiver* end, where all the data is collected and processed). For these types of applications, one needs a data collection paradigm which is as "dumb" as possible (and which is also robust with respect to, say, the loss of 10% of the sensors, or with respect to various types of noise or data corruption).

This is where *compressed sensing* comes in. The guiding philosophy is this: if one only needs $100,000$ components to recover most of the image, why not just take $100,000$ measurements instead of 2 million? (In practice, we would allow a safety margin, e.g. taking $300,000$ measurements, to allow for all sorts of issues, ranging from noise to aliasing to breakdown of the recovery algorithm.) In principle, this could lead to a power consumption saving of up to an order of magnitude, which may not mean much for consumer photography but can be of real importance in sensor networks.

But, as I said before, the camera does not know in advance which hundred thousand of the two million wavelet coefficients are the important ones that one needs to save. What if the camera selects a completely different set of $100,000$ (or $300,000$) wavelets, and thus loses all the interesting information in the image?

The solution to this problem is both simple and unintuitive. It is to make $300,000$ measurements which are *totally unrelated* to the wavelet basis - despite all that I have said above regarding how this is the best basis in which to view and compress images. In fact, the best types of measurements to make are (pseudo-)*random* measurements - generating, say, $300,000$ random "mask" images and measuring the extent to which the actual image resembles each of the masks. Now, these measurements (or "correlations") between the image and the masks are likely to be all very small, and very random. But - and this is the key point - each one of the 2 million possible wavelets which comprise the image will generate their own distinctive "signature" inside these random measurements, as they will correlate positively against some of the masks, negatively against others, and be uncorrelated with yet more masks. But (with overwhelming probability) each of the 2 million signatures will be distinct; furthermore, it turns out

that arbitrary linear combinations of up to a hundred thousand of these signatures will still be distinct from each other (from a linear algebra perspective, this is because two randomly chosen $100,000$-dimensional subspaces of a $300,000$ dimensional ambient space will be almost certainly disjoint from each other). Because of this, it is possible *in principle* to recover the image (or at least the $100,000$ most important components of the image) from these $300,000$ random measurements. In short, we are constructing a linear algebra analogue of a *hash function*.

There are however two technical problems with this approach. Firstly, there is the issue of noise: an image is not perfectly the sum of $100,000$ wavelet coefficients, but also has small contributions from the other 1.9 million coefficients also. These small contributions could conceivably disguise the contribution of the $100,000$ wavelet signatures as coming from a completely unrelated set of $100,000$ wavelet signatures; this is a type of "aliasing" problem. The second problem is how to use the $300,000$ measurements obtained to recover the image.

Let us focus on the latter problem first. If we knew which $100,000$ of the 2 million wavelets involved were, then we could use standard linear algebra methods (Gaussian elimination, least squares, etc.) to recover the signal. (Indeed, this is one of the great advantages of linear encodings - they are much easier to invert than nonlinear ones. Most hash functions are practically impossible to invert - which is an advantage in cryptography, but not in signal recovery.) However, as stated before, we don't know in advance which wavelets are involved. How can we find out? A naive least-squares approach gives horrible results which involve all 2 million coefficients and thus lead to very noisy and grainy images. One could perform a brute-force search instead, applying linear algebra once for each of the possible set of $100,000$ key coefficients, but this turns out to take an insanely impractical amount of time (there are roughly $10^{170,000}$ combinations to consider!) and in any case this type of brute-force search turns out to be NP-complete in general (it contains problems such as *subset-sum* as a special case). Fortunately, however, there are two much more feasible ways to recover the data:

- *Matching pursuit*: locate a wavelet whose signature seems to correlate with the data collected; remove all traces of that signature from the data; and repeat until we have totally "explained" the data collected in terms of wavelet signatures.

- *Basis pursuit* (or l^1 minimisation): Out of all the possible combinations of wavelets which would fit the data collected, find the one which is "sparsest" in the sense that the total sum of the magnitudes of all the coefficients is as small as possible. (It turns out that this particular minimisation tends to force most of the coefficients to vanish.) This type of minimisation can be computed in reasonable time via convex optimisation methods such as the *simplex method*.

Note that these image recovery algorithms do require a non-trivial (though not ridiculous) amount of computer processing power, but this is not a problem for applications such as sensor networks since this recovery is done on the receiver end (which has access to powerful computers) rather than the sensor end (which does not).

There are now rigorous results [CaRoTa2006, GiTr2008, CaTa2006, Do2006, RuVe2006] which show that these approaches can reconstruct the original signals perfectly or

almost-perfectly with very high probability of success, given various compressibility or sparsity hypotheses on the original image. The matching pursuit algorithm tends to be somewhat faster, but the basis pursuit algorithm seems to be more robust with respect to noise. Exploring the exact range of applicability of these methods is still a highly active current area of research. (Sadly, there does not seem to be an application to $P \neq NP$; the type of sparse recovery problems which are NP-complete are the total opposite (as far as the measurement matrix is concerned) with the type of sparse recovery problems which can be treated by the above methods.)

As compressed sensing is still a fairly new field (especially regarding the rigorous mathematical results), it is still a bit premature to expect developments here to appear in actual sensors. However, there are proof-of-concept prototypes already, most notably the single-pixel camera[BaKe2008] developed at Rice.

Finally, I should remark that compressed sensing, being an abstract mathematical idea rather than a specific concrete recipe, can be applied to many other types of contexts than just imaging. Some examples include:

- *Magnetic resonance imaging* (MRI). In medicine, MRI attempts to recover an image (in this case, the water density distribution in a human body) by taking a large but finite number of measurements (basically taking a discretised Radon transform (or x-ray transform) of the body), and then reprocessing the data. Because of the large number of measurements needed, the procedure is lengthy for the patient. Compressed sensing techniques can reduce the number of measurements required significantly, leading to faster imaging (possibly even to real-time imaging, i.e. MRI videos rather than static MRI). Furthermore, one can trade off the number of measurements against the quality of the image, so that by using the same number of measurements as one traditionally does, one may be able to get much finer scales of resolution.

- *Astronomy*. Many astronomical phenomena (e.g. pulsars) have various frequency oscillation behaviours which make them very sparse or compressible in the frequency domain. Compressed sensing techniques then allow one to measure these phenomena in the time domain (i.e. by recording telescope data) and being able to reconstruct the original signal accurately even from incomplete and noisy data (e.g. if weather, lack of telescope time, or simply the rotation of the earth prevents a complete time-series of data).

- *Linear coding*. Compressed sensing also gives a simple way for multiple transmitters to combine their output in an error-correcting way, so that even if a significant fraction of the output is lost or corrupted, the original transmission can still be recovered. For instance, one can transmit 1000 bits of information by encoding them using a random linear code into a stream of 3000 bits; and then it will turn out that even if, say, 300 of the bits (chosen adversarially) are then corrupted, the original message can be reconstructed perfectly with essentially no chance of error. The relationship with compressed sensing arises by viewing the corruption itself as the sparse signal (it is only concentrated on 300 of the 3000 bits).

Many of these applications are still only theoretical, but nevertheless the potential of these algorithms to impact so many types of measurement and signal processing is rather exciting. From a personal viewpoint, it is particularly satisfying to see work arising from pure mathematics (e.g. estimates on the determinant or singular values of Fourier minors) end up having potential application to the real world.

2.2.2 Notes

This article was originally posted on April 13, 2007 at

terrytao.wordpress.com/2007/04/13

For some explicit examples of how compressed sensing works on test images, see

www.acm.caltech.edu/l1magic/examples.html

2.3 Soft analysis, hard analysis, and the finite convergence principle

In the field of analysis, it is common to make a distinction between "hard", "quantitative", or "finitary" analysis on one hand, and "soft", "qualitative", or "infinitary" analysis on the other. "Hard analysis" is mostly concerned with finite quantities (e.g. the cardinality of finite sets, the measure of bounded sets, the value of convergent integrals, the norm of finite-dimensional vectors, etc.) and their *quantitative* properties (in particular, upper and lower bounds). "Soft analysis", on the other hand, tends to deal with more infinitary objects (e.g. sequences, measurable sets and functions, σ-algebras, Banach spaces, etc.) and their *qualitative* properties (convergence, boundedness, integrability, completeness, compactness, etc.). To put it more symbolically, hard analysis is the mathematics[1] of ε, N, $O()$, and $\leq$; soft analysis is the mathematics of 0, ∞, $\in$, and $\to$.

At first glance, the two types of analysis look very different; they deal with different types of objects, ask different types of questions, and seem to use different techniques in their proofs. They even use[2] different axioms of mathematics; the axiom of infinity, the axiom of choice, and the Dedekind completeness axiom for the real numbers are often invoked in soft analysis, but rarely in hard analysis. (As a consequence, there are occasionally some finitary results that can be proven easily by soft analysis but are in fact *impossible* to prove via hard analysis methods; the Paris-Harrington theorem[PaHa1977] provides a famous example.)

Because of all these differences, it is common for analysts to specialise in only one of the two types of analysis. For instance, as a general rule (and with notable exceptions), discrete mathematicians, computer scientists, real-variable harmonic analysts, and analytic number theorists tend to rely on "hard analysis" tools, whereas operator algebraists, abstract harmonic analysts, and ergodic theorists tend to rely on "soft analysis" tools[3].

[1] One can subdivide hard analysis into further subcategories by inspecting what kind of inequalities are used. There is "exact hard analysis" where one really uses $\leq$; "quasi-exact hard analysis" in which one is willing to lose absolute constants (and so one sees notation such as $O()$, $\lesssim$, or $\ll$); "logarithmically coarse hard analysis" in which one is willing to lose quantities such as $\log^{O(1)} N$ which are "logarithmic" in some key parameter N; and "polynomially coarse hard analysis" in which one is willing to lose quantities such as $N^{O(1)}$ which are polynomial in key parameters. Finally, there is *coarse analysis* in which one is willing to lose arbitrary functions of key parameters. The relationships between these flavours of hard analysis are interesting, but will have to wait to be discussed elsewhere.

[2] One can use these axioms to make finer distinctions, for instance "strongly finitary" analysis, in which one is not even willing to use real numbers, but instead only works with finite complexity numbers (e.g. rationals), and "strongly infinitary" analysis, in which one freely uses the axiom of choice (or related concepts such as ultrafilters, see Section 2.5). There are also hybrids between finitary and infinitary analysis, such as "pre-infinitary" analysis, in which one takes sequences of increasingly large or complex objects, and uses phrases such as "passing to a subsequence if necessary" frequently, but does not actually "jump to the limit"; we also have "pseudo-finitary" analysis, of which *non-standard analysis* is the most prominent example, in which infinitary methods are re-expressed using infinitesimals or other pseudo-finitary objects. See Section 2.5 for further discussion.

[3] Partial differential equations (PDE) is an interesting intermediate case in which *both* types of analysis are popular and useful, though many practitioners of PDE still prefer to primarily use just one of the two types. Another interesting transition occurs on the interface between point-set topology, which largely uses soft analysis, and metric geometry, which largely uses hard analysis. Also, the ineffective bounds which crop

2.3.1 Correspondences between hard and soft analysis

It is fairly well known that the *results* obtained by hard and soft analysis respectively can be connected to each other by various "correspondence principles" or "compactness principles". It is however my belief that the relationship between the two types of analysis is in fact much closer[4] than just this; in many cases, qualitative analysis can be viewed as a convenient abstraction of quantitative analysis, in which the precise dependencies between various finite quantities has been efficiently concealed from view by use of infinitary notation. Conversely, quantitative analysis can often be viewed as a more precise and detailed refinement of qualitative analysis. Furthermore, a method from hard analysis often has some analogue in soft analysis and vice versa, though the language and notation of the analogue may look completely different from that of the original. I therefore feel that it is often profitable for a practitioner of one type of analysis to learn about the other, as they both offer their own strengths, weaknesses, and intuition, and knowledge of one gives more insight[5] into the workings of the other. I wish to illustrate this point here using a simple but not terribly well known result, which I shall call the "finite convergence principle"[6]. It is the finitary analogue of an utterly trivial infinitary result - namely, that every bounded monotone sequence converges - but sometimes, a careful analysis of a trivial result can be surprisingly revealing, as I hope to demonstrate here.

Before I discuss this principle, let me first present an informal, incomplete, and inaccurate "dictionary" between soft and hard analysis, to try to give a rough idea of the (partial) correspondences between the two:

---

up from time to time in analytic number theory are a sort of hybrid of hard and soft analysis. Finally, there are examples of evolution of a field from soft analysis to hard (e.g. Banach space geometry) or vice versa (e.g. recent developments in extremal combinatorics, particularly in relation to the regularity lemma).

[4]There are rigorous results from proof theory, such as Herbrand's theorem[He1930], which can allow one to automatically convert certain types of qualitative arguments into quantitative ones. There has recently been some activity in applying the ideas from this and other *proof mining* results to various basic theorems in analysis; see [Ko2008].

[5]For instance, in my result with Ben Green[GrTa2008] establishing arbitrarily long arithmetic progressions of primes, the argument was (necessarily) finitary in nature, but it was absolutely essential for us to be aware of the infinitary arguments and intuition that had been developed in ergodic theory, as we had to adapt such arguments to the finitary setting in order to conclude our proof, and it would have far less evident how to discover such arguments if we were always restricted to looking at finitary settings. In general, it seems that infinitary methods are good for "long-range" mathematics, as by ignoring all quantitative issues one can move more rapidly to uncover qualitatively new kinds of results, whereas finitary methods are good for "short-range" mathematics, in which existing "soft" results are refined and understood much better via the process of making them increasingly sharp, precise, and quantitative. I feel therefore that these two methods are complementary, and are both important to deepening our understanding of mathematics as a whole.

[6]Thanks to Ben Green for suggesting this name; Jennifer Chayes has also suggested the "metastability principle".

Soft analysis	Hard analysis
x finite	x bounded (e.g. $x = O(1)$)
x vanishes	x small (e.g. $\|x\| \leq \varepsilon$)
x infinite	x large (e.g. $\|x\| \geq N$)
$x_n \to 0$	Quantitative decay bound (e.g. $x_n = O(n^{-c})$)
x_n is convergent	x_n is metastable (see below)
f uniformly continuous	Lipschitz or Hölder bound on f (e.g. $\|f(x) - f(y)\| = O(\|x-y\|)$
$f \in X$	$\|f\|_X = O(1)$
E compact	Metric entropy bound on E
E is Lebesgue measurable	E is (quantitatively) approximated by bounded complexity sets
V is generated by S	V is an algorithm initialised by S
u locally extremises F	u has no nearby competitor with significantly better value of F

One can draw two conclusions from this table:

- Soft analysis statements can often be stated both succinctly and rigorously, by using precisely defined and useful concepts (e.g. compactness, measurability, etc.). In hard analysis, one usually has to sacrifice one or the other: either one is rigorous but verbose (using lots of parameters such as ε, N, etc.), or succinct but "fuzzy" (using intuitive but vaguely defined concepts such as "size", "complexity", "nearby", etc.).

- A single concept in soft analysis can have multiple hard analysis counterparts. In particular, a "naive" translation of a statement in soft analysis into hard analysis may be incorrect. (In particular, one should not use the above table blindly to convert from one to the other.)

2.3.2 The finite convergence principle

Anyway, back to the finite convergence principle. The infinite convergence principle is well known, though perhaps not by this name, to anyone who has taken undergraduate analysis:

Proposition 2.1 (Infinite convergence principle)**.** *Every bounded monotone sequence x_n of real numbers is convergent.*

This basic principle - essentially equivalent to the Dedekind completeness axiom for the real numbers - is of course fundamental to many parts of infinitary analysis, most obviously in the theory of infinite sequences and series, but it also is implicit in just about any context in which one studies an "infinite" object (e.g. an infinite-dimensional vector space) by expressing it as a monotone limit of "finite" objects. It is undoubtedly an important tool in soft analysis. What, then, is its counterpart in hard analysis?

We will answer this question presently, but let us first make the infinite convergence principle a bit more quantitative. We may as well normalise the bounded sequence x_n to lie between 0 and 1. Expanding out the "epsilon-delta" definition of convergence, we obtain

Proposition 2.2 (Infinite convergence principle (again)). *If $0 \leq x_1 \leq x_2 \leq \ldots \leq 1$, then there exists a real number x such that for every $\varepsilon > 0$, there exists an N such that $|x_n - x| \leq \varepsilon$ for all $n \geq N$.*

There are quite a lot of quantifiers here. One can cut down the complexity a little bit by replacing the notion of a convergent sequence with that of a Cauchy sequence. This lets us eliminate the need for a limit x, which does not have an obvious finitary counterpart. This leaves us with

Proposition 2.3 (Infinite convergence principle (yet again)). *If $\varepsilon > 0$ and $0 \leq x_1 \leq x_2 \leq \ldots \leq 1$, there exists an N such that $|x_n - x_m| \leq \varepsilon$ for all $n, m \geq N$.*

Note now that one does not need the real number system to make this principle both meaningful and non-trivial; the principle already works quite well when restricted to the rationals. (Exercise: prove this principle for the rationals without constructing the real number system.) Informally speaking, this principle asserts that every bounded monotone sequence is eventually stable up to error ε.

Now let's try to find the finitary (quantitative) equivalent of this principle. The most naive thing to do is simply to replace the infinite sequence by a finite sequence, thus

Proposition 2.4 (Finite convergence principle (first attempt)). *If $\varepsilon > 0$ and $0 \leq x_1 \leq x_2 \leq \ldots \leq x_M \leq 1$, there exists an N such that $|x_n - x_m| \leq \varepsilon$ for all $N \leq n, m \leq M$.*

But this proposition is trivially true; one can simply set N equal to M (or any number larger than M). So one needs to strengthen the claim. What about making N be independent of M, and only dependent on ε?

Proposition 2.5 (Finite convergence principle (second attempt)). *If $\varepsilon > 0$ and $0 \leq x_1 \leq x_2 \leq \ldots \leq x_M \leq 1$, there exists an $N = N(\varepsilon)$ depending only on ε such that $|x_n - x_m| \leq \varepsilon$ for all $N \leq n, m \leq M$.*

But this is trivially false; consider for instance a sequence x_i which equals zero except at $i = M$, at which point we jump up to $x_M = 1$. We are not going to get the Cauchy property unless we set N to be as large as M... but we can't do that if we only want N to depend on ε.

So, is there anything non-trivial that one can say at all about finite bounded monotone sequences? Well, we have the pigeonhole principle:

Proposition 2.6 (Pigeonhole principle). *If $\varepsilon > 0$ and $0 \leq x_1 \leq x_2 \leq \ldots \leq x_M \leq 1$ is such that $M \geq 1/\varepsilon + 1$, there exists an $1 \leq N < M$ such that $|x_{N+1} - x_N| \leq \varepsilon$.*

Indeed, if the gaps between each element x_N of the sequence and the next x_{N+1} were always larger than ε, then $x_M - x_1$ would exceed $(M-1)\varepsilon \geq 1$, a contradiction. This principle is true, but it is too weak to be considered a true finitary version of the infinite convergence principle; indeed, we see that the pigeonhole principle easily implies

Proposition 2.7 (Weak infinite convergence principle). *If $0 \leq x_1 \leq x_2 \leq \ldots \leq 1$, then $\liminf_{n\to\infty} |x_{n+1} - x_n| = 0$.*

but does not obviously imply the full infinite convergence principle.

The problem is that the pigeonhole principle only establishes *instantaneous* stability of the sequence at some point n, whereas the infinite convergence principle concludes the *permanent* stability of the sequence after some point N. To get a better finitary match to the infinite convergence principle, we need to extend the region of stability that the pigeonhole principle offers. Now, one can do some trivial extensions such as

Proposition 2.8 (Pigeonhole principle (second version)). *If $\varepsilon > 0$ and $k \geq 1$ and $0 \leq x_1 \leq x_2 \leq \ldots \leq x_M \leq 1$ is such that $M \geq k/\varepsilon + 1$, there exists $1 \leq N < N+k \leq M$ such that $|x_n - x_m| \leq \varepsilon$ for all $N \leq n, m \leq N+k$.*

which one can quickly deduce from the first pigeonhole principle by considering the sparsified sequence $x_k, x_{2k}, x_{3k}, \ldots$. But this is only a little bit better, as it now gives the infinitary statement

Proposition 2.9 (Slightly less weak infinite convergence principle). *If $0 \leq x_1 \leq x_2 \leq \ldots \leq 1$, then $\liminf_{n\to\infty} |x_{n+k} - x_n| = 0$ for all k.*

but is still not strong enough to imply the infinite convergence principle in its full strength. Nevertheless, it shows that we can extend the realm of stability offered by the pigeonhole principle. One can for instance sparsify further, replacing $n+k$ with $2n$:

Proposition 2.10 (Pigeonhole principle (third version)). *If $\varepsilon > 0$ and $k \geq 1$ and $0 \leq x_1 \leq x_2 \leq \ldots \leq x_M \leq 1$ is such that $M \geq 2^{1/\varepsilon} + 1$, there exists $1 \leq N < 2N \leq M$ such that $|x_n - x_m| \leq \varepsilon$ for all $N \leq n, m \leq 2N$.*

This can be proven by applying the first version of the pigeonhole principle to the sparsified sequence $x_1, x_2, x_4, x_8, \ldots$. This corresponds to an infinite convergence principle in which the conclusion is that $\liminf_{n\to\infty} |x_{2n} - x_n| = 0$.

One can of course keep doing this, achieving various sparsified versions of the pigeonhole principle which each capture part of the infinite convergence principle. To get the full infinite convergence principle, one cannot use any single such sparsified version of the pigeonhole principle, but instead must take *all of them at once*. This is the full strength of the finite convergence principle:

Proposition 2.11 (Finite convergence principle). *If $\varepsilon > 0$ and $F : \mathbf{Z}_+ \to \mathbf{Z}_+$ is a function and $0 \leq x_1 \leq x_2 \leq \ldots \leq x_M \leq 1$ is such that M is sufficiently large depending on F and ε, then there exists $1 \leq N < N + F(N) \leq M$ such that $|x_n - x_m| \leq \varepsilon$ for all $N \leq n, m \leq N + F(N)$.*

This principle is easily proven by appealing to the first pigeonhole principle with the sparsified sequence $x_{i_1}, x_{i_2}, x_{i_3}, \ldots$ where the indices are defined recursively by $i_1 := 1$ and $i_{j+1} := i_j + F(i_j)$. This gives an explicit bound on M as $M := i_{\lfloor 1/\varepsilon \rfloor + 1}$. Note that the first pigeonhole principle corresponds to the case $F(N) \equiv 1$, the second pigeonhole principle to the case $F(N) \equiv k$, and the third to the case $F(N) \equiv N$. A particularly useful case for applications is when F grows exponentially in N, in which case M grows tower-exponentially in $1/\varepsilon$.

Informally, the above principle asserts that any sufficiently long (but finite) bounded monotone sequence will experience arbitrarily high-quality amounts of metastability with a specified error tolerance ε, in which the duration $F(N)$ of the metastability exceeds the time N of onset of the metastability by an arbitrary function F which is specified in advance.

Let us now convince ourselves that this is the true finitary version of the infinite convergence principle, by deducing them from each other:

The finite convergence principle implies the infinite convergence principle. Suppose for contradiction that the infinite convergence principle failed. Untangling the quantifiers, this asserts that there is an infinite sequence $0 \leq x_1 \leq x_2 \leq \ldots \leq 1$ and an $\varepsilon > 0$ with the property that, given any positive integer N, there exists a larger integer $N + F(N)$ such that $x_{N+F(N)} - x_N > \varepsilon$. But this contradicts the finite convergence principle. □

The infinite convergence principle implies the finite convergence principle. Suppose for contradiction that the finite convergence principle failed. Untangling the quantifiers, this asserts that there exists $\varepsilon > 0$ and a function F, together with a collection $0 \leq x_1^{(i)} \leq \ldots \leq x_{M_i}^{(i)} \leq 1$ of bounded monotone sequences whose length M_i goes to infinity, such that for each one of these sequences, there does ¡b¿not¡/b¿ exist $1 \leq N < N + F(N) \leq M_i$ such that $|x_n^{(i)} - x_m^{(i)}| \leq \varepsilon$ for all $N \leq n, m \leq N + F(N)$. Let us extend each of the finite bounded sequences to infinite bounded sequences in some arbitrary manner, e.g. defining $x_n^{(i)} = 1$ whenever $n > M_i$. The space of all bounded sequences is well-known[7] to be sequentially compact in the product topology, thus after refining the i labels to a subsequence if necessary, we can assume that the sequences $(x_n^{(i)})_{n=1}^{\infty}$ converge in the product topology (i.e. pointwise) to a new limit sequence $(x_n)_{n=1}^{\infty}$. Since each of the original sequences were bounded in the interval $[0,1]$ and monotone, we see that the limit sequence is also. Furthermore, we claim that there does not exist any $N \geq 1$ for which $|x_n - x_m| < \varepsilon$ for all $N \leq n, m \leq N + F(N)$. Indeed, if this were the case, then by pointwise convergence we would also have $|x_n^{(i)} - x_m^{(i)}| < \varepsilon$ for all $N \leq n, m \leq N + F(N)$ and all sufficiently large i, but this contradicts the construction of the $x_n^{(i)}$. But now we see that this infinite bounded monotone sequence $(x_n)_{n=1}^{\infty}$ contradicts the infinite convergence principle. □

One can draw some morals from the above discussion:

- The finitary version of an infinitary statement can be significantly more verbose and ugly-looking than the infinitary original, and the arrangement of quantifiers becomes crucial.

[7] This result is of course a consequence of Tychonoff's theorem, but because we require sequential compactness here rather than topological compactness, the result here is in fact much closer in spirit to the Arzelá-Ascoli theorem. In particular, the axiom of choice is not actually used here, instead one can repeatedly use the Bolzano-Weierstrass theorem for the interval $[0,1]$ followed by a diagonalisation argument to establish sequential compactness. The astute reader here will observe that the Bolzano-Weierstrass theorem is essentially equivalent to the infinite convergence principle! Fortunately, there is no circularity here, because we are only using this theorem in order to deduce the finite convergence principle from the infinite, and not the other way around.

- The "naive" finitisation of an infinitary statement is often not the correct one.
- While the finitary version of an infinitary statement is indeed quantitative, the bounds obtained can be quite poor (e.g. tower-exponential or worse).
- The deduction of the infinitary statement from the finitary one is quite short, as long as one is willing to work indirectly (arguing by contradiction).
- The deduction of the finitary statement from the infinitary one is a bit more complicated, but still straightforward, and relies primarily on compactness.
- In particular, the equivalence of the finitary and infinitary formulations requires a non-trivial amount of infinitary mathematics (though in this particular case, we can at least leave the ultrafilters out of it).

These morals apply not only to the finite and infinite convergence principle, but to many other pairs of finitary and infinitary statements, for instance *Szemerédi's theorem*[Sz1975] on one hand and the *Furstenberg recurrence theorem*[Fu1977] on the other; see Section 3.1.2 for more discussion. In these contexts, the correspondence between the finitary and infinitary statements is known as the *Furstenberg correspondence principle*.

2.3.3 Applications

So, we've now extracted a quantitative finitary equivalent of the infinitary principle that every bounded monotone sequence converges. But can we actually *use* this finite convergence principle for some non-trivial finitary application? The answer is a definite *yes*: the finite convergence principle (implicitly) underlies the famous Szemerédi regularity lemma[Sz1975], which is a major tool in graph theory, and also underlies several other regularity lemmas, such as the arithmetic regularity lemma of Green[Gr2005] and the "strong" regularity lemma in [AlFiKrSz2000]. More generally, this principle seems to often arise in any finitary application in which tower-exponential bounds are inevitably involved.

Before plunging into these applications, let us first establish a Hilbert space version[8] of the convergence principle. Given a (closed) subspace X of a Hilbert space H, and a vector $v \in H$, let $\pi_X v$ be the orthogonal projection from v onto X. If X is finite dimensional, then this projection can be defined in a finitary way, for instance by applying the Gram-Schmidt orthogonalisation procedure to X. If X is infinite dimensional, then even the existence of the orthogonal projection is not completely trivial, and in fact relies ultimately on the infinite convergence principle. Closely related to the existence of this projection is the following monotone continuity property:

Proposition 2.12 (Hilbert space infinite convergence principle)**.** *Let $0 \subset X_1 \subset X_2 \subset \ldots \subset H$ be a nested sequence of subspaces of a Hilbert space H, and let $X := \overline{\bigcup_{n=1}^\infty X_n}$ be the monotone closed limit of the X_n. Then for any vector v, $\pi_{X_n} v$ converges strongly in H to $\pi_X v$.*

[8] One could also view this as a "noncommutative" or "quantum" version of the convergence principle, but this is somewhat of an abuse of terminology, despite the presence of the Hilbert space, since we don't actually have any noncommutativity or any other quantum weirdness going on.

As with the infinite convergence principle in $[0,1]$, there is a Cauchy sequence version which already captures the bulk of the content:

Proposition 2.13 (Hilbert space infinite convergence principle (again)). *Let $0 \subset X_1 \subset X_2 \subset \ldots \subset H$ be a nested sequence of subspaces of a Hilbert space H, and let $\varepsilon > 0$. Then for any vector v there exists N such that $\|\pi_{X_n} v - \pi_{X_m} v\|_H^2 \leq \varepsilon$ for all $n, m \geq N$.*

One can deduce this principle from the analogous principle in $[0,1]$ by first normalising $\|v\|_H = 1$, and then observing from Pythagoras' theorem that $\|\pi_{X_n} v\|_H^2$ (which one should view as the *energy* of X_n as measured relative to v) is a bounded monotone sequence from 0 to 1. Applying the infinite convergence principle, followed by Pythagoras' theorem yet again, we obtain the claim. Once one sees this, one immediately concludes that there is also a finitary equivalent:

Proposition 2.14 (Hilbert space finite convergence principle). *If $\varepsilon > 0$ and $F : \mathbf{Z}_+ \to \mathbf{Z}_+$, and $0 \subset X_1 \subset X_2 \subset \ldots X_M \subset H$ is such that M is sufficiently large depending on F and ε, then for any vector v with $\|v\|_H \leq 1$ there exists $1 \leq N \leq N + F(N) \leq M$ such that $\|\pi_{X_n} v - \pi_{X_m} v\|_H^2 \leq \varepsilon$ for all $N \leq n, m \leq N + F(N)$.*

Informally, given a long enough sequence of nested subspaces, and a given bounded vector v, one can find an arbitrarily good region of metastability in the orthogonal projections of v into these subspaces.

From this principle one can then quickly deduce the Szemerédi regularity lemma[Sz1975] as follows. Let $G = (V, E)$ be a graph. One can think of the adjacency matrix 1_E of this graph as an element of the (finite-dimensional) Hilbert space $L^2(V \times V)$, where the product space $V \times V$ is given normalised counting measure (and the discrete sigma-algebra $2^V \times 2^V$). We can construct a nested sequence $\mathscr{B}_0 \subset \mathscr{B}_1 \subset \mathscr{B}_2 \subset \ldots$ of σ-algebras in V (which one can think of as a sequence of increasingly fine partitions of V), together with the attendant sequence $L^2(\mathscr{B}_0 \times \mathscr{B}_0) \subset L^2(\mathscr{B}_1 \times \mathscr{B}_1) \subset \ldots$ of subspaces (this corresponds to functions on $V \times V$ which are constant on any product of pair of cells in the partition), by the following greedy algorithm[9]:

- Step 0. Initialise $\mathscr{B}_0 := \{\emptyset, V\}$ to be the trivial σ-algebra (i.e. the trivial partition).
- Step 1. Given $\mathscr{B}_n$, let $f_n := \mathbb{E}(1_E | \mathscr{B}_n \times \mathscr{B}_n)$ be the orthogonal projection of 1_E to the space $L^2(\mathscr{B}_n \times \mathscr{B}_n)$ (thus the value on any product of cells is just the edge density between that pair of cells), and let $g_n := 1_E - f_n$ be the deviation of the graph from its density.
- Step 2. Let A_n, B_n be sets in V which maximise the discrepancy $\frac{1}{|V|^2} \sum_{a \in A_n} \sum_{b \in B_n} g_n(a,b)$.
- Step 3. Let $\mathscr{B}_{n+1}$ be the σ-algebra generated by $\mathscr{B}_n$ and A_n, B_n. Now increment n to $n+1$ and return to Step 2.

[9] One can also replace this greedy algorithm by a random algorithm, in which each $\mathscr{B}_{n+1}$ is obtained from $\mathscr{B}_n$ by adding a neighbourhood $N(v_n)$ of a randomly chosen vertex v_n in V. This use of randomisation appears in the infinitary setting in [Ta2007e], and in the finitary setting in [Is2008].

Let $\varepsilon > 0$ and $F : \mathbf{Z}_+ \to \mathbf{Z}_+$ be a function. Applying the Hilbert space finite convergence principle to the above sequence of vector spaces, we obtain some N with some bounded size (depending on ε and F) such that

$$\|f_n - f_m\|_{L^2(V\times V)} \leq \varepsilon^2 \tag{2.5}$$

for all $N \leq n,m \leq N + F(N)$. By a further application of the pigeonhole principle (for Hilbert spaces), one can find $N \leq n \leq N + F(N)$ such that

$$\|f_{n+1} - f_n\|_{L^2(V\times V)} \leq \varepsilon^2/F(N).$$

What this basically means is that the partition $\mathscr{B}_n$ is very regular, in that even the greediest way to refine this partition does not significantly capture any more of the fluctuations of the graph G. By choosing F to be a suitable exponentially growing function, one can make the regularity of this partition exceed the number of cells (which is basically 2^{2^N}) in the partition $\mathscr{B}_N$, which is "within epsilon" of the partition $\mathscr{B}_n$ in the sense of (2.5). Putting all this together, one can get a strong version of the Szemerédi regularity lemma, which implies the usual formulation by a simple argument; see [Ta2006h] for further discussion. The choice of F being exponential is what results in the notorious tower-exponential bounds in this regularity lemma (which are necessary, thanks to a result of Gowers[Go1997]). But one can reduce F to, say, a polynomial, resulting in more civilised bounds but with a weaker regularity conclusion. Such a "weak regularity lemma" was for instance established by Frieze and Kannan[FrKa1999], and also underlies the "generalised Koopman von Neumann theorem" which is a key component of my result with Ben Green[GrTa2008] establishing long arithmetic progressions in the primes. In the opposite direction, various flavours of "strong regularity lemma" have appeared in the literature [AlFiKrSz2000], [RoSc2007], [Ta2006h], and also turn out to be convenient ways to formulate hypergraph versions of the regularity lemma of adequate strength to imply non-trivial theorems (such as Szemerdi's theorem).

Rather than using sets which maximise discrepancy, one can also use sublevel sets of the eigenvectors of the adjacency matrix corresponding to the largest eigenvalues of the matrix to generate the partition; see [FrKa1999] for details of a closely related construction.

The appearance of spectral theory (eigenvalues and eigenvectors) into this topic brings one in contact with Fourier analysis, especially if one considers circulant matrices (which correspond in graph-theoretic terms to Cayley graphs on a cyclic group). This leads us towards the arithmetic regularity lemma of Green[Gr2005], which regularises a bounded function f on a finite abelian group G in terms of a partition generated by the sublevel sets (Bohr sets) of a bounded number of characters; the precise formulation is a bit lengthy to state properly, although it simplifies substantially in the "dyadic model" case (see Section 2.6) when G is a vector space over a small finite field (e.g. $\mathbf{F}_2$). This arithmetic regularity lemma can also be established using the finite convergence principle (in either the numerical form or the Hilbert space form). Indeed, if we let $H = L^2(G)$ and let V_n be the vector space generated by the characters associated to the n largest Fourier coefficients of f, then by applying the finite convergence principle (with $v = f$) we can locate a metastable region, where there is not much going on (in an L^2 sense) between V_N and $V_{N+F(N)}$ for some (exponentially growing) function F,

thus there is a "spectral gap" of sorts between the N largest Fourier coefficients and the coefficients ranked $N + F(N)$ and beyond. The sublevel sets of characters associated to the N largest coefficients can then be used to regularise the original function f. Similar ideas also appear in [Bo1986], [GrKo2006]. See also my survey [Ta2007f] for a general discussion of structural theorems of this type.

2.3.4 A non-finitisable statement

One may conclude from the above discussion that every infinitary statement concerning, say, the natural numbers, has a finitary analogue which is equivalent to it. It turns out that this is not quite the case (and in fact some subtle issues in logic come in), even for very "natural" and "non-pathological" infinitary statements. In particular, the following innocuous-looking infinitary statement is basically non-finitisable:

Proposition 2.15 (Infinite pigeonhole principle)**.** *If the natural numbers are divided into finitely many colour classes, then one of the colour classes is infinite.*

This principle is of course a triviality in the infinitary world, and there are some obvious candidates for finitary versions (e.g. Proposition 2.6), but they are not equivalent to the infinitary principle the same way that the finite convergence principle is equivalent to the infinite convergence principle; there is a failure of compactness here. There is a finitary version (of sorts), but it is somewhat difficult to state. Define a *set function* to be any function f which takes a finite set A of natural numbers as input, and returns a natural number $F(A)$ as output. Let us say that a set function F is *asymptotically stable*[10] if, given any nested sequence $A_1 \subset A_2 \subset A_3 \subset \ldots$ of finite sets, the numbers $F(A_n)$ are constant for sufficiently large n. For instance, the "least element" set function $f(A) = \inf(A)$ (adopting the unusual convention that the infimum of the empty set is 0) is asymptotically stable, but the "cardinality" set function $f(A) = |A|$ is not. Anyway, the correct "finitary" version of the infinite pigeonhole principle is

Proposition 2.16 ("Finitary" infinite pigeonhole principle)**.** *Let F be an asymptotically stable set function, and let k be a positive integer. Then there exists a positive integer N with the property that whenever the set $\{1,\ldots,N\}$ is divided into k colour classes, at least one of the colour classes A has the property that $|A| > F(A)$.*

It is a good exercise to see that this principle is equivalent to the infinite pigeonhole principle. Note that Proposition 2.6 essentially corresponds to the case when F is a constant function - and is thus definitely a very special case. The case when

[10] This concept is *very* roughly equivalent to the notion of a function $F(A)$ defined for all sets of integers (both finite and infinite) which can always be "computed" in "finite time". But one should take this informal definition with a large grain of salt: while there is indeed an algorithm for computing $F(A)$ for any given set A which will eventually give the right answer, you might not be able to tell when the algorithm has finished! A good example is the asymptotically stable function $F(A) := \inf(A)$: you can "compute" this function for any set A by initialising the answer to 0, running a counter n from 0 to infinity, and resetting the answer permanently to n the first time n lies in A. As long as A is non-empty, this algorithm terminates in finite time with the correct answer; if A is empty, the algorithm gives the right answer from the beginning, but you can never be sure of this fact! In contrast, the cardinality $|A|$ of a possibly infinite set A cannot be computed even in this rather unsatisfactory sense of having a running "provisional answer" which is guaranteed to eventually be correct.

$F(A) = f(\inf(A))$ for some fixed function f is already very interesting - it is the 1-uniform case of a "strong Ramsey theorem" and is barely provable by finitary means[11], although the general case of that theorem is not finitarily provable, even if it is an immediate consequence of Proposition 2.16; this assertion is essentially the celebrated *Paris-Harrington theorem*. The assumption of asymptotic stability of F is necessary, as one can see by considering the counterexample $F(A) := |A|$.

I am enclosing "finitary" in quotes in Proposition 2.16, because while most of the assertion of this principle is finitary, one part still is not, which is the notion of "asymptotically stable". This is a notion which cannot be precisely formulated in a purely finitary manner, even though the notion of a set function is basically a finitary concept (ignoring for now a subtle issue about what "function" means). If one insists on working in a finitary setting, then one can recast the infinite pigeonhole principle as a *schema* of finitary principles, one for each asymptotically stable set function F, but in order to work out exactly which set functions are asymptotically stable or not requires infinitary mathematics. (And for some (constructible, well-defined) set functions, the asymptotic stability is *undecidable*; this fact is closely related to the undecidability of the halting problem and is left as an exercise to the reader.)

The topic of exactly which statements in infinitary mathematics are "truly infinitary" is a fascinating one, and is basically a question in *reverse mathematics*, but we will not be able to discuss it here.

2.3.5 Notes

This article was originally posted on May 23, 2007 at

`terrytao.wordpress.com/2007/05/23`

I am indebted to Harvey Friedman for discussions on the Paris-Harrington theorem and the infinite pigeonhole principle, and to Henry Towsner, Ulrich Kohlenbach, and Steven Simpson for pointing out the connections to proof theory and reverse mathematics.

Richard Borcherds pointed out that the distinction between hard and soft analysis was analogous to the distinction between first-order and second-order logic.

JL pointed out the paper of Freedman[Fr1998], in which a limiting process is proposed to convert problems in complexity theory to some infinitary counterpart in decidability theory.

Thanks to Liu Xiao Chuan for corrections.

[11] Try it, say for $k = 10$ and $F(A) := \inf(A) + 10$. What quantitative bound for N do you get?

2.4 The Lebesgue differentiation theorem and the Szemerédi regularity lemma

This article is a sequel of sorts to Section 2.3. Here, I want to discuss a well-known theorem in infinitary soft analysis - the *Lebesgue differentiation theorem* - and whether there is any meaningful finitary version of this result. Along the way, it turns out that we will uncover a simple analogue of the Szemerédi regularity lemma, for subsets of the interval rather than for graphs. (Actually, regularity lemmas seem to appear in just about any context in which fine-scaled objects can be approximated by coarse-scaled ones.) The connection between regularity lemmas and results such as the Lebesgue differentiation theorem was recently highlighted by Elek and Szegedy[ElSz2008], while the connection between the finite convergence principle and results such as the pointwise ergodic theorem (which is a close cousin of the Lebesgue differentiation theorem) was recently detailed by Avigad, Gerhardy, and Towsner[AvGeTo2008].

The Lebesgue differentiation theorem has many formulations, but we will avoid the strongest versions and just stick to the following model case for simplicity:

Theorem 2.17 (Lebesgue differentiation theorem)**.** *If $f : [0,1] \to [0,1]$ is Lebesgue measurable, then for almost every $x \in [0,1]$ we have $f(x) = \lim_{r\to 0} \frac{1}{r} \int_x^{x+r} f(y)\ dy$. Equivalently, the fundamental theorem of calculus $f(x) = \frac{d}{dy} \int_0^y f(z)dz|_{y=x}$ is true for almost every x in $[0,1]$.*

Here we use the oriented definite integral, thus $\int_x^y = -\int_y^x$. Specialising to the case where $f = 1_A$ is an indicator function, we obtain the Lebesgue density theorem as a corollary:

Corollary 2.18 (Lebesgue density theorem)**.** *Let $A \subset [0,1]$ be Lebesgue measurable. Then for almost every $x \in A$, we have $\frac{|A\cap[x-r,x+r]|}{2r} \to 1$ as $r \to 0^+$, where $|A|$ denotes the Lebesgue measure of A.*

In other words, almost all the points x of A are points of density of A, which roughly speaking means that as one passes to finer and finer scales, the immediate vicinity of x becomes increasingly saturated with A. (Points of density are like robust versions of interior points, thus the Lebesgue density theorem is an assertion that measurable sets are almost like open sets. This is Littlewood's first principle.) One can also deduce the Lebesgue differentiation theorem back from the Lebesgue density theorem by approximating f by a finite linear combination of indicator functions; we leave this as an exercise.

The Lebesgue differentiation and density theorems are qualitative in nature: they assert that $\frac{1}{r} \int_x^{x+r} f(y)\ dy$ eventually gets close to $f(x)$ for almost every x, or that A will eventually occupy most of $[x-r,x+r]$ for almost every x in A, by taking r small enough, but does not give a quantitative bound for how small r has to be. The following simple example shows why there is a problem. Let n be a large integer, and partition $[0,1]$ into 2^n dyadic intervals $[0,1/2^n],[1/2^n,2/2^n],\ldots,[1-1/2^n,1]$ (never mind about the overlaps on the boundary, as they have zero measure, and the Lebesgue philosophy in analysis is to always ignore anything which happens on sets of measure zero). Let

A_n be the union of every other interval:

$$A_n = [0,1/2^n] \cup [2/2^n, 3/2^n] \cup \ldots \cup [1-2/2^n, 1-1/2^n].$$

One then sees that if x is any element of A_n which is not on the boundary, then it is indeed true that the local density $\frac{|A_n \cap [x-r,x+r]|}{2r}$ of A_n will eventually converge to 1, but one has to wait until r is of size $1/2^n$ or smaller before one sees this; for scales much larger than this, the local density will remain stubbornly close to $1/2$. A similar phenomenon holds for the indicator functions $f_n := 1_{A_n}$: the local average $\frac{1}{r}\int_x^{x+r} f_n(y)\,dy$ will eventually get close to $f_n(x)$, which is either 0 or 1, but when $|r| \gg 1/2^n$, these averages will also stay close to $1/2$. (Closely related to this is the fact that the functions f_n converge weakly to $1/2$, despite only taking values in $\{0,1\}$.)

Intuitively, what is going on here is that while each set A_n is certainly Lebesgue measurable, these sets are getting increasingly "less measurable" as n gets large, and the rate of convergence in the Lebesgue differentiation and density theorems depends on how measurable the sets A_n are. One can illustrate this by considering (non-rigorously) the limiting case $n = \infty$ as follows. Suppose we select a random subset A_∞ of $[0,1]$ by requiring each real number x in $[0,1]$ to lie in A_∞ with an independent probability of $1/2$ (thus we are flipping an uncountable number of coins to determine this set!). The law of large numbers (applied very non-rigorously!) then suggests that with probability 1, A_∞ should have density $1/2$ in every single interval I in $[0,1]$, thus $|A_\infty \cap I| = \frac{1}{2}|I|$. This would seem to violate the Lebesgue density theorem; but what is going on here is that the set A_∞ is in fact almost surely non-measurable (indeed, the Lebesgue density theorem provides a proof of this fact, modulo the issues of justifying several non-rigorous claims in this paragraph).

So, it seems that to proceed further we need to quantify the notion of measurability, in order to decide which sets or functions are "more measurable" than others. There are several ways to make such a quantification. Here are some typical proposals:

Definition 2.19. A set $A \subset [0,1]$ is (ε,n)-*measurable* if there exists a set B which is the union of dyadic intervals $[j/2^n,(j+1)/2^n]$ at scale 2^{-n}, such that A and B only differ on a set of Lebesgue measure (or outer measure) ε.

Definition 2.20. A function $f : [0,1] \to [0,1]$ is (ε,n)-measurable if there exists a function g which is constant on the dyadic intervals $[j/2^n,(j+1)/2^n]$, which differs from f in L^1-norm by at most ε, thus $\int_0^1 |f(x)-g(x)|\,dx \leq \varepsilon$.

Remark 2.21. One can phrase these definitions using the σ-algebra generated by the dyadic intervals of length 2^{-n}; we will not do so here, but these σ-algebras are certainly underlying our discussion. Their presence is particularly prominent in the "ergodic theory" approach to this circle of ideas, which we are not focusing on here.

One can now obtain the following quantitative results:

Theorem 2.22 (Quantitative Lebesgue differentiation theorem)**.** *Let $f : [0,1] \to [0,1]$ be (ε,n)-measurable. Then for all x in $[0,1]$ outside of a set of measure $O(\sqrt{\varepsilon})$ we have $\frac{1}{r}\int_x^{x+r} f(y)\,dy = f(x) + O(\sqrt{\varepsilon})$ for all $0 < r < \sqrt{\varepsilon}2^{-n}$.*

Theorem 2.23 (Quantitative Lebesgue density theorem)**.** *Let $A \subset [0,1]$ be (ε,n)-measurable. Then for all x in A outside of a set of measure $O(\sqrt{\varepsilon})$ we have $|A \cap [x-r,x+r]| = (1-O(\sqrt{\varepsilon}))2r$ for all $0 < r < \sqrt{\varepsilon}2^{-n}$.*

These results follow quickly from the Hardy-Littlewood maximal inequality, and by exploiting the low "complexity" of the structured objects B and g which approximate A and f respectively; we omit the standard arguments.

To connect these quantitative results back to their qualitative counterparts, one needs to connect the quantitative notion of (ε,N)-measurability with the traditional qualitative notion of Lebesgue measurability. The relevant result that provides this connection is

Theorem 2.24 (Lebesgue approximation theorem, first version)**.** *Let $A \subset [0,1]$ be measurable. Then for every $\varepsilon > 0$ there exists n such that A is (ε,n)-measurable.*

Theorem 2.25 (Lebesgue approximation theorem, second version)**.** *Let f :[0,1]$\to$ $[0,1]$ be measurable. Then for every $\varepsilon > 0$ there exists n such that f is (ε,n)-measurable.*

These two results are easily seen to be equivalent. Let us quickly recall the proof of the first version:

Proof of Theorem 2.24. The claim is easily verified when A is the finite union of dyadic intervals, and then by monotone convergence one also verifies the claim when A is compact (or open). One then verifies that the claim is closed under countable unions, intersections, and complements, which then gives the claim for all Borel-measurable sets. The claim is also obviously true for null sets, and thus true for Lebesgue-measurable sets. □

So, we're done, right? Well, there is still an unsatisfactory issue: the Lebesgue approximation theorems guarantee, for any given ε, that a measurable set A or a measurable function f will eventually be (ε,n)-measurable by taking n large enough, but don't give any *bound* as to what this n will be. In a sense, this is unavoidable, even if we consider "nice" objects such as compact sets A or piecewise constant functions f; the example of the set A_n or the function f_n discussed previously show that for fixed ε, one can be forced to take n to be arbitrarily large.

However, we can start looking for substitutes for these theorems which do have quantitative bounds. Let's focus on the first version of the Lebesgue approximation theorem, and in particular in the case when A is compact. Then, we can write $A = \bigcap_{n=1}^{\infty} A^{(n)}$, where $A^{(n)}$ is the union of all the (closed) dyadic intervals which intersect A. The measures $|A^{(n)}|$ are a monotone decreasing sequence of numbers between 0 and 1, and thus (by Proposition 2.1!) they have a limit, which (by the upper continuity of Lebesgue measure) is just $|A|$. Thus, for every $\varepsilon > 0$ we have $|A^{(n)}| - |A| < \varepsilon$ for all sufficiently large n, which explains why A is (ε,n)-measurable for all large n.

So now we see where the lack of a bound on n is coming from - it is the fact that the infinite convergence principle also does not provide an effective bound on the rate of convergence. But in Section 2.3, we saw how the finite convergence theorem (Proposition 2.11) did provide an effective substitute for the infinite convergence principle. If we apply it directly to this sequence $|A^{(n)}|$, this is what we get:

Theorem 2.26 (Effective Lebesgue approximation theorem)**.** *Let $F : \mathbf{N} \to \mathbf{N}$ be any function, and let $\varepsilon > 0$. Then there exists an integer N with the following property: given any subset A of $[0,1]$, there exists $1 \leq n \leq N$ such that $A^{(n+F(n))}$ is (ε, n)-measurable.*

This theorem does give a specific upper bound on the scale n one has to reach in order to get quantitative measurability. The catch, though, is that the measurability is not attained for the original set A, but instead on some discretisation $A^{(n+F(n))}$ of A. However, we can make the scale at which we are forced to discretise to be arbitrarily finer than the scale at which we have the measurability.

Nevertheless, this theorem is still a little unsatisfying, because it did not directly say too much about the original set A. There is an alternate approach which gives a more interesting result. In the previous results, the goal was to try to approximate an arbitrary object (a set or a function) by a "structured" or 'low-complexity" one (a finite union of intervals, or a piecewise constant function), thus trying to move away from "pseudorandom" or "high-complexity" objects (such as the sets A_n and functions f_n discussed earlier). Of course, the fact that these pseudorandom objects actually exist is what is making this goal difficult to achieve satisfactorily. However, one can adopt a different philosophy, namely to embrace both the structured and the pseudorandom aspects of these objects, and focus instead on creating an efficient decomposition of arbitrary objects into the structured and pseudorandom components.

To do this, we need to understand what "pseudorandom" means. One clue is to look at the examples A_n and f_n discussed earlier. Observe that if one averages f_n on any reasonable sized interval J, one gets something very close to the global average of f_n, i.e. $1/2$. In other words, the integral of f_n on an interval J is close to the global average of f_n times $|J|$. (This is also true when J is a small interval, since in this case both expressions are small.) This motivates the following definition:

Definition 2.27. A function $f : [0,1] \to [0,1]$ is said to be *ε-regular* on a dyadic interval I if we have $|\int_J f(y)\,dy - \frac{|J|}{|I|}\int_I f(y)\,dy| \leq \varepsilon |I|$ for all dyadic subintervals $J \subset I$.

Thus, for instance, f_n is 2^{-n}-regular on $[0,1]$. We then have an analogue of the Szemerédi regularity lemma for subsets of the interval, which I will dub the "Lebesgue regularity lemma":

Lemma 2.28 (Lebesgue regularity lemma)**.** *If $\varepsilon > 0$ and $f : [0,1] \to [0,1]$ is measurable, then there exists an positive integer $n = O_\varepsilon(1)$ (i.e. n is bounded by a quantity depending only on ε), such that f is ε-regular on all but at most $\varepsilon 2^n$ of the 2^n dyadic intervals of length 2^{-n}.*

Proof. As with the proof of many other regularity lemmas, we shall rely primarily on the *energy increment argument* (the energy is also known as the *index* in some literature). For minor notational reasons we will take ε to be a negative power of 2. For each integer n, let $f^{(n)} : [0,1] \to [0,1]$ be the conditional expectation of f to the dyadic intervals of length 2^{-n}, thus on each such interval I, $f^{(n)}$ is equal to the constant value $\frac{1}{I}\int_I f$ (again, we are ignoring sets of measure zero). An easy application of Pythagoras's theorem (for $L^2([0,1])$) shows that the energies $E_n := \int_0^1 |f^{(n)}(x)|^2\,dx$ are

an increasing sequence in n, and bounded between 0 and 1. Applying (a special case of) the finite convergence principle, we can find $n = O_\varepsilon(1)$ such that we have the energy metastability

$$E_{n+\log_2 1/\varepsilon} - E_n \leq \varepsilon^3.$$

(Indeed, we obtain a fairly civilised bound of $n \leq \varepsilon^3 \log_2 1/\varepsilon$). Applying Pythagoras' theorem again, we conclude

$$\int_0^1 |f^{(n+\log_2 1/\varepsilon)}(x) - f^{(n)}(x)|^2 dx \leq \varepsilon^3$$

which by Markov's inequality implies that

$$\int_I |f^{(n+\log_2 1/\varepsilon)}(x) - f^{(n)}(x)|^2 dx \leq \varepsilon^2 |I|$$

for all but $\varepsilon 2^n$ of the 2^n dyadic intervals I of length 2^{-n}. The Cauchy-Schwarz inequality then quickly shows that f is ε-regular on all of these intervals. □

One can of course specialise the above lemma to indicator functions $f = 1_A$ to obtain a regularity lemma for subsets of the interval, whose formulation I will leave to the reader as an exercise.

An inspection of the proof shows that the full power of the finite convergence principle was not tapped here, because we only used it to get a metastability region of $\log_2 1/\varepsilon$. One can get a stronger result (at the cost of worsening the bound on n) by extending this region of stability. To motivate this stronger version, first observe that if a function f is ε-regular on an interval I, then on that interval we have a decomposition $f = c + h$ on I where $c = \frac{1}{|I|}\int_I f(y)dy$ is the mean of f on I (in particular, c is constant), and h has small averages in the sense that $|\int_J h(y)\,dy| \leq \varepsilon |I|$ for all dyadic subintervals J of I. We can do better than this:

Definition 2.29. A function f :[0,1]→ $[0,1]$ is said to be *strongly* (ε, m)*-regular* on a dyadic interval I if there exists a decomposition $f = c + e + h$ on I, where $c = \frac{1}{|I|}\int_I f(y)dy$ is the mean of f on I, e is small in the sense that $\frac{1}{|I|}\int_I |e(y)|\,dy \leq \varepsilon$, and h has vanishing averages in the sense that $\int_J h(y)dy = 0$ for all dyadic subintervals $J \subset I$ with $|J| \geq 2^{-m}|I|$.

Remark 2.30. Note that strong (ε, m)-regularity implies 2ε-regularity as long as $2^{-m} \leq \varepsilon$. Strong (ε, m)-regularity offers much better control on the fluctuation of f at finer scales, as long as the scale is not too fine (this is where the parameter m comes in).

Lemma 2.31 (Strong Lebesgue regularity lemma)**.** *If $\varepsilon > 0$, $F : \mathbf{N} \to \mathbf{N}$, and $f : [0,1] \to [0,1]$ is measurable, then there exists an positive integer $n = O_{\varepsilon,F}(1)$ such that f is $(\varepsilon, F(n))$-regular on all but at most $\varepsilon 2^n$ of the 2^n dyadic intervals of length 2^{-n}.*

The bound on n is rather poor; it is basically a $1/\varepsilon^3$-fold iteration of the function $n \mapsto n + F(n)\log_2 1/\varepsilon$ applied to 1, so for instance if one wanted F to be exponential in nature then n might be as large as a tower of exponentials of height $1/\varepsilon^3$ or so. (A

very similar situation occurs for the Szemerédi regularity lemma, which has a variety of such strong versions [AlFiKrSz2000], [RoSc2007], [Ta2006h].)

We can now return to the Lebesgue differentiation theorem, and use the strong regularity lemma to obtain a more satisfactory quantitative version of that theorem:

Theorem 2.32 (Quantitative Lebesgue differentiation theorem)**.** *If $\varepsilon > 0$, $F : \mathbf{N} \to \mathbf{N}$, and f :[0,1]$\to [0,1]$ is measurable, then there exists an positive integer $n = O_{\varepsilon,F}(1)$ such that for all x in $[0,1]$ outside of a set of measure $O(\varepsilon)$ we have the Cauchy sequence property $|\frac{1}{r}\int_x^{x+r} f(y)\,dy - \frac{1}{s}\int_x^{x+s} f(y)\,dy| \leq \varepsilon$ for all $2^{-n-F(n)} < r,s < 2^{-n}$.*

This theorem can be deduced fairly easily by combining the strong regularity lemma with the Hardy-Littlewood maximal inequality (to deal with the errors e), and by covering (most of) the non-dyadic intervals $[x,x+r]$ or $[x,x+s]$ by dyadic intervals and using the boundedness of f to deal with the remainder. We leave the details as an exercise.

One sign that this is a true finitary analogue of the infinitary differentiation theorem is that this finitary theorem implies most of the infinitary theorem; namely, it shows that for any measurable f, and almost every x, the sequence $\frac{1}{r}\int_x^{x+r} f(y)$ is a Cauchy sequence as $r \to 0$, although it does not show that the limit is equal to $f(x)$. (Finitary statements can handle Cauchy sequences - which make sense even in the rationals - but have some trouble actually evaluating the limits of such sequences - which need the (infinite precision) real numbers and thus not truly finitary.) Conversely, using weak compactness methods one can deduce the quantitative differentiation theorem from the infinitary one, in much the same way that the finite and infinite convergence principles can be deduced from each other.

The strong Lebesgue regularity lemma can also be used to deduce the (one-dimensional case of the) Rademacher differentiation theorem, namely that a Lipschitz continuous function from $[0,1]$ to the reals is almost everywhere differentiable. To see this, suppose for contradiction that we could find a function g which was Lipschitz continuous but failed to be differentiable on a set of positive measure, thus for every x in this set, the (continuous) sequence $\frac{g(x+r)-g(x)}{r}$ is not a Cauchy sequence as r goes to zero. We can normalise the Lipschitz constant of g to equal 1. Then by standard arguments we can find $\varepsilon > 0$ and a function $F : \mathbf{N} \to \mathbf{N}$ such that for every x in a set of measure greater than ε, and every n, the sequence $\frac{g(x+r)-g(x)}{r}$ fluctuates by at least ε in the range $2^{-n-F(n)} < r < 2^{-n}$. Now let M be a very large integer (depending on ε and F) and discretise g to scale 2^{-M} to create a piecewise linear approximant g_M, which is the antiderivative of a bounded function f which is constant on dyadic intervals of length 2^{-M}. We apply the strong Lebesgue regularity lemma to f and find a scale $n = O_{F,\varepsilon}(1)$ for which the conclusion of that lemma holds; by choosing n large enough we can ensure that $M \geq n + F(n) \geq n$. It is then not hard to see that the lemma contradicts the previous assertion that $\frac{g(x+r)-g(x)}{r}$ fluctuates for certain ranges of x and r.

I used several of the above ideas in [Ta2008c] to establish a quantitative version of the Besicovitch projection theorem.

2.4.1 Notes

This article was originally posted on Jun 18, 2007 at

terrytao.wordpress.com/2007/06/18

Jeremy Avigad mentioned some connections with Steinhaus's classic theorem that if A, B are subsets of $\mathbf{R}$ with positive measure, then the set $A+B$ contains an interval, for which effective versions have been recently established.

2.5 Ultrafilters, nonstandard analysis, and epsilon management

This article is in some ways an antithesis of Section 2.3. There, the emphasis was on taking a result in soft analysis and converting it into a hard analysis statement (making it more "quantitative" or "effective"); here we shall be focusing on the reverse procedure, in which one harnesses the power of infinitary mathematics - in particular, ultrafilters and nonstandard analysis - to facilitate the proof of finitary statements.

Arguments in hard analysis are notorious for their profusion of "epsilons and deltas". In the more sophisticated arguments of this type, one can end up having an entire army of epsilons $\varepsilon_1, \varepsilon_2, \varepsilon_3, \ldots$ that one needs to manage, in particular choosing each epsilon carefully to be sufficiently small compared to other parameters (including other epsilons), while of course avoiding an impossibly circular situation in which a parameter is ultimately required to be small with respect to itself, which is absurd. This art of *epsilon management*, once mastered, is not terribly difficult - it basically requires one to mentally keep track of which quantities are "small", "very small", "very very small", and so forth - but when these arguments get particularly lengthy, then epsilon management can get rather tedious, and also has the effect of making these arguments unpleasant to read. In particular, any given assertion in hard analysis usually comes with a number of unsightly quantifiers (For every ε there exists an N...) which can require some thought for a reader to parse. This is in contrast with soft analysis, in which most of the quantifiers (and the epsilons) can be cleanly concealed via the deployment of some very useful terminology; consider for instance how many quantifiers and epsilons are hidden within, say, the Heine-Borel theorem (a subset of a Euclidean space is compact if and only if it is closed and bounded).

For those who practice hard analysis for a living (such as myself), it is natural to wonder if one can somehow "clean up" or "automate" all the epsilon management which one is required to do, and attain levels of elegance and conceptual clarity comparable to those in soft analysis, hopefully without sacrificing too much of the "elementary" or "finitary" nature of hard analysis in the process.

One important step in this direction has been the development of various types of *asymptotic notation*, such as the Hardy notation of using unspecified constants C, the Landau notation of using $O()$ and $o()$, or the Vinogradov notation of using symbols such as $\ll$ or $\lesssim$; each of these symbols, when properly used, absorbs one or more of the ambient quantifiers in a hard analysis statement, thus making these statements easier to read. But, as useful as these notations are, they still fall a little short of fully capturing one's intuition regarding orders of magnitude. For instance, we tend to think of any quantity of the form $O(1)$ as being "bounded", and we know that bounded objects can be combined to form more bounded objects; for instance, if $x = O(1)$ and $y = O(1)$, then $x+y = O(1)$ and $xy = O(1)$. But if we attempt to formalise this by trying to create the set $A := \{x \in \mathbf{R} : x = O(1)\}$ of all bounded numbers, and asserting that this set is then closed under addition and multiplication, we are speaking nonsense; the $O()$ notation cannot be used within the *axiom schema of specification*, and so the above definition of A is meaningless.

There is however, a way to make concepts such as "the set of all bounded numbers"

precise and meaningful, by using *non-standard analysis*, which is the most well-known of the "pseudo-finitary" approaches to analysis, in which one adjoins additional numbers to the standard number system. Similarly for "bounded" replaced by "small", "polynomial size", etc.. Now, in order to set up non-standard analysis one needs a (non-principal) *ultrafilter* (or an equivalent gadget), which tends to deter people from wanting to hear more about the subject. Because of this, many treatments of non-standard analysis tend to gloss over the actual *construction* of non-standard number systems, and instead emphasise the various *benefits* that these systems offer, such as a rigorous supply of infinitesimals, and a general *transfer principle* that allows one to convert statements in standard analysis into equivalent ones in non-standard analysis. This transfer principle (which requires the ultrafilter to prove) is usually recommended to be applied only at the very beginning and at the very end of an argument, so that the bulk of the argument is carried out purely in the non-standard universe.

I feel that one of the reasons that non-standard analysis is not embraced more widely is because the transfer principle, and the ultrafilter that powers it, is often regarded as some sort of "black box" which mysteriously bestows some certificate of rigour on non-standard arguments used to prove standard theorems, while conveying no information whatsoever on what the quantitative bounds for such theorems should be. Without a proper understanding of this black box, a mathematician may then feel uncomfortable with *any* non-standard argument, no matter how impressive and powerful the result.

The purpose of this post is to try to explain this black box from a "hard analysis" perspective, so that one can comfortably and productively transfer into the non-standard universe whenever it becomes convenient to do so (in particular, it can become cost-effective to do this whenever the burden of epsilon management becomes excessive, and one is willing to not make certain implied constants explicit).

2.5.1 What is an ultrafilter?

In order to do all this, we have to tackle head-on the notorious concept of a non-principal ultrafilter. Actually, these ultrafilters are not as impossible to understand as their reputation suggests; they are basically a consistent set of rules which allow one to always take limits (or make similar decisions) whenever necessary.

To motivate them, let us recall some of the properties of convergent sequences from undergraduate real analysis. If x_n is a convergent sequence of real numbers (where n ranges in the natural numbers), then we have a limit $\lim x_n$, which is also a real number. In addition to the usual analytical interpretations, we can also interpret the concept of a limit as a voting system, in which the natural numbers n are the voters, which are each voting for a real number x_n, and the limit $\lim x_n$ is the elected "winner" emerging from all of these votes. One can also view the limit (somewhat non-rigorously) as the expected value of x_n when n is a "randomly chosen" natural number. Ignoring for now the objection that the natural numbers do not admit a uniform probability measure, it is intuitively clear that such a "randomly chosen" number is almost surely going to be larger than any fixed finite number, and so almost surely x_n will be arbitrarily close to the limit $\lim x_n$ (thus we have a sort of "concentration of measure").

These limits obey a number of laws, including

1. (Algebra homomorphism) If x_n, y_n are convergent sequences, and c is a real number, then $\lim 1 = 1$, $\lim cx_n = c\lim x_n$, $\lim x_n + y_n = \lim x_n + \lim y_n$, and $\lim x_n y_n = (\lim x_n)(\lim y_n)$. (In particular, all sequences on the left-hand side are convergent.)

2. (Boundedness) If x_n is a convergent sequence, then $\inf x_n \leq \lim x_n \leq \sup x_n$. (In particular, if x_n is non-negative, then so is $\lim x_n$.)

3. (Non-principality) If x_n and y_n are convergent sequences which differ at only finitely many values of n, then $\lim x_n = \lim y_n$. [Thus, no individual voter has any influence on the outcome of the election!]

4. (Shift invariance) If x_n is a convergent sequence, then for any natural number h we have $\lim x_{n+h} = \lim x_n$.

These properties are of course very useful in computing the limits of various convergent sequences. It is natural to wonder if it is possible to generalise the notion of a limit to cover various non-convergent sequences, such as the class $l^\infty(\mathbf{N})$ of *bounded* sequences. There are of course many ways to do this in the literature (e.g. if one considers series instead of sequences, one has Cesáro summation, zeta function regularisation, etc.), but (as observed by Euler) one has to give up at least one of the above four limit laws if one wants to evaluate the limit of sequences such as $0,1,0,1,0,1,\ldots$. Indeed, if this sequence had a limit x, then the algebra homomorphism laws force $x^2 = x$ and thus x is either 0 or 1; on the other hand, the algebra homomorphism laws also show us that $1,0,1,0,\ldots$ has a limit $1-x$, and hence by shift invariance we have $x = 1-x$, which is inconsistent with the previous discussion. In the voting theory interpretation, the problem here is one of lack of consensus: half of the voters want 0 and the other half want 1, and how can one consistently and fairly elect a choice from this? Similarly, in the probabilistic interpretation, there is no concentration of measure; a randomly chosen x_n is not close to its expected value of $1/2$, but instead fluctuates randomly between 0 and 1.

So, to define more general limits, we have to give up something. We shall give up shift-invariance (property 4). In the voting theory interpretation given earlier, this means that we abandon the pretense that the election is going to be "fair"; some voters (or groups of voters) are going to be treated differently than others, due to some arbitrary choices made in designing the voting system. (This is the first hint that the axiom of choice will be involved.) Similarly, in the probabilistic interpretation, we will give up the notion that the "random number" n we will choose has a shift-invariant distribution, thus for instance n could have a different distribution than $n+1$.

Suppose for the moment that we managed to have an improved concept of a limit which assigned a number, let's call it $p-\lim x_n$, to any bounded sequence, which obeyed the properties 1-3. It is then easy to see that this p-limit extends the ordinary notion of a limit, because of a sequence x_n is convergent, then after modifying the sequence on finitely many elements we can keep the sequence within ε of $\lim x_n$ for any specified $\varepsilon > 0$, which implies (by properties 2, 3) that $p-\lim x_n$ stays within ε of $\lim x_n$, and the claim follows.

Now suppose we consider a *Boolean sequence* x_n - one which takes only the values 0 and 1. Since $x_n^2 = x_n$ for all n, we see from property that $(p-\lim x_n)^2 = p-\lim x_n$, thus

$p-\lim x_n$ must also be either 0 or 1. From a voting perspective, the p-limit is a *voting system*: a mechanism for extracting a yes-no answer out of the yes-no preferences of an infinite number of voters.

Let p denote the collection of all subsets A of the natural numbers such that the indicator sequence of A (i.e. the boolean sequence x_n which equals 1 when n lies in A and equals 0 otherwise) has a p-limit of 1; in the voting theory language, p is the collection of all voting blocs who can decide the outcome of an election by voting in unison, while in the probability theory language, p is the collection of all sets of natural numbers of measure 1. It is easy to verify that p has four properties:

1. (Monotonicity) If A lies in p, and B contains A, then B lies in p.
2. (Closure under intersection) If A and B lie in p, then $A \cap B$ also lies in p.
3. (Dichotomy) If A is any set of natural numbers, either A or its complement lies in p, but not both.
4. (Non-principality) If one adds (or deletes) a finite number of elements to (or from) a set A, this does not affect whether the set A lies in p.

A collection p obeying properties 1 and 2 is called a *filter*; a collection obeying 1,2, and 3 is called an *ultrafilter*, and a collection obeying 1,2,3, and 4 is a *non-principal ultrafilter*]footnoteIn contrast, a principal ultrafilter is one which is controlled by a single index n_0 in the sense that $p = \{A : n_0 \in A\}$. In voting theory language, this is a scenario in which n_0 is a dictator; in probability language, the random variable n is now a deterministic variable taking the values of n_0.

A property $A(n)$ pertaining to a natural number n can be said to be *p-true* if the set $\{n : A(n) \text{ true}\}$ lies in p, and *p-false* otherwise; for instance any tautologically true statement is also p-true. Using the probabilistic interpretation, these notions are analogous to those of "almost surely true" and "almost surely false" in probability theory. (Indeed, one can view p as being a probability measure on the natural numbers which always obeys a zero-one law, though one should caution that this measure is only finitely additive rather than countably additive, and so one should take some care in applying measure-theoretic technology directly to an ultrafilter.)

Properties 1-3 assert that this notion of "p-truth" obeys the usual laws of propositional logic; for instance property 2 asserts that if A is p-true and B is p-true, then so is "A and B", while property 3 is the familiar law of the excluded middle and property 1 is *modus ponens*. This is actually rather remarkable: it asserts that ultrafilter voting systems cannot create voting paradoxes, such as those guaranteed by Arrow's theorem. There is no contradiction here, because Arrow's theorem only applies to *finite* (hence *compact*) electorates of voters, which do not support any non-principal ultrafilters. At any rate, we now get a hint of why ultrafilters are such a useful concept in logic and model theory.

We have seen how the notion of a p-limit creates a non-principal ultrafilter p. Conversely, once one has a non-principal ultrafilter p, one can uniquely recover the p-limit operation. This is easiest to explain using the voting theory perspective. With the ultrafilter p, one can ask yes-no questions of an electorate, by getting each voter to answer yes or no and then seeing whether the resulting set of "yes" voters lies in p. To take

a p-limit of a bounded sequence x_n, say in $[0,1]$, what is going on is that each voter n has his or her own favourite candidate number x_n between 0 and 1, and one has to elect a real number x from all these preferences. One can do this by an infinite electoral version of "Twenty Questions": one asks all the voters whether x should be greater than $1/2$ or not, and uses p to determine what the answer should be; then, if x is to be greater than $1/2$, one asks whether x should be greater than $3/4$, and so on and so forth. This eventually determines x uniquely; the properties $1-4$ of the ultrafilter can be used to derive properties $1-3$ of the p-limit.

A modification of the above argument also lets us take p-limits of any sequence in a compact metric space (or slightly more generally, in any compact Hausdorff first-countable topological space[12]). These p-limits then behave in the expected manner with respect to operations in those categories, such as composition with continuous functions or with direct sum. As for unbounded real-valued sequences, one can still extract a p-limit as long as one works in a suitable compactification of the reals, such as the extended real line.

The reconstruction of p-limits from the ultrafilter p is also analogous to how, in probability theory, the concept of expected value of a (say) non-negative random variable X can be reconstructed from the concept of probability via the integration formula $\mathbf{E}(X) = \int_0^\infty \mathbf{P}(X \geq \lambda)\, d\lambda$. Indeed, one can define $p-\lim x_n$ to be the supremum of all numbers x such that the assertion $x_n > x$ is p-true, or the infimum of all numbers y that $x_n < y$ is p-true.

We have said all these wonderful things about non-principal ultrafilters, but we haven't actually shown that these amazing objects actually exist. There is a good reason for this - the existence of non-principal ultrafilters requires the axiom of choice (or some slightly weaker versions of this axiom, such as the boolean prime ideal theorem). Let's give two quick proofs of the existence of a non-principal ultrafilter:

Proof 1. Let q be the set of all cofinite subsets of the natural numbers (i.e. sets whose complement is finite). This is clearly a filter which is *proper* (i.e. it does not contain the empty set $\emptyset$). Since the union of any chain of proper filters is again a proper filter, we see from Zorn's lemma that q is contained in a maximal proper filter p. It is not hard to see that p must then be a non-principal ultrafilter. □

Proof 2. Consider the StoneCech compactification $\beta\mathbf{N}$ of the natural numbers. Since $\mathbf{N}$ is not already compact, there exists an element p of this compactification which does not lie in $\mathbf{N}$. Now note that any bounded sequence x_n on the natural numbers is a bounded continuous function on $\mathbf{N}$ (since $\mathbf{N}$ is discrete) and thus, by definition of $\beta\mathbf{N}$, extends uniquely to a bounded continuous function on $\beta\mathbf{N}$, in particular one can evaluate this function at p to obtain a real number x_p. If one then defines $p-\lim x_n := x_p$ one easily verifies the properties 1-4 of a p-limit, which by the above discussion creates a non-principal ultrafilter (which by abuse of notation is also referred to as p; indeed, $\beta\mathbf{N}$ is canonically identifiable with the space of all ultrafilters). □

These proofs are short, but not particularly illuminating. A more informal, but perhaps more instructive, explanation of why non-principal ultrafilters exist can be given

[12]Note however that Urysohn's metrisation theorem implies that any compact Hausdorff first-countable space is metrisable.

as follows. In the voting theory language, our task is to design a complete and consistent voting system for an infinite number of voters. In the cases where there is near-consensus, in the sense that all but finitely many of the voters vote one way or another, the decision is clear - go with the option which is preferred by the infinite voting bloc. But what if an issue splits the electorate with an infinite number of voters on each side? Then what one has to do is make an arbitrary choice - pick one side to go with and completely *disenfranchise* all the voters on the other side, so that they will have no further say in any subsequent votes. By performing this disenfranchisement, we increase the total number of issues for which our electoral system can reach a consistent decision; basically, any issue which has the consensus of all but finitely many of those voters not yet disenfranchised can now be decided upon in a consistent (though highly unfair) manner. We now continue voting until we reach another issue which splits the remaining pool of voters into two infinite groups, at which point we have to make another arbitrary choice, and disenfranchise another infinite set of voters. Very roughly speaking, if one continues this process of making arbitrary choices "ad infinitum", then at the end of this transfinite process we eventually exhaust the (uncountable) number of issues one has to decide, and one ends up[13] with the non-principal ultrafilter. (If at any stage of the process one decided to disenfranchise all but finitely many of the voters, then one would quickly end up with a principal ultrafilter, i.e. a dictatorship.)

With this informal discussion, it is now rather clear why the axiom of choice (or something very much like that axiom) needs to play a role in constructing non-principal ultrafilters. However, one may wonder whether one really needs the full strength of an ultrafilter in applications; to return once again to the voting analogy, one usually does not need to vote on every single conceivable issue (of which there are uncountably many) in order to settle some problem; in practice, there are often only a countable or even finite number of tricky issues which one needs to put to the ultrafilter to decide upon. Because of this, many of the results in soft analysis which are proven using ultrafilters can instead be established using a "poor man's non-standard analysis" (or "pre-infinitary analysis") in which one simply does the "voter disenfranchisement" step mentioned above by hand. This step is more commonly referred to as the trick of "passing to a subsequence whenever necessary", and is particularly popular in the soft analysis approach to PDE and calculus of variations. For instance, to minimise some functional, one might begin with a minimising sequence. This sequence might not converge in any reasonable topology, but it often lies in a sequentially compact set in some weak topology (e.g. by using the sequential version of the Banach-Alaoglu theorem), and so by passing to a subsequence one can force the sequence to converge in this topology. One can continue passing to a subsequence whenever necessary to force more and more types of convergence, and can even diagonalise using the Arzela-Ascoli argument to achieve a countable number of convergences at once (this is of course the sequential Banach-Alaoglu theorem in disguise); in many cases, one gets such a strong convergence that one can then pass to the limit. Most of these types of

[13] One should take this informal argument with a grain of salt; it turns out that after one has made an infinite number of choices, the infinite number of disenfranchised groups, while individually having no further power to influence elections, can begin having some *collective* power, basically because property 2 of a filter only guarantees closure under finite intersections and not infinite intersections, and things begin to get rather complicated. At this point, I recommend abandoning the informal picture and returning to Zorn's lemma.

arguments could also be equivalently performed by selecting an ultrafilter p at the very beginning, and replacing the notions of limit by p-limit throughout; roughly speaking, the ultrafilter has performed all the subsequence-selection for you in advance, and all your sequences in compact spaces will automatically converge without the need to pass to any further subsequences. (For much the same reason, ultrafilters can be used to simplify a lot of infinitary Ramsey theory, as all the pigeonholing has been done for you in advance.) On the other hand, the "by hand" approach of selecting subsequences explicitly tends to be much more constructive (for instance, it can often be performed without any appeal to the axiom of choice), and can also be more easily converted to a quantitative "hard analysis" argument (for instance, by using the finite convergence principle from Section 2.3).

As a concrete example from my own experience, in [CoKeStTaTa2008], we had a rather severe epsilon management problem in our "hard analysis" arguments, requiring *seven* (!) very different small quantities $1 \gg \eta_0 \gg \ldots \gg \eta_6 > 0$, with each η_i extremely small compared with the previous one. (As a consequence of this and of our inductive argument, our eventual bounds, while quantitative, were extremely large, requiring a *nine*-fold iterated Knuth arrow!) This epsilon management also led to the paper being unusually lengthy (85 pages). Subsequently, (inspired by [KeMe2006]), I learnt how the use of the above "poor man's non-standard analysis" could conceal almost all of these epsilons (indeed, due to concentration-compactness one can soon pass to a limiting object in which most of the epsilons get sent to zero). Partly because of this, a later paper of myself, Visan, and Zhang[TaViZh2008] on a very similar topic, which adopted this softer approach, was significantly shorter (28 pages, although to be fair this paper also relies on an auxiliary 30-page paper[TaViZh2008b]), though to compensate for this it becomes much more difficult to extract any sort of quantitative bound from the argument.

For the purposes of non-standard analysis, one non-principal ultrafilter is much the same as any other. But it turns out that if one wants to perform additive operations on the index set n, then there is a special (and very useful) class of non-principal ultrafilters that one can use, namely the *idempotent ultrafilters*. These ultrafilters p almost recover the shift-invariance property (which, as remarked earlier, cannot be perfectly attained for ultrafilters) in the following sense: for p-almost all h, the ultrafilter p is equal to its translate $p+h$, or equivalently that $p-\lim_h p-\lim_n x_{n+h} = p-\lim_n x_n$ for all bounded sequences. (In the probability theory interpretation, in which p-limits are viewed as an expectation, this is analogous to saying that the probability measure associated to p is idempotent under convolution, hence the name). Such ultrafilters can, for instance, be used to give a short proof of Hindman's theorem[Hi1974], which is otherwise rather unpleasant to prove. There are even more special ultrafilters known as *minimal idempotent ultrafilters*, which are quite useful in infinitary Ramsey theory, but these are now rather technical and I will refer the reader to [Be2003] for details. I will note however one amusing feature of these objects; whereas "ordinary" non-principal ultrafilters require an application of Zorn's lemma (or something similar) to construct them, these more special ultrafilters require *multiple* applications of Zorn's lemma - i.e. a nested transfinite induction! Thus these objects are truly deep in the "infinitary" end of the finitary-infinitary spectrum of mathematics.

2.5.2 Non-standard models

We have now thoroughly discussed non-principal ultrafilters, interpreting them as voting systems which can extract a consistent series of decisions out of a countable number of independent voters. With this we can now discuss non-standard models of a mathematical system. There are a number of ways to build these models, but we shall stick to the most classical (and popular) construction.

Throughout this discussion we fix a single non-principal ultrafilter p. Now we make the following general definition.

Definition 2.33. Let X be any set. The *ultrapower* ${}^*X$ of X is defined to be the collection of all sequences (x_n) with entries in X, modulo the equivalence that two sequences $(x_n), (y_n)$ are considered equal if they agree p-almost surely (i.e. the statement $x_n = y_n$ is p-true).

If X is a class of "standard" objects, we shall view ${}^*X$ as the corresponding class of "nonstandard" objects. Thus, for instance, $\mathbf{R}$ is the class of standard real numbers, and ${}^*\mathbf{R}$ is the class of non-standard real (or *hyperreal*) numbers, with each non-standard real number being uniquely representable (up to p-almost sure equivalence) as an arbitrary sequence of standard real numbers (not necessarily convergent or even bounded). What one has done here is "democratised" the class X; instead of declaring a single object x in X that everyone has to work with, one allows each voter n in a countable electorate to pick his or her own object $x_n \in X$ arbitrarily, and the voting system p will then be used later to fashion a consensus as to the properties of these objects; this is why we can identify any two sets of voter choices which are p-almost surely identical. We shall abuse notation a little bit and use sequence notation (x_n) to denote a non-standard element, even though strictly speaking one should deal with equivalence classes of sequences (just like how an element of an L^p space is not, strictly speaking, a single function, but rather an equivalence class of functions that agree almost everywhere).

One can embed any class X of standard objects in its nonstandard counterpart ${}^*X$, by identifying an element x with the constant sequence $x_n := x$; thus standard objects correspond to unanimous choices of the electorate. This identification is obviously injective. On the other hand, it is rather clear that ${}^*X$ is likely to be significantly larger than X itself.

Any operation or relation on a class (or several classes) of standard objects can be extended to the corresponding class(es) of nonstandard objects, simply by working pointwise on each n separately. For instance, the sum of two non-standard real numbers (x_n) and (y_n) is simply $(x_n + y_n)$; each voter in the electorate performs the relevant operation (in this case, addition) separately. Note that the fact that these sequences are only defined p-almost surely does not create any ambiguity. Similarly, we say that one non-standard number (x_n) is less than another (y_n), if the statement $x_n < y_n$ is p-true. And so forth. There is no direct interaction between different voters (which, in view of the lack of shift invariance, is a good thing); it is only through the voting system p that there is any connection at all between all of the individual voters.

For similar reasons, any property that one can define on a standard object, can also be defined on a non-standard object. For instance, a non-standard integer $m = (m_n)$ is prime iff the statement "m_n is prime" is p-true; a non-standard function $f = (f_n)$

is continuous iff the statement "f_n is continuous" is p-true; and so forth. Basically, if you want to know anything about a non-standard object, go put your question to all the voters, and then feed the answers into the ultrafilter p to get the answer to your question. The properties 1-4 (actually, just 1-3) of the ultrafilter ensure that you will always get a consistent answer out of this.

It is then intuitively obvious that any "simple" property that a class of standard objects has, will be automatically inherited by its nonstandard counterpart. For instance, since addition is associative in the standard real numbers, it will be associative in the non-standard real numbers. Since every non-zero standard real number is invertible in the standard real numbers, so is every non-zero non-standard real number (why?). Because (say) Fermat's last theorem is true for standard natural numbers, it is true for non-standard natural numbers (why?). And so forth. Now, what exactly does "simple" mean? Roughly speaking, any statement in *first-order logic* will transfer over from a standard class to a non-standard class, as long as the statement does not itself use p-dependent terms such as "standard" or "non-standard" anywhere. One could state a formal version of this principle here, but I find it easier just to work through examples such as the ones given above to get a sense of why this should be the case.

Now the opposite is also true; any statement in first-order logic, avoiding p-dependent terms such as standard and non-standard, which is true for non-standard classes of objects, is automatically true for standard classes also. This follows just from applying the above principle to the *negation* of the statement one is interested in. Suppose for instance that one has somehow managed to prove the twin prime conjecture (say) for non-standard natural numbers. To see why this then implies the twin prime conjecture for standard natural numbers, we argue by contraduction. If the statement "the twin prime conjecture failed" was true for standard natural numbers, then it would also be true for non-standard natural numbers (it is instructive to work this out explicitly[14]), a contradiction.

That's the *transfer principle* in a nutshell; informally, everything which avoids p-dependent terminology and which is true in standard mathematics, is also true in non-standard mathematics, and vice versa. Thus the two models are *syntactically* equivalent, even if they are *semantically* rather different. So, if the two models of mathematics are equivalent, why bother working in the latter, which looks much more complicated? It is because in the non-standard model one acquires some additional useful adjectives, such as "standard". Some of the objects in one's classes are standard, and others are not. One can use this new adjective (and some others which we will define shortly) to perform manipulations in the non-standard universe which have no obvious counterpart in the standard universe. One can then hope to use those manipulations to eventually end up at a non-trivial new theorem in the standard world, either by arriving at a statement in the non-standard world which no longer uses adjectives such as "standard" and can thus be fed into the transfer principle, or else by using some other principles (such as the overspill principle) to convert a non-standard statement involving p-dependent adjectives into a standard statement. It's similar to how, say, one can find a real root of a real polynomial by embedding the real numbers in the complex numbers, performing

[14]In order to avoid some conceptual issues regarding non-standard set theory, I recommend using this formulation of the twin prime conjecture: for every integer N, there exists a prime $p > N$ such that $p+2$ is also prime.

some mathematical manipulations in the complex domain, and then verifying that the complex-valued answer one gets is in fact real-valued.

Let's give an example of a non-standard number. Let ω be the non-standard natural number (n), i.e. the sequence $0,1,2,3,\ldots$ (up to p-almost sure equivalence, of course). This number is larger than any standard number; for instance, the standard number 5 corresponds to the sequence $5,5,5,\ldots$; since n exceeds 5 for all but finitely many values of n, we see that $n > 5$ is p-true and hence $\omega > 5$. More generally, let us say that a non-standard number is *limited* if its magnitude is bounded by a standard number, and *unlimited* otherwise; thus ω is unlimited. The notion of "limited" is analogous to the notion of being $O(1)$ discussed earlier, but unlike the $O()$ notation, there are no implicit quantifiers that require care to manipulate (though as we shall see shortly, the difficulty has not gone away completely).

One also sees, for instance, that 2ω is larger than the sum of ω and any limited number, that ω^2 is larger than the product of ω with any limited number, and so forth. It is also clear that the sum or product of any two limited numbers is limited. The number $1/\omega$ has magnitude smaller than any positive standard real number and is thus considered to be an *infinitesimal*. Using p-limits, we quickly verify that every limited number x can be uniquely expressed as the sum of a standard[15] number $\mathrm{st}(x)$ and an infinitesimal number $x - \mathrm{st}(x)$. The set of standard numbers, the set of limited numbers, and the set of infinitesimal numbers are all subrings of the set of all non-standard numbers. A non-zero number is infinitesimal if and only if its reciprocal is unlimited.

Now at this point one might be suspicious that one is beginning to violate some of the axioms of the natural numbers or real numbers, in contradiction to the transfer principle alluded to earlier. For instance, the existence of unlimited non-standard natural numbers seems to contradict the well-ordering property: if one defines $S \subset {}^*\mathbf{N}$ to be the set of all unlimited non-standard natural numbers, then this set is non-empty, and so the well-ordering property should then provide a minimal unlimited non-standard number $\inf(S) \in {}^*\mathbf{N}$. But then $\inf(S) - 1$ must be unlimited also, a contradiction. What's the problem here?

The problem here is rather subtle: a set of non-standard natural numbers is not quite the same thing as a non-standard set of natural numbers. In symbols: if $2^X := \{A : A \subset X\}$ denotes the power set of X, then $2^{{}^*\mathbf{N}} \not\equiv {}^*(2^{\mathbf{N}})$. Let's look more carefully. What is a non-standard set $A \in {}^*(2^{\mathbf{N}})$ of natural numbers? This is basically a sequence (A_n) of sets of natural numbers, one for each voter. Any given non-standard natural number $m = (m_n)$ may belong to A or not, depending on whether the statement $m_n \in A_n$ is p-true or not. We can collect all the non-standard numbers m which do belong in A, and call this set $\tilde{A}$; this is thus an element of $2^{{}^*\mathbf{N}}$. The map $A \mapsto \tilde{A}$ from ${}^*(2^{\mathbf{N}})$ to $2^{{}^*\mathbf{N}}$ turns out to be injective (why? this is the transferred axiom of extensionality), but it is not surjective; there are some sets of non-standard natural numbers which are not non-standard sets of natural numbers, and as such the well-ordering principle, when transferred over from standard mathematics, does not apply to them. This subtlety is all rather confusing at first, but a good rule of thumb is that as long as your set (or function, or whatever) is not defined using p-dependent terminology such as "standard"

[15] The map $x \mapsto \mathrm{st}(\log_\omega x)$, by the way, is a homomorphism from the semiring of non-standard positive reals to the tropical semiring $(\mathbf{R}, \min, +)$, and thus encodes the correspondence principle between ordinary rings and tropical rings.

or "limited", it will be a non-standard set (or a non-standard function, etc.); otherwise it will merely[16] be a set of non-standard objects (or a function from one non-standard set to another, etc.).

It is worth comparing the situation here with that with the $O()$ notation. With $O()$, the axiom schema of specification is simply inapplicable; one cannot form a set using O() notation inside the definition (though I must admit that I have occasionally been guilty of abusing notation and violating the above rule in my own papers). In non-standard analysis, in contrast, one *can* use terminology such as "limited" to create sets of non-standard objects, which then enjoy some useful structure (e.g. the set of limited numbers is a ring). It's just that these sets are not themselves non-standard, and thus not subject to the transfer principle.

2.5.3 Example: calculus via infinitesimals

Historically, one of the original motivations of non-standard analysis was to make rigorous the manipulations of infinitesimals in calculus. While this is not the main focus of my post here, I will give just one small example of how non-standard analysis is applied in differential calculus. If x and y are two non-standard real numbers, with y positive, we write $x = o(y)$ if x/y is infinitesimal. The key lemma is

Lemma 2.34. *Let $f : \mathbf{R} \to \mathbf{R}$ be a standard function, and let x, L be standard real numbers. We identify f with a non-standard function in the usual manner. Then the following are equivalent:*

1. *f is differentiable at x with derivative $f'(x) = L$.*
2. *For any infinitesimal h, we have $f(x+h) = f(x) + hf'(x) + o(|h|)$.*

This lemma looks very similar to linear Taylor expansion, but note that there are no limits involved (despite the suggestive $o()$ notation); instead, we have the concept of an infinitesimal. The implication of (2) from (1) follows easily from the definition of derivative, the transfer principle, and the fact that infinitesimals are smaller in magnitude than any positive standard real number. The implication of (1) from (2) can be seen by contradiction; if f is *not* differentiable at x with derivative L, then (by the axiom of choice) there exists a sequence h_n of standard real numbers going to zero, such that the Newton quotient $(f(x+h_n) - f(x))/h_n$ is bounded away from L by a standard positive number. One now forms the non-standard infinitesimal $h = (h_n)$ and obtains a contradiction to (2).

Using this equivalence, one can now readily deduce the usual laws of differential calculus, e.g. the chain rule, product rule, and mean value theorem; the proofs are algebraically almost identical to the usual proofs (especially if one rewrites those proofs in $o()$ notation), but one does not need to deal explicitly with epsilons, deltas, and limits (the ultrafilter has in some sense already done all that for you). The epsilon management is done invisibly and automatically; one does not need to keep track of whether one has to choose epsilon first before selecting delta, or vice versa. In particular, most

[16] The situation here is similar to that with the adjective "constructive"; not every function from the constructive numbers to the constructive numbers is itself a constructive function, and so forth.

of the existential quantifiers ("... there exists ε such that ...") have been eliminated, leaving only the more pleasant universal quantifiers ("for every infinitesimal h ...").

There is one caveat though: Lemma 2.34 only works when x is *standard*. For instance, consider the standard function $f(x) := x^2 \sin(1/x^3)$, with the convention $f(0) = 0$. This function is everywhere differentiable, and thus extending to non-standard numbers we have $f(x+h) = f(x) + hf'(x) + o(|h|^2)$ for all standard x and infinitesimal h. However, the same claim is not true for arbitrary non-standard x; consider for instance what happens if one sets $x = -h$.

One can also obtain an analogous characterisation of the Riemann integral: a standard function f is Riemann integrable on an interval $[a,b]$ with integral A if and only if one has

$$A = \sum_{1 \leq i < n} f(x_i^*)(x_{i+1} - x_i) + o(1)$$

for any non-standard sequence

$$a = x_1 \leq x_1^* \leq x_2 \leq \ldots \leq x_{n-1} \leq x_{n-1}^* \leq x_n = b$$

with $\sup_{1 \leq i < n}(x_{i+1} - x_i)$ infinitesimal. One can then reprove the usual basic results, such as the fundamental theorem of calculus, in this manner; basically, the proofs are the same, but the limits have disappeared, being replaced by infinitesimals.

2.5.4 Big $O()$ notation

Big $O()$ notation[17] in standard mathematics can be translated easily into the non-standard setting, as follows.

Lemma 2.35. *Let $f : \mathbf{N} \to \mathbf{C}$ and $g : \mathbf{N} \to \mathbf{R}^+$ be standard functions (which can be identified with non-standard functions in the usual manner). Then the following are equivalent.*

1. *$f(m) = O(g(m))$ in the standard sense, i.e. there exists a standard positive real constant C such that $|f(m)| \leq Cg(m)$ for all standard natural numbers n.*
2. *$|f(m)|/g(m)$ is limited for every non-standard natural number m.*

This lemma is proven similarly to Lemma 2.34; the implication of (2) from (1) is obvious from the transfer principle, while to the implication of (2) from (1) is again by contradiction, converting a sequence of increasingly bad counterexamples to (1) to a counterexample to (2). Lemma 2.35 is also a special case of the "overspill principle" in non-standard analysis, which asserts that a non-standard set of numbers which contains arbitrarily large standard numbers, must also contain an unlimited non-standard number (thus the large standard numbers "spill over" to contain some non-standard numbers). The proof of the overspill principle is related to the (specious) argument discussed above in which one tried to derive a contradiction from the set of unlimited natural numbers, and is left as an exercise.

[17] In some texts, the notation $f = O(g)$ only requires that $|f(m)| \leq Cg(m)$ for all *sufficiently large* m. The nonstandard counterpart to this is the claim that $|f(m)|/g(m)|$ is limited for every *unlimited* non-standard m.

Because of the above lemma, it is now natural to define the non-standard counterpart of the $O()$ notation: if x, y are non-standard numbers with y positive, we say that $x = O(y)$ if $|x|/y$ is limited. Then the above lemma says that the standard and non-standard $O()$ notations agree for standard functions of one variable. Note how the non-standard version of the$O()$ notation does not have the existential quantifier ("... there exists C such that ...") and so the epsilon management is lessened. If we let $\mathscr{L}$ denote the subring of ${}^*\mathbf{R}$ consisting of all limited numbers, then the claim $x = y + O(z)$ can be rewritten as $x = y \bmod z\mathscr{L}$, thus we see how the $O()$ notation can be viewed algebraically as the operation of quotienting the (non-standard) real numbers by various dilates of the subring $\mathscr{L}$.

One can convert many other order-of-magnitude notions to non-standard notation. For instance, suppose one is performing some standard hard analysis involving some large parameter $N > 1$, e.g. one might be studying a set of N points in some group or Euclidean space. One often wants to distinguish between quantities which are of polynomial size in N and those which are super-polynomial in size; for instance, these N points might lie in a finite group G, where G has size much larger than N, and one's application is such that any bound which depends on the size of G will be worthless. Intuitively, the set of quantities which are of polynomial size in N should be closed under addition and multiplication and thus form a sort of subring of the real numbers, though in the standard universe this is difficult to formalise rigorously. But in non-standard analysis, it is not difficult: we make N non-standard (and G too, in the above example), and declare any non-standard quantity x to be of polynomial size if we have $x = O(N^{O(1)})$, or equivalently if $\log(1+|x|)/\log N$ is limited. We can then legitimately form the set $\mathscr{P}$ of all non-standard numbers of polynomial size, and this is in fact a subring of the non-standard real numbers; as before, though, we caution that $\mathscr{P}$ is not a non-standard set of reals, and in particular is not a non-standard subring of the reals. But since $\mathscr{P}$ *is* a ring, one can then legitimately apply whatever results from ring theory one pleases to P, bearing in mind though that any sets of non-standard objects one generates using that theory may not necessarily be non-standard objects themselves. At the end of the day, we then use the transfer principle to go back to the original problem in which N is standard.

As a specific example of this type of thing from my own experience, in [TaVu2007], we had a large parameter n, and had at some point to introduce the somewhat fuzzy notion of a "highly rational number", by which we meant a rational number a/b whose numerator and denominator were both at most $n^{o(n)}$ in magnitude. Such numbers looked like they were forming a field, since the sum, difference, product, or quotient of two highly rational numbers was again highly rational (but with a slightly different rate of decay in the $o()$ notation). Intuitively, one should be able to do any algebraic manipulation on highly rational numbers which is legitimate for true fields (e.g. using Cramer's rule to invert a non-singular matrix) and obtain an output which is also highly rational, as long as the number of algebraic operations one uses is $O(1)$ rather than, say, $O(n)$. We did not actually formalise this rigorously in our standard notation, and instead resorted to informal English sentences to describe this; but one can do everything perfectly rigorously in the non-standard setting by letting n be non-standard, and defining the field F of non-standard rationals a/b where $a, b = O(n^{o(n)})$; F is genuinely a field of non-standard rationals (but not a non-standard field of rationals), and so using Cramer's

rule here (but only for matrices of standard size) would be perfectly legitimate. (We did not actually write our argument in this non-standard manner, keeping everything in the usual standard hard analysis setting, but it would not have been difficult to rewrite the argument non-standardly, and there would be some modest simplifications.)

2.5.5 A hierarchy of infinitesimals

We have seen how, by selecting an ultrafilter p, we can extend the standard real numbers $\mathbf{R}$ to a larger system ${}^*\mathbf{R}$, in which the original number system $\mathbf{R}$ becomes a real totally ordered subfield. (Exercise: is $\mathbf{R}$ complete? The answer depends on how one defines one's terms.) This gives us some new objects, such as the infinitesimal η_0 given by the sequence $1, 1/2, 1/4, 1/8, \ldots$. This quantity is smaller than any standard positive number, in particular it is infinitesimally smaller than any quantity depending (via standard operations) on standard constants such as 1. One may think of ${}^*\mathbf{R}$ as the non-standard extension of $\mathbf{R}$ generated by adjoining η_0; this is similar to the field extension $\mathbf{R}(\eta_0)$, but is much larger, because field extensions are only closed under arithmetic operations, whereas non-standard extensions are closed under *all* definable operations. For instance, $\exp(1/\eta_0)$ lies in ${}^*\mathbf{R}$ but not in $\mathbf{R}(\eta_0)$.

Now it is possible to iterate this process, by introducing an non-standard ultrafilter ${}^*p$ on the non-standard natural numbers ${}^*\mathbf{N}$, and then embedding the field ${}^*\mathbf{R}$ inside an even larger system ${}^{**}\mathbf{R}$, whose elements can be identified (modulo ${}^*p$-almost sure equivalence) with non-standard sequences (x_n) of non-standard numbers in ${}^*\mathbf{R}$ (where n now ranges over the non-standard natural numbers ${}^*\mathbf{N}$); one could view these as "doubly non-standard numbers". This gives us some "even smaller" infinitesimals, such as the "doubly infinitesimal" number η_1 given by the non-standard sequence $1, \eta_0, \eta_0^2, \eta_0^3, \ldots$. This quantity is smaller than any standard or (singly) non-standard number, in particular infinitesimally smaller than any positive quantity depending (via standard or singly non-standard operations) on standard or singly non-standard constants such as 1 or η_0. For instance, it is smaller than $1/A(\lfloor 1/\eta_0 \rfloor)$, where A is the Ackermann function, since the sequence that defines η_1 is indexed over the non-standard natural numbers and η_0^n will drop below $1/A(\lfloor 1/\eta_0 \rfloor)$ for sufficiently large non-standard n.

One can continue in this manner, creating a triply infinitesimal quantity η_2 which is infinitesimally smaller than anything depending on 1, η_0, or η_1, and so forth. Indeed one can iterate this construction an absurdly large number of times, though in most applications one only needs an explicitly finite number of elements from this hierarchy. Having this hierarchy of infinitesimals, each one of which is guaranteed to be infinitesimally small compared to *any* quantity formed from the preceding ones, is quite useful: it lets one avoid having to explicitly write a lot of epsilon-management phrases such as "Let η_2 be a small number (depending on η_0 and η_1) to be chosen later" and "... assuming η_2 was chosen sufficiently small depending on η_0 and η_1", which are very frequent in hard analysis literature, particularly for complex arguments which involve more than one very small or very large quantity. (The paper [CoKeStTaTa2008] referred to earlier is of this type.)

2.5.6 Conclusion

I hope I have shown that non-standard analysis is not a totally "alien" piece of mathematics, and that it is basically only "one ultrafilter away" from standard analysis. Once one selects an ultrafilter, it is actually relatively easy to swap back and forth from the standard universe and the non-standard one (or to doubly non-standard universes, etc.). This allows one to rigorously manipulate things such as "the set of all small numbers", or to rigorously say things like "η_1 is smaller than anything that involves η_0", while greatly reducing epsilon management issues by automatically concealing many of the quantifiers in one's argument. One has to take care as to which objects are standard, non-standard, sets of non-standard objects, etc., especially when transferring results between the standard and non-standard worlds, but as long as one is clearly aware of the underlying mechanism used to construct the non-standard universe and transfer back and forth (i.e. as long as one understands what an ultrafilter is), one can avoid difficulty. The main drawbacks to use of non-standard notation (apart from the fact that it tends to scare away some of your audience) is that a certain amount of notational setup is required at the beginning, and that the bounds one obtains at the end are rather ineffective (though, of course, one can always, after painful effort, translate a non-standard argument back into a messy but quantitative standard argument if one desires).

2.5.7 Notes

This article was originally posted on Jun 25, 2007 at

`terrytao.wordpress.com/2007/06/25`

Theo Johnson-Freyd, noted that the use of ultrafilters was not completely identical to the trick of passing to subsequences whenever necessary; for instance, there exist ultrafilters with the property that not every sequence in a compact set (e.g. $[0,1]$) admits a large convergent subsequence. (Theo learned about this observation from Ken Ross.)

Eric Wofsey, answering a question of Theo, pointed out that, thanks to a cardinality argument, there exist a pair of non-principal ultrafilters which are not permutations of each other, despite the fact that one non-principal ultrafilter tends to be just as good as any other for most applications. On the other hand, if one assumes the continuum hypothesis, then any two ultrapowers (chosen using different ultrafilters of $\mathbf{N}$) of a structure with at most the cardinality of the continuum and with a countable language are isomorphic (meaning that there is a bijection between the ultrapowers that preserves all interpretations of the language symbols), although in many applications in non-standard analysis one needs to take extensions of uncountably many objects and so this equivalence does not always apply.

Alejandro Rivero pointed out Connes' non-commutative variant of non-standard analysis, which used compact operators for infinitesimals and avoided the use of ultrafilters (or the axiom of choice), though as a consequence the transfer principle was not fully present.

Michael Greineckerer pointed out that an explicit link between Arrow's theorem and ultrafilters appears in the papers [KiSo1972], [Ha1976].

Thanks to Liu Xiao Chuan for corrections.

2.6 Dyadic models

One of the oldest and most fundamental concepts in mathematics is the *line*. Depending on exactly what mathematical structures we want to study (algebraic, geometric, topological, order-theoretic, etc.), we model lines nowadays by a variety of standard mathematical objects, such as the real line $\mathbf{R}$, the complex line $\mathbf{C}$, the projective line $\mathbb{RP}^1$, the extended real line $[-\infty,+\infty]$, the affine line $\mathbb{A}^1$, the continuum c, the long line L, etc. We also have discrete versions of the line, such as the natural numbers $\mathbf{N}$, the integers $\mathbf{Z}$, and the ordinal ω, as well as compact versions of the line, such as the unit interval $[0,1]$ or the unit circle $\mathbf{T} := \mathbf{R}/\mathbf{Z}$. Finally we have discrete *and* compact versions of the line, such as the cyclic groups $\mathbf{Z}/N\mathbf{Z}$ and the discrete intervals $\{1,\ldots,N\}$ and $\{0,\ldots,N-1\}$. By taking Cartesian products we then obtain higher-dimensional objects such as Euclidean space $\mathbf{R}^n$, the standard lattice $\mathbf{Z}^n$, the standard torus $\mathbf{T}^n = \mathbf{R}^n/\mathbf{Z}^n$, and so forth. These objects of course form the background on which a very large fraction of modern mathematics is set.

Broadly speaking, the line has three major families of structures on it:

1. *Geometric structures*, such as a metric or a measure, completeness, scales (coarse and fine), rigid motions (translations and reflection), similarities (dilation, affine maps), and differential structures (tangent bundle, etc.);

2. *Algebraic structures*, such group, ring, or field structures, and everything else that comes from those categories (e.g. subgroups, homomorphisms, involutions, etc.); and

3. *One-dimensional structures*, such as order, a length space structure (in particular, path-connectedness structure), a singleton generator, the Archimedean property, the ability to use mathematical induction (i.e. well-ordering), convexity, or the ability to disconnect the line by removing a single point.

Of course, these structures are inter-related, and it is an important phenomenon that a mathematical concept which appears to be native to one structure, can often be equivalently defined in terms of other structures. For instance, the absolute value $|n|$ of an integer n can be defined *geometrically* as the distance from 0 to n, *algebraically* as the index of the subgroup $\langle n\rangle = n\cdot\mathbf{Z}$ of the integers $\mathbf{Z}$ generated by n, or *one-dimensionally* as the number of integers between 0 and n (including 0, but excluding n). This equivalence of definitions becomes important when one wants to work in more general contexts in which one or more of the above structures is missing or otherwise weakened.

What I want to talk about today is an important toy model for the line (in any of its incarnations), in which the geometric and algebraic structures are enhanced (and become neatly nested and recursive), at the expense of the one-dimensional structure (which is largely destroyed). This model has many different names, depending on what field of mathematics one is working in and which structures one is interested in. In harmonic analysis it is called the *dyadic model*, the *Walsh model*, or the Cantor group model; in number theory and arithmetic geometry it is known as the *function field model*; in topology it is the *Cantor space model*; in probability it is the *martingale*

model; in metric geometry it is the *ultrametric*, *tree*, or *non-Archimedean* model; in algebraic geometry it is the *Puiseux series model*; in additive combinatorics it is the *bounded torsion* or *finite field model*; in computer science and information theory it is the *Hamming cube model*; in representation theory it is the *Kashiwara crystal model*; and so forth. Let me arbitrarily select one of these terms, and refer to all of these models as *dyadic models* for the line (or of objects derived from the line). While there is often no direct link between a dyadic model and a non-dyadic model, dyadic models serve as incredibly useful laboratories in which to gain insight and intuition for the "real-world" non-dyadic model, since one has much more powerful and elegant algebraic and geometric structure to play with in this setting (though the loss of one-dimensional structure can be a significant concern). Perhaps the most striking example of this is the three-line proof of the Riemann hypothesis in the function field model of the integers, which I will discuss a little later.

2.6.1 Dyadic integers and reals

Very broadly speaking, one of the key advantages that dyadic models offer over non-dyadic models is that they do not have any "spillover" from one scale to the next. This spillover is introduced to us all the way back in primary school, when we learn about the algorithms for decimal notation arithmetic: long addition, long subtraction, long multiplication, and long division. In decimal notation, the notion of scale is given to us by powers of ten (with higher powers corresponding to coarse scales, and lower powers to fine scales), but in order to perform arithmetic properly in this notation, we must constantly "carry" digits from one scale to the next coarser scale, or conversely to "borrow" digits from one scale to the next finer one. These interactions between digits from adjacent scales (which in modern terminology, would be described as iterated cocycles over the base space $\mathbf{Z}/10\mathbf{Z}$) make the arithmetic operations look rather complicated in decimal notation, although one can at least isolate the fine-scale behaviour from the coarse-scale digits (but not vice versa) through modular arithmetic. (To put it a bit more algebraically, the integers or real numbers can quotient out the coarse scales via normal subgroups (or ideals) such as $N \cdot \mathbf{Z}$, but do not have a corresponding normal subgroup or ideal to quotient out the fine scales.)

It is thus natural to look for models of arithmetic in which this spillover is not present. One is first exposed to such models in high school, when the arithmetic of polynomials in one unknown t is introduced (i.e. one works with rings such as $\mathbf{Z}[t]$ or $\mathbf{R}[t]$ rather than $\mathbf{Z}$ or $\mathbf{R}$). For instance, to quotient one polynomial by another, one uses the polynomial long division (or *synthetic division*) algorithm, which is formally identical to long division for integers in decimal notation but without all the borrowing from one scale to the next. Here scales are represented by powers of t, rather than powers of 10. As with the reals or integers, the coarse scales can be contained in a normal subgroups (and ideals) such as $t^d \cdot \mathbf{R}[t]$, but now the fine scales can *also* be contained[18] in normal subgroups (though not ideals) such as $\langle 1, t, \ldots, t^{d-1} \rangle$, the group generated by $1, t, \ldots, t^{d-1}$ (i.e. the group of polynomials of degree less than d).

[18] From a homological algebra perspective, things are better here because various short exact sequences involving the scales are now split.

Now, polynomial rings such as $\mathbf{Z}[t]$ or $\mathbf{R}[t]$ are a bit "too big" to serve as models for $\mathbf{Z}$ or $\mathbf{R}$ (unless one adjoins some infinitesimals, as in Section 2.5, but that's another story), as they have one more dimension. One can get a more accurate model by considering the decimal representation again, which identifies natural numbers as polynomials over the space of digits $\{0,1,\ldots,9\}$. This space is not closed under addition (which is what causes spillover in the first place); but we can remedy this by replacing this space of digits with the cyclic group $\mathbf{Z}/10\mathbf{Z}$. This gives us the model $(\mathbf{Z}/10\mathbf{Z})[t]$ for the integers; this is the decimal representation without the operation of carrying. If we follow the usual decimal notation and identify polynomials in $(\mathbf{Z}/10\mathbf{Z})[t]$ with strings of digits in the usual manner (e.g. identifying $3t+2$ with 32) then we obtain a number system which is similar, but not quite identical, to the integers. For instance, $66+77$ now equals 33 rather than 143; $25*4$ now equals 80 rather than 100; and so forth. Note that unlike the natural numbers, the space of polynomials is already closed under negation and so there is no need to introduce negative numbers; for instance, in this system we have $-12 = 98$. I'll refer to $(\mathbf{Z}/10\mathbf{Z})[t]$ as the "base 10 dyadic" model for the integers (somewhat annoyingly, the term "10-adic" is already taken to mean something slightly different).

There is also a base 10 dyadic model for the real numbers, in which we allow infinitely many negative powers of t but only finitely many positive powers in t; in other words, the model is $(\mathbf{Z}/10\mathbf{Z})((1/t))$, the ring of formal Laurent series in $1/t$. This ring again differs slightly from the reals; for instance, $0.999\ldots$ is now no longer equal to $1.000\ldots$ (in fact, they differ by $1.111\ldots$). So the decimal notation maps $(\mathbf{Z}/10\mathbf{Z})((1/t))$ onto the positive real axis $\mathbf{R}^+$, but there is a small amount of non-injectivity caused by this map.

The base 10 dyadic models for the reals and integers are not particularly accurate, due to the presence of zero divisors in the underlying base ring $\mathbf{Z}/10\mathbf{Z}$. For instance, we have $2 \times 5 = 0$ in this model. One can do a lot better by working over a finite field F, such as the field $\mathbf{F}_2$ of two elements. This gives us dyadic models $F[t]$ and $F((1/t))$ for the integers and reals respectively which turn out to be much closer analogues than the base 10 model. For instance, $F[t]$, like the integers, is a Euclidean domain, and $F((1/t))$ is a field. (In the binary case $F = \mathbf{F}_2$, the addition operation is just bitwise XOR, and multiplication is bitwise convolution.) We can also model many other non-dyadic objects, as the following table illustrates:

Non-dyadic	Dyadic
Integers $\mathbf{Z}$	Polynomials $F[t]$
Rationals $\mathbf{Q}$	Rational functions $F(t)$
Reals $\mathbf{R}$	Laurent polynomials $F((1/t))$
Unit circle $\mathbf{R}/\mathbf{Z}$	$F((1/t))/F[t] \equiv \frac{1}{t}F[\frac{1}{t}]$
$\|F\|^d \cdot \mathbf{Z}$	$t^d \cdot F[t]$
Cyclic group $\mathbf{Z}/\|F\|^d \cdot \mathbf{Z}$	Vector space F^d
Finite field $\mathbf{Z}/p \cdot \mathbf{Z}$	Finite field $F[t]/p(t) \cdot F[t]$
Absolute value	(Exponential of) degree
Plane wave	Walsh function
Wavelet	Haar wavelet
Gaussian	Step function
Ball	Dyadic interval
Heat operators	Martingale conditional expectations
Band-limited	Locally constant
Interval / arithmetic progression	Subspace / subgroup
Bohr set	Hyperplane

Recall that we can define the absolute value (or norm) of an integer n as the index of the subgroup $\langle n \rangle$ of the integers. Exactly the same definition can be applied to the dyadic model $F[t]$ of the integers; the absolute value of an element $n \in F[t]$ can then be seen to equal $|n| = |F|^{\deg(n)} \in \mathbf{Z}^+$, where $\deg(n)$ is the degree of t in n (with the convention that 0 has a degree of $-\infty$ and thus an absolute value of 0). For instance, in the binary case, t^3+t+1 (or 1011) has a norm of 8. Like the absolute value on the integers, the absolute value on the dyadic model $F[t]$ of the integers is multiplicative and obeys the triangle inequality, giving rise to a metric on $F[t]$ by the usual formula $d(n,m) := |n-m|$. In fact, we have something better than a metric, namely an *ultrametric*; in the dyadic world, the triangle inequality

$$d(x,z) \leq d(x,y) + d(y,z)$$

can be strengthened to the *ultratriangle inequality*

$$d(x,z) \leq \max(d(x,y), d(y,z)).$$

One can then uniquely extend this absolute value multiplicatively to the dyadic model $F((1/t))$ of the reals, which is given by the same formula $|n| = |F|^{\deg(n)} \in \mathbf{R}^+$, where $\deg(n)$ is now understood to be the highest exponent of t which appears in the expansion of n (or $-\infty$ of no such exponent appears). Thus for instance in the binary case $1/t + 1/t^2 + 1/t^3 + \ldots$ (or $0.111\ldots$) has a norm of 1/2. Just as with the real line, this absolute value turns the dyadic real line $F((1/t))$ into a complete metric space. The metric space then generates balls $B(x,r) := \{y \in F((1/t)) : |y-x| < r\}$, which in the binary case are identifiable with *dyadic intervals*. The fact that we have an ultrametric instead of a metric means that the balls enjoy a very useful *nesting property*, which is unavailable in the non-dyadic setting: if two balls intersect, then the larger one must necessarily contain the smaller one.

On the other hand, most of the "one-dimensional" structure of the real line is lost when one passes to the dyadic model. For instance, the dyadic real line is still locally

compact, but not locally connected; the topology is instead locally that of a Cantor space. There is no natural notion of order on the dyadic integers or real line, and the metric is non-Archimedean. Related to this, mathematical induction no longer applies to the dyadic integers. Nevertheless, and somewhat counter-intuitively, one can go remarkably far in mimicking many features of the integers and real numbers without using any one-dimensional structure. I'll try to illustrate this in a number of contexts.

2.6.2 Dyadic models in harmonic analysis

Let us first compare the harmonic analysis of the dyadic and non-dyadic models. Lebesgue measure dx is the unique Haar measure on the real line which assigns a measure of 1 to the unit interval $[0,1]$. Similarly, the dyadic real line $F((1/t))$, there is a unique Haar measure dx which assigns a measure of 1 to the unit ball $B(0,1)$. Indeed, this measure can be defined by pulling back Lebesgue measure on the positive real axis $\mathbf{R}^+$ via the decimal map which maps elements of $F((1/t))$ to the corresponding base $|F|$ expansion in the reals (e.g. in the binary case, $t^2 + 1/t$ would be mapped to $100.1_2 = 4.5$).

The general theory of harmonic analysis on locally compact abelian groups then shows that there is a theory of the Fourier transform on the dyadic real line $F((1/t))$, which turns out to be closely analogous to that on the non-dyadic real line $\mathbf{R}$. (There is also a Fourier theory relating the dyadic integers $F[t]$ with the dyadic unit circle $F((1/t))/F[t] \equiv \frac{1}{t} \cdot F[\frac{1}{t}]$, which we leave to the reader.) Recall that the Fourier transform on the real line is built out of the 1-periodic character $e : \mathbf{R} \to \mathbf{C}$ defined by $e(x) := e^{2\pi i x}$ by the formula

$$\hat{f}(\xi) := \int_{\mathbf{R}} f(x) e(-x\xi)\, dx$$

for all well-behaved $f : \mathbf{R} \to \mathbf{C}$ (e.g. absolutely integrable f). Similarly, the Fourier transform on $F((1/t))$ (assuming F to be a prime field $F = \mathbf{F}_p \equiv \mathbf{Z}/p\mathbf{Z}$ for simplicity) can be built out of the 1-periodic character $e_p : F((1/t)) \to \mathbf{C}$ defined by

$$e_p(\sum_{j=-\infty}^{d} a_j t^j) := e(a_{-1}/p)$$

(which would be a square wave in the binary case) using almost exactly the same formula, namely

$$\hat{f}(\xi) := \int_{F((1/t))} f(x) e_p(-x\xi)\, dx$$

for all well-behaved $f : F((1/t)) \to \mathbf{C}$. One can then show that this dyadic Fourier transform (known as the *Walsh-Fourier transform* in the binary case) enjoys all the usual algebraic properties that the non-dyadic Fourier transform does - for instance, it reacts with convolution, translation, modulation, and dilation in exactly the same way as its non-dyadic counterpart, and also enjoys a perfect analogue of Plancherel's theorem. (It also has a more pleasant fast Fourier transform algorithm than its non-dyadic counterpart, as one no longer needs the additional step of taking care of the

spillover from one scale to the next.) In fact, the dyadic structure makes the harmonic analysis on $F((1/t))$ somewhat simpler than that on $\mathbf{R}$, because of the ability to have *perfect phase space localisation*. In the real line, it is well-known that a function and its Fourier transform cannot simultaneously be compactly supported without vanishing completely (because if a function was compactly supported, then its Fourier transform would be a real analytic function, which cannot be compactly supported without vanishing completely, due to analytic continuation). However, analytic continuation is a highly "one-dimensional" property (among other things, it exploits connectedness). Furthermore, it is not a robust property, and it is possible to have functions f on the real line such that f and its Fourier transform are "almost compactly supported", or more precisely *rapidly decreasing*; the Gaussian function $f(x) = \exp(-\pi|x|^2)$, which is its own Fourier transform, is a particularly good example. In the dyadic world, the analogue of the Gaussian function is the step function $1_{B(0,1)}$, which is also its own Fourier transform, and thus demonstrates that it is possible for a function and its Fourier transform to both be compactly supported. More generally, it is possible for a function $f : F((1/t)) \to \mathbf{C}$ to be supported on a dyadic interval I, and for its Fourier transform to be supported on another dyadic interval J, as long as the uncertainty principle $|I||J| \geq 1$ is respected. One can use these "Walsh wave packets" (which include the *Haar wavelets* and *Radamacher functions* as special cases) to elegantly and efficiently perform time-frequency analysis in the dyadic setting. This has proven to be an invaluable model to work with before tackling the more interesting time-frequency problems in the non-dyadic setting (such as those relating to Carleson's theorem[Ca1966], or to various multilinear singular integrals), as many technical headaches (such as those involving "Schwartz tails") are absent in the dyadic setting, while the time-frequency combinatorics (which is really the heart of the matter) stays largely intact[19]. See [Pe2001], [Ta2001]

In some cases one can in fact deduce a non-dyadic harmonic analysis result directly from a dyadic one via some sort of averaging argument (or the $1/3$ *translation trick* of Michael Christ[Ch1988], which is the observation that every non-dyadic interval (in, say, $[0,1]$) is contained either in a dyadic interval of comparable size, or the $1/3$ translation of a dyadic interval of comparable size). In particular the "Bellman function" approach to harmonic analysis often proceeds via this averaging, as the Bellman function method requires a recursive dyadic structure (or a continuous heat kernel-type structure) in order to work properly. In general, though, the dyadic argument only serves as a model "road map" for the non-dyadic argument, rather than a formal component. There are only a few cases known where a dyadic result in harmonic analysis has not shown the way towards proving the non-dyadic analogue; one of these exceptions is the problem of establishing a nonlinear analogue of Carleson's theorem, which was achieved in the dyadic setting[MuTaTh2003b] but remains open in the non-dyadic setting.

[19] To give just one example, the *Shannon sampling theorem* collapses in the dyadic setting to the trivial statement that a function which is locally constant on dyadic intervals of length 2^{-n}, can be reconstructed exactly by sampling that function at intervals of 2^{-n}

2.6.3 Dyadic models in PDE

Let us now leave harmonic analysis and turn to dyadic and non-dyadic models of other parts of mathematics. I should briefly discuss PDE, which is one field in which the dyadic models have proven to have only a limited impact (though the Katz-Pavlovic dyadic model[KaPa2005] for the Euler and Navier-Stokes equations is perhaps a counterexample). This is partly because, in contrast to harmonic analysis, the analysis of PDEs *does* heavily exploit the one-dimensionality of the real line (and in particular, the time axis), for instance via the use of continuity arguments or monotonicity formulae. Nevertheless one can still obtain some partial analogues of various PDE objects, most notably those connected to the heat equation, as long as one is willing to work with dyadic notions of time. For instance, in the binary case $F = \mathbf{F}_2$, the dyadic analogue of the heat operator $e^{t\Delta}$ when $t = 2^{2n}$ is a power of 4 would be the conditional expectation operator to the σ-algebra generated by dyadic intervals of length 2^n. These conditional expectations are nested in n, yielding a martingale structure. There is in fact a very strong (and well known) analogy between heat operators and conditional expectations with respect to a martingale; to give just one example, the sharp Young's inequality on the real line can be proven by a heat flow method (ensuring that a certain multilinear expression is monotone along heat flow), and the corresponding sharp Young's inequality on $F((1/t))$ or F^n can similarly be proven (with a somewhat shorter proof) by a nested conditional expectation argument.

2.6.4 Dyadic models in additive combinatorics

We now briefly turn to additive combinatorics. Here, it is often convenient for combinatorial reasons to work not on an infinite additive group such as $\mathbf{R}$ or $\mathbf{Z}$, but in a finite additive group. In the non-dyadic setting, one usually uses a cyclic group such as $\mathbf{Z}/N\mathbf{Z}$; in the dyadic setting, one uses a vector space such as F^n (one should think of F as being fixed, e.g. $F = \mathbf{F}_2$ or $F = \mathbf{F}_3$, and n as being large). A general philosophy is that as long as these two groups have roughly the same size (i.e. $N \approx |F|^n$), then the additive combinatorics of these two groups will be broadly analogous, even if algebraically the groups are rather different (the former can be generated by a single generator, but most elements have large order, whereas the latter needs many generators, and all elements have small order). But the dyadic model tends to be significantly more tractable for a number of reasons. Most obviously, the group F^n is also a vector space, and thus one can apply the powerful tools of linear algebra. A cyclic group only has some partial (and messy) analogues of linear algebraic tools; for instance, in cyclic groups generalised arithmetic progressions are somewhat analogous to vector spaces spanned by a given set of vectors, has a lot of subgroups; dually, Bohr sets (level sets of one or more characters) play a role in cyclic groups analogous to the intersection of one or more hyperplanes in F^n. See [Gr2005b] for further discussion. Another useful feature of the group F^n is the presence of flags of subspaces, e.g. the coordinate flag

$$\{0\} = F^0 \subset F^1 \subset \ldots \subset F^n$$

which allows one in some cases to prove combinatorial facts in F^n by an induction on dimension, and to use tools related to such flags such as the combinatorial technique of

compression (see e.g. [GrTa2008e].

Very recently, though, Ben Green and I have discovered that the lack of one-dimensionality in the finite field model can make that model *less* tractable than the cyclic group model in certain technical senses. For instance, Gowers' celebrated proof[Go2001] of Szemerédi's theorem[Sz1975] does not quite work in finite field models because of this little wrinkle. This seems to have something to do with the (still poorly understood) analogy between nilsequences in the non-dyadic setting, and polynomials in the dyadic setting. Hopefully we'll have more to say about this in the future.

2.6.5 Dyadic models in algebraic combinatorics

I'll now touch briefly on the role of dyadic models in algebraic combinatorics. Here it seems that many algebraic questions that are over the field **R** or **C** collapse to purely combinatorial questions once one "tropicalises" by passing to a dyadic model field, such as the field of Puiseux series $\mathbf{C}\{t\}$. Furthermore (and rather miraculously), certain questions have *identical* answers in the dyadic and non-dyadic settings, thus allowing one to use combinatorial gadgets to solve algebraic problems. For instance, the question of understanding the possible relationships between eigenvalues of a sum $A+B$ of Hermitian matrices, with the eigenvalues of A and B separately, is a non-trivial algebraic problem; but by passing to the dyadic model of Puiseux series and extracting the leading exponents (which does not affect the final answer) it was shown by Speyer[Sp2005] that the problem collapses to the combinatorial problem of locating a *honeycomb* (see Section 1.7) between the three sets of putative eigenvalues. (This fact, as well as the one below, had been established earlier by Knutson and myself by a much more indirect method.) There is also a discrete (representation-theoretic) counterpart to this phenomenon; the question of computing tensor product multiplicities for the unitary group $U(n)$ is also a non-trivial algebraic problem, in large part due to the "mixing" between basis elements of irreducible representations when one takes tensor products and then decomposes again into irreducibles (in the case $n = 2$, this mixing is described by the Clebsch-Gordan coefficients; the situation is more complicated in higher n due to multiplicities in the decomposition). But if one replaces the notion of a representation with the "dyadic" model of a crystal representation [Ka1990], the mixing is eliminated (without affecting the multiplicities), and it was shown by Henriques and Kamnitzer [HeKa2006] that the problem of computing tensor product multiplicities again collapses to a honeycomb problem. It would be interesting to get a more general explanation of why these phenomena occur.

2.6.6 Dyadic models in number theory

Finally, I want to discuss the role of dyadic models in number theory - a topic which has in fact been the subject of at least one entire graduate text [Ro2002]. In contrast to the other settings discussed above, there is a fantastic disparity in number theory between our understanding of the dyadic model and that of the non-dyadic model; several of the most famous problems in non-dyadic number theory (e.g. Riemann hypothesis, Fermat's last theorem, twin prime conjecture, abc conjecture, factorisation of large numbers) are surprisingly easy to solve in the dyadic world, but nobody knows how

to convert the dyadic arguments to the non-dyadic setting (although the converse step of converting non-dyadic arguments to dyadic ones is usually rather straightforward). One notable exception here is the parity problem (Section 1.10, which has resisted progress in both dyadic and non-dyadic settings.

Let's now turn to the Riemann hypothesis. Classically, number theory has focused on the multiplicative structure of the ring of integers $\mathbf{Z}$. After factoring out the group of units $\{-1,+1\}$, we usually restrict attention to the positive integers $\mathbf{Z}^+$. In the dyadic model, we study the multiplicative structure of the ring $F[t]$ of polynomials for some finite field F. After factoring out the group $F^\times$ of units, we can restrict attention to the monic polynomials $F[t]_m$. As the ring of polynomials is a Euclidean domain, it has unique factorisation, and in particular every monic polynomial can be expressed uniquely (up to permutation) as the product of irreducible monic polynomials, which we shall call *prime* polynomials.We can analyse the problem of counting primes in $F[t]$ by using zeta functions, in complete analogy with the integer case. The Riemann zeta function is of course given by

$$\zeta(s) := \sum_{n\in\mathbf{Z}^+} \frac{1}{n^s}$$

(for $\mathrm{Re}(s) > 1$) and we introduce the analogous zeta function

$$\zeta_{F[t]}(s) := \sum_{n\in F[t]_m} \frac{1}{|n|^s}.$$

In the integers, unique factorisation gives the identity

$$\log n = \sum_{d|n} \Lambda(d)$$

where $\Lambda(d)$ is the von Mangoldt function, defined to equal $\log p$ when d is the power of a prime p and 0 otherwise. Taking the Mellin transform of this identity, we conclude that

$$-\frac{\zeta'(s)}{\zeta(s)} = \sum_{n\in\mathbf{Z}^+} \frac{\Lambda(n)}{n^s},$$

which is the fundamental identity linking the zeroes of the zeta function to the distribution of the primes. We can do the same thing in the dyadic case, obtaining the identity

$$-\frac{\zeta'_{F[t]}(s)}{\zeta_{F[t]}(s)} = \sum_{n\in F[t]_m} \frac{\Lambda_{F[t]}(n)}{|n|^s}, \tag{2.6}$$

where the von Mangoldt function $\Lambda_{F[t]}(n)$ for F[t] is defined as $\log|p|$ when n is the power of a prime polynomial p, and 0 otherwise.

So far, the dyadic and non-dyadic situations are very closely analogous. But now we can do something special in the dyadic world: we can compute the zeta function explicitly by summing by degree. Indeed, we have

$$\zeta_{F[t]}(s) = \sum_{d=0}^{\infty} \sum_{n\in F[t]_m:\deg(n)=d} \frac{1}{|F|^{ds}}.$$

The number of monic polynomials of degree d is $|F|^d$. Summing the geometric series, we obtain an exact formula for the zeta function:

$$\zeta_{F[t]}(s) = (1 - |F|^{s-1})^{-1}.$$

In particular, the Riemann hypothesis for F[t] is a triviality - there are clearly no zeroes whatsoever! Inserting this back into (2.6) and comparing coefficients, one soon ends up with an *exact* prime number theorem for $F[t]$:

$$\sum_{n \in F[t]_m : \deg(n) = d} \Lambda(n) = |F|^d \log |F|$$

which quickly implies that the number of prime polynomials of degree d is $\frac{1}{d}|F|^d + O(|F|^{d/2})$. (One can generalise the above analysis to other varieties over finite fields, leading ultimately to the (now-proven) Weil conjectures, which include the "Riemann hypothesis for function fields".)

Another example of a problem which is hard in non-dyadic number theory but trivial in dyadic number theory is *factorisation*. In the integers, it is not known whether a number which is n digits long can be factored (probabilistically) in time polynomial in n (the best known algorithm for large n, the *number field sieve*, takes a little longer than $\exp(O(\log^{1/3} n))$ time, according to standard heuristics); indeed, the presumed hardness of factoring underlies many popular cryptographic protocols such as RSA. However, in $F[t]$ with F fixed, a polynomial f of degree n can be factored (probabilistically) in time polynomial in n by the following three-stage algorithm:

1. Compute the gcd of f and its derivative f' using the Euclidean algorithm (which is polynomial time in the degree). This locates all the repeated factors of f, and lets one quickly reduce to the case when f is squarefree. (This trick is unavailable in the integer case, due to the lack of a good notion of derivative.)

2. Observe (from Cauchy's theorem) that for any prime polynomial g of degree d, we have $t^{|F|^d} = t \bmod g$. Thus the polynomial $t^{|F|^d} - t$ contains the product of all the primes of this degree (and of all primes of degree dividing d); indeed, by the exact prime number theorem and a degree count, these are the only possible factors of $t^{|F|^d} - t$. It is easy to compute the remainder of $t^{|F|^d} - t$ modulo f in polynomial time, and then one can compute gcd of f with $t^{|F|^d} - t$ in polynomial time also. This essentially isolates the prime factors of a fixed degree, and quickly lets one reduce to the case when f is the product of distinct primes of the same degree d. (Here we have exploited the fact that there are many primes with exactly the same norm - which is of course highly false in the integers. Similarly in Step 3 below.)

3. Now we apply the *Cantor-Zassenhaus algorithm*. Let us assume that $|F|$ is odd (the case $|F| = 2$ can be treated by a modification of this method). By computing $g^{(|F|^d-1)/2} \bmod f$ for randomly selected g, we can generate some random square roots a of 1 modulo f (thanks to Cauchy's theorem and the Chinese remainder theorem; there is also a small chance we generate a non-invertible element, but

this is easily dealt with). These square roots a will be either $+1$ or -1 modulo each of the prime factors of f. If we take the gcd of f with $a+1$ or $a-1$ we have a high probability of splitting up the prime factors of f; doing this a few times one soon isolates all the prime factors separately.

2.6.7 Conclusion

As the above whirlwind tour hopefully demonstrates, dyadic models for the integers, reals, and other "linear" objects show up in many different areas of mathematics. In some areas they are an oversimplified and overly easy toy model; in other areas they get at the heart of the matter by providing a model in which all irrelevant technicalities are stripped away; and in yet other areas they are a crucial component in the analysis of the non-dyadic case. In all of these cases, though, it seems that the contribution that dyadic models provide in helping us understand the non-dyadic world is immense.

2.6.8 Notes

This article was originally posted on July 27, 2007 at

`terrytao.wordpress.com/2007/07/27`

John Armstrong noted some analogies between the cocycles which make the non-dyadic world more complicated than the dyadic world, and the commutators which make the non-commutative world more complicated than the commutative world, though the former seem to be more "nilpotent" than the latter.

2.7 "Math doesn't suck", and the Chayes-McKellar-Winn theorem

As you may already know, Danica McKellar, the actress and UCLA mathematics alumnus, has recently launched her book "Math Doesn't Suck"[McK2007], which is aimed at pre-teenage girls and is a friendly introduction to middle-school mathematics, such as the arithmetic of fractions. The book has received quite a bit of publicity, most of it rather favourable, and is selling quite well; at one point, it even made the Amazon top 20 bestseller list, which is a remarkable achievement for a mathematics book. (The current Amazon rank can be viewed in the product details of the Amazon page for this book.)

I'm very happy that the book is successful for a number of reasons. Firstly, I got to know Danica for a few months (she took my Introduction to Topology class way back in 1997), and it is always very heartening to see a former student put her or his mathematical knowledge to good use. Secondly, Danica is a wonderful role model and it seems that this book will encourage many school-age kids to give maths a chance. But the final reason is that the book is, in fact, rather good; the mathematical content is organised in a logical manner (for instance, it begins with prime factorisation, then covers least common multiples, then addition of fractions), well motivated, and interleaved with some entertaining, insightful, and slightly goofy digressions, anecdotes, and analogies. (To give one example: to motivate why dividing 6 by $1/2$ should yield 12, she first discussed why 6 divided by 2 should give 3, by telling a story about having to serve lattes to a whole bunch of actors, where each actor demands two lattes each, but one could only carry the weight of six lattes at a time, so that only $6/2 = 3$ actors could be served in one go; she then asked what would happen instead of each actor only wanted half a latte instead of two. Danica also gives a very clear explanation of the concept of a variable (such as x), by using the familiar concept of a nickname given to someone with a complicated real name as an analogy.)

While I am not exactly in the target audience for this book, I can relate to its pedagogical approach. When I was a kid myself, one of my favourite maths books was a very obscure (and now completely out of print) book called "Creating Calculus"[HoPy1974], which introduced the basics of single-variable calculus via concocting a number of slightly silly and rather contrived stories which always involved one or more ants. For instance, to illustrate the concept of a derivative, in one of these stories one of the ants kept walking up a mathematician's shin while he was relaxing against a tree, but started slipping down at a point where the slope of the shin reached a certain threshold; this got the mathematician interested enough to compute that slope from first principles. The humour in the book was rather corny, involving for instance some truly awful puns, but it was perfect for me when I was 11: it inspired me to play with calculus, which is an important step towards improving one's understanding of the subject beyond a superficial level. (Two other books in a similarly playful spirit, yet still full of genuine scientific substance, are "Darwin for beginners"[MiVa2003] and "Mr. Tompkins in paperback"[Ga1993], both of which I also enjoyed very much as a kid. They are of course no substitute for a serious textbook on these subjects, but they complement such treatments excellently.)

Anyway, Danica's book has already been reviewed in several places, and there's not much more I can add to what has been said elsewhere. I thought however that I could talk about another of Danica's contributions to mathematics, namely her paper "Percolation and Gibbs states multiplicity for ferromagnetic Ashkin-Teller models on $\mathbf{Z}^2$"[ChMcKWi1998], joint with Brandy Winn and my colleague Lincoln Chayes. This paper is noted from time to time in the above-mentioned publicity, and its main result is sometimes referred to there as the "Chayes-McKellar-Winn theorem", but as far as I know, no serious effort has been made to explain exactly what this theorem is, or the wider context the result is placed in. So I'll give it a shot; this allows me an opportunity to talk about some beautiful topics in mathematical physics, namely statistical mechanics, spontaneous magnetisation, and percolation.

2.7.1 Statistical mechanics

To begin the story, I would like to quickly review the theory of *statistical mechanics*. This is the theory which bridges the gap between the microscopic (particle physics) description of many-particle systems, and the macroscopic (thermodynamic) description, giving a semi-rigorous explanation of the empirical laws of the latter in terms of the fundamental laws of the former.

Statistical mechanics is a remarkably general theory for describing many-particle systems - for instance it treats classical and quantum systems in almost exactly the same way! But to simplify things I will just discuss a toy model of the microscopic dynamics of a many-particle system S - namely a finite Markov chain model. In this model, time is discrete, though the interval between discrete times should be thought of as extremely short. The *state space* is also discrete; at any given time, the number of possible *microstates* that the system S could be in is finite (though extremely large - typically, it is exponentially large in the number N of particles). One should view the state space of S as a directed graph with many vertices but relatively low degree. After each discrete time interval, the system may move from one microstate to an adjacent one on the graph, where the transition probability from one microstate to the next is independent of time, or on the past history of the system. We make the key assumption that the counting measure on microstates is invariant, or equivalently that the sum of all the transition probabilities that lead one away from a given microstate equals the sum of all the transition probabilities that lead one towards that microstate. (In classical systems, governed by Hamiltonian mechanics, the analogue of this assumption is Liouville's theorem; in quantum systems, governed by Schrdinger's equation, the analogue is the unitarity of the evolution operator.) We also make the mild assumption that the transition probability across any edge is positive.

If the graph of microstates was connected (i.e. one can get from any microstate to any other by some path along the graph), then after a sufficiently long period of time, the probability distribution of the microstates will converge towards normalised counting measure, as can be seen by basic Markov chain theory. However, if the system S is *isolated* (i.e. not interacting with the outside world), conservation laws intervene to disconnect the graph. In particular, if each microstate x had a total energy $H(x)$, and one had a law of conservation of energy which meant that microstates could only transition to other microstates with the same energy, then the probability distribution

could be trapped on a single energy surface, defined as the collection $\{x : H(x) = E\}$ of all microstates of S with a fixed total energy E.

Physics has many conservation laws, of course, but to simplify things let us suppose that energy is the only conserved quantity of any significance (roughly speaking, this means that no other conservation law has a significant impact on the entropy of possible microstates). In fact, let us make the stronger assumption that the energy surface is connected; informally, this means that there are no "secret" conservation laws beyond the energy which could prevent the system evolving from one side of the energy surface to the other.

In that case, Markov chain theory lets one conclude that if the solution started out at a fixed total energy E, and the system S was isolated, then the limiting distribution of microstates would just be the uniform distribution on the energy surface $\{x : H(x) = E\}$; every state on this surface is equally likely to occur at any given instant of time (this is known as the *fundamental postulate of statistical mechanics*, though in this simple Markov chain model we can actually derive this postulate rigorously). This distribution is known as the microcanonical ensemble of S at energy E. It is remarkable that this ensemble is largely independent of the actual values of the transition probabilities; it is only the energy E and the function H which are relevant. (This analysis is perfectly rigorous in the Markov chain model, but in more realistic models such as Hamiltonian mechanics or quantum mechanics, it is much more difficult to rigorously justify convergence to the microcanonical ensemble. The trouble is that while these models appear to have a chaotic dynamics, which should thus exhibit very pseudorandom behaviour (similar to the genuinely random behaviour of a Markov chain model), it is very difficult to demonstrate this pseudorandomness rigorously; the same difficulty, incidentally, is present in the Navier-Stokes regularity problem, see Section 1.4.)

In practice, of course, a small system S is almost never truly isolated from the outside world S', which is a far larger system; in particular, there will be additional transitions in the combined system $S \cup S'$, through which S can exchange energy with S'. In this case we do not expect the S-energy $H(x)$ of the combined microstate (x,x') to remain constant; only the global energy $H(x) + H'(x')$ will equal a fixed number E. However, we can still view the larger system $S \cup S'$ as a massive isolated system, which will have some microcanonical ensemble; we can then project this ensemble onto S to obtain the canonical ensemble for that system, which describes the distribution of S when it is in thermal equilibrium with S'. (Of course, since we have not yet experienced heat death, the entire outside world is not yet in the microcanonical ensemble; but in practice, we can immerse a small system in a heat bath, such as the atmosphere, which accomplishes a similar effect.)

Now it would seem that in order to compute what this canonical ensemble is, one would have to know a lot about the external system S', or the total energy E. Rather astonishingly, though, as long as S' is much larger than S, and obeys some plausible physical assumptions, we can specify the canonical ensemble of S using only a single scalar parameter, the *temperature* T. To see this, recall in the microcanonical ensemble of $S \cup S'$, each microstate (x,x') with combined energy $H(x) + H'(x') = E$ has an equal probability of occurring at any given time. Thus, given any microstate x of S, the probability that x occurs at a given time will be proportional to the cardinality of the

set $\{x' : H(x') = E - H(x)\}$. Now as the outside system S' is very large, this set will be be enormous, and presumably very complicated as well; however, the key point is that it only depends on E and x through the quantity $E - H(x)$. Indeed, we conclude that the canonical ensemble distribution of microstates at x is proportional to $\Omega(E - H(x))$, where $\Omega(E')$ is the number of microstates of the outside system S' with energy E'.

Now it seems that it is hopeless to compute $\Omega(E')$ without knowing exactly how the system S' works. But, in general, the number of microstates in a system tends to grow exponentially in the energy in some fairly smooth manner, thus we have $\Omega(E') = \exp(F(E'))$ for some smooth increasing function F of E' (although in some rare cases involving population inversion, F may be decreasing). Now, we are assuming S' is much larger than S, so E should be very large compared with $H(x)$. In such a regime, we expect Taylor expansion to be reasonably accurate, thus $\Omega(E - H(x)) \approx \exp(F(E) - \beta H(x))$, where $\beta := F'(E)$ is the derivative of F at E (or equivalently, the log-derivative of Ω); note that β is positive by assumption. The quantity $\exp(F(E))$ doesn't depend on x, and so we conclude that the canonical ensemble is proportional to counting measure, multiplied by the function $\exp(-\beta H(x))$. Since probability distributions have total mass 1, we can in fact describe the probability $P(x)$ of the canonical ensemble being at x exactly as

$$P(x) = \frac{1}{Z} e^{-\beta H(x)}$$

where Z is the partition function

$$Z := \sum_x e^{-\beta H(x)}.$$

The canonical ensemble is thus specified completely except for a single parameter $\beta > 0$, which depends on the external system S' and on the total energy E. But if we take for granted the laws of thermodynamics (particularly the zeroth law), and compare S' with an ideal gas, we can obtain the relationship $\beta = 1/kT$, where T is the temperature of S' and k is Boltzmann's constant. Thus the canonical ensemble of a system S is completely determined by the temperature, and on the energy functional H. The underlying transition graph and transition probabilities, while necessary to ensure that one eventually attains this ensemble, do not actually need to be known in order to compute what this ensemble is, and can now (amazingly enough) be discarded. (More generally, the microscopic laws of physics, whether they be classical or quantum, can similarly be discarded almost completely at this point in the theory of statistical mechanics; the only thing one needs those laws of physics to provide is a description of all the microstates and their energy, though in some situations one also needs to be able to compute other conserved quantities, such as particle number.)

At the temperature extreme $T \to 0$, the canonical ensemble becomes concentrated at the minimum possible energy $E_{\min}$ for the system (this fact, incidentally, inspires the numerical strategy of simulated annealing); whereas at the other temperature extreme $T \to \infty$, all microstates become equally likely, regardless of energy.

2.7.2 Gibbs states

One can of course continue developing the theory of statistical mechanics and relate temperature and energy to other macroscopic variables such as volume, particle number, and entropy (see for instance Schrödinger's classic little book [Sc1989]), but I'll now turn to the topic of Gibbs states of infinite systems, which is one of the main concerns of [ChMcKWi1998].

A *Gibbs state* is simply a distribution of microstates of a system which is invariant under the dynamics of that system, physically, such states are supposed to represent an equilibrium state of the system. For systems S with finitely many degrees of freedom, the microcanonical and canonical systems given above are examples of Gibbs states. But now let us consider systems with infinitely many degrees of freedom, such as those arising from an infinite number of particles (e.g. particles in a lattice). One cannot now argue as before that the entire system is going to be in a canonical ensemble; indeed, as the total energy of the system is likely to be infinite, it is not even clear that such an ensemble still exists. However, one can still argue that any localised portion S_1 of the system (with finitely many degrees of freedom) should still be in a canonical ensemble, by treating the remaining portion $S \backslash S_1$ of the system as a heat bath that S_1 is immersed in. Furthermore, the zeroth law of thermodynamics suggests that all such localised subsystems should be at the same temperature T. This leads to the definition of a Gibbs state at temperature T for a global system S: it is any probability distribution of microstates whose projection to any local subsystem S_1 is in the microcanonical ensemble at temperature T. (To make this precise, one needs probability theory on infinite dimensional spaces, but this can be put on a completely rigorous footing, using the theory of product measures. There are also some technical issues regarding compatibility on the boundary between S_1 and $S \backslash S_1$ which I will ignore here.)

For systems with finitely many degrees of freedom, there is only one canonical ensemble at temperature T, and thus only one Gibbs state at that temperature. However, for systems with infinitely many degrees of freedom, it is possible to have more than one Gibbs state at a given temperature. This phenomenon manifests itself physically via phase transitions, the most familiar of which involves transitions between solid, liquid, and gaseous forms of matter, but also includes things like spontaneous magnetisation or demagnetisation. Closely related to this is the phenomenon of spontaneous symmetry breaking, in which the underlying system (and in particular, the energy functional H) enjoys some symmetry (e.g. translation symmetry or rotation symmetry), but the Gibbs states for that system do not. For instance, the laws of magnetism in a bar of iron are rotation symmetric, but there are some obviously non-rotation-symmetric Gibbs states, such as the magnetised state in which all the iron atoms have magnetic dipoles oriented with the north pole pointing in (say) the upward direction. [Of course, as the universe is finite, these systems do not truly have infinitely many degrees of freedom, but they do behave analogously to such systems in many ways.]

It is thus of interest to determine, for any given physical system, under what choices of parameters (such as the temperature T) one has non-uniqueness of Gibbs states. For "real-life" physical systems, this question is rather difficult to answer, so mathematicians have focused attention instead on some simpler toy models. One of the most popular of these is the *Ising model*, which is a simplified model for studying phenom-

ena such as spontaneous magnetism. A slight generalisation of the Ising model is the Potts model; the *Ashkin-Teller model*, which is studied in [ChMcKWi1998], is an interpolant betwen a certain Ising model and a certain Potts model.

2.7.3 Ising, Potts, and Ashkin-Teller models

All three of these models involve particles on the infinite two-dimensional lattice $\mathbf{Z}^2$, with one particle at each lattice point (or site). (One can also consider these models in other dimensions, of course; the behaviour is quite different in different dimensions.) Each particle can be in one of a finite number of states, which one can think of as "magnetisations" of that particle. In the classical Ising model there are only two states ($+1$ and -1), though in [ChMcKWi1998], four-state models are considered. The particles do not move from their designated site, but can change their state over time, depending on the state of particles at nearby sites.

As discussed earlier, in order to do statistical mechanics, we do not actually need to specify the exact mechanism by which the particles interact with each other; we only need to describe the total energy of the system. In these models, the energy is contained in the bonds between adjacent sites on the lattice (i.e. sites which are a unit distance apart). The energy of the whole system is then the sum of the energies of all the bonds, and the energy of each bond depends only on the state of the particles at the two endpoints of the bond. (The total energy of the infinite system is then a divergent sum, but this is not a concern since one only needs to be able to compute the energy of finite subsystems, in which one only considers those particles within, say, a square of length R.) The Ising, Potts, and Ashkin-Teller models then differ only in the number of states and the energy of various bond configurations. Up to some harmless normalisations, we can describe them as follows:

1. In the *classical Ising model*, there are two magnetisation states ($+1$ and -1); the energy of a bond between two particles is $-1/2$ if they are in the same state, and $+1/2$ if they are in the opposite state (thus one expects the states to align at low temperatures and become non-aligned at high temperatures);

2. In the *four-state Ising model*, there are four magnetisation states $(+1,+1)$, $(+1,-1)$, $(-1,+1)$, and $(-1,-1)$ (which can be viewed as four equally spaced vectors in the plane), and the energy of a bond between two particles is the sum of the classical Ising bond energy between the first component of the two particle states, and the classical Ising bond energy between the second component. Thus for instance the bond energy between particles in the same state is -1, particles in opposing states is $+1$, and particles in orthogonal states (e.g. $(+1,+1)$ and $(+1,-1)$) is 0. This system is equivalent to two non-interacting classical Ising models, and so the four-state theory can be easily deduced from the two-state theory.

3. In the *degenerate Ising model*, we have the same four magnetisation states, but now the bond energy between particles is $+1$ if they are in the same state or opposing state, and 0 if they are in an orthogonal state. This model essentially

collapses to the two-state model after identifying $(+1,+1)$ and $(-1,-1)$ as a single state, and identifying $(+1,-1)$ and $(-1,+1)$ as a single state.

4. In the *four-state Potts model*, we have the same four magnetisation states, but now the energy of a bond between two particles is -1 if they are in the same state and 0 otherwise.

5. In the *Ashkin-Teller model*, we have the same four magnetisation states; the energy of a bond between two particles is -1 if they are in the same state, 0 if they are orthogonal, and ε if they are in opposing states. The case $\varepsilon = +1$ is the four-state Ising model, the case $\varepsilon = 0$ is the Potts model, and the cases $0 < \varepsilon < 1$ are intermediate between the two, while the case $\varepsilon = -1$ is the degenerate Ising model.

For the classical Ising model, there are two minimal-energy states: the state where all particles are magnetised at $+1$, and the state where all particles are magnetised at -1. (One can of course also take a probabilistic combination of these two states, but we may as well restrict attention to pure states here.) Since one expects the system to have near-minimal energy at low temperatures, we thus expect to have non-uniqueness of Gibbs states at low temperatures for the Ising model. Conversely, at sufficiently high temperatures the differences in bond energy should become increasingly irrelevant, and so one expects to have uniqueness of Gibbs states at high energy. (Nevertheless, there is an important duality relationship between the Ising model at low and high energies.)

Similar heuristic arguments apply for the other models discussed above, though for the degenerate Ising model there are many more minimal-energy states and so even at very low temperatures one only expects to obtain partial ordering rather than total ordering in the magnetisations.

For the Askhin-Teller models with $0 < \varepsilon < 1$, it was known for some time that there is a unique critical temperature T_c (which has a physical interpretation as the *Curie temperature*), below which one has non-unique and magnetised Gibbs states (thus the expected magnetisation of any given particle is non-zero), and above which one has unique (non-magnetised) Gibbs states. (For ε close to -1 there are two critical temperatures, describing the transition from totally ordered magnetisation to partially ordered, and from partially ordered to unordered.) The problem of computing this temperature T_c exactly, and to describe the nature of this transition, appears to be rather difficult, although there are a large number of partial results. What was shown in [ChMcKWi1998], though, is that this critical temperature T_c is also the critical temperature T_p for a somewhat simpler phenomenon, namely that of *site percolation*. Let us denote one of the magnetised states, say $(+1,+1)$, as "blue". We then consider the Gibbs state for a bounded region (e.g. an $N \times N$ square), subject to the boundary condition that the entire boundary is blue. In the zero temperature limit $T \to 0$ the entire square would then be blue; in the high temperature limit $T \to +\infty$ each particle would have an independent random state. Consider the probability p_N that a particle at the center of this square is part of the blue "boundary cluster"; in other words, the particle is not only blue, but there is a path of bond edges connecting this particle to the boundary which only goes through blue vertices. Thus we expect this probability to be

close to 1 at very low temperatures, and close to 0 at very high temperatures. And indeed, standard percolation theory arguments show that there is a critical temperature T_p below which $\lim_{N\to\infty} p_N$ is positive (or equivalently, the boundary cluster has density bounded away from zero), and below which $\lim_{N\to\infty} p_N = 0$ (thus the boundary cluster has asymptotic density zero). The "Chayes-McKellar-Winn theorem" is then the claim that $T_c = T_p$.

This result is part of a very successful program, initiated by Fortuin and Kasteleyn [FoKa1972], to analyse the statistical mechanics of site models such as the Ising, Potts, and Ashkin-Teller models via the random clusters generated by the bonds between these sites. (One of the fruits of this program, by the way, was the FKG correlation inequality, which asserts that any two monotone properties on a lattice are positively correlated. This inequality has since proven to be incredibly useful in probability, combinatorics and computer science.) The claims $T_c \leq T_p$ and $T_p \leq T_c$ are proven separately. To prove $T_c \leq T_p$ (i.e. multiple Gibbs states implies percolation), the main tool is a theorem of Chayes and Machta[ChMa1995] that relates the non-uniqueness of Gibbs states to positive magnetisation (the existence of states in which the expected magnetisation of a particle is non-zero). To prove $T_p \leq T_c$ (i.e. percolation implies multiple Gibbs states), the main tool is a theorem of Gandolfi, Keane, and Russo[GaKeRu1998], which studied percolation on the infinite lattice and who showed that under certain conditions (in particular, that a version of the FKG inequality is satisfied), there can be at most one infinite cluster; basically, one can use the colour of this cluster (which will exist if percolation occurs) to distinguish between different Gibbs states. (The fractal structure of this infinite cluster, especially near the critical temperature, is quite interesting, but that's a whole other story.) One of the main tasks in [ChMcKWi1998] paper is to verify the FKG inequality for the Ashkin-Teller model; this is done by viewing that model as a perturbation of the Ising model, and expanding the former using the random clusters of the latter.

2.7.4 Executive summary

When one heats an iron bar magnet above a certain special temperature - the Curie temperature - the iron bar will cease to be magnetised; when one cools the bar again below this temperature, the bar can once again spontaneously magnetise in the presence of an external magnetic field. This phenomenon is still not perfectly understood; for instance, it is difficult to predict the Curie temperature precisely from the fundamental laws of physics, although one can at least prove that this temperature exists. However, Chayes, McKellar, and Winn have shown that for a certain simplified model for magnetism (known as the Ashkin-Teller model), the Curie temperature is equal to the critical temperature below which percolation can occur; this means that even when the bar is unmagnetised, enough of the iron atoms in the bar spin in the same direction that they can create a connected path from one end of the bar to another. Percolation in the Ashkin-Teller model is not fully understood either, but it is a simpler phenomenon to deal with than spontaneous magnetisation, and so this result represents an advance in our understanding of how the latter phenomenon works.

2.7.5 Notes

This article was originally posted on Aug 20, 2007 at

terrytao.wordpress.com/2007/08/20

See also an explanation by John Baez at

lem.ph.utexas.edu/category/2007/08/gerbes_in_the_guardian.html{\#}c011

2.8 Nonfirstorderisability

I recently came across the phenomenon of *nonfirstorderisability* in mathematical logic: there are perfectly meaningful and useful statements in mathematics which cannot be phrased within the confines of first order logic (combined with the language of set theory, or any other standard mathematical theory). In order to phrase such statements rigorously, one must use a more powerful language such as second order logic instead. This phenomenon is very well known among logicians, but I hadn't learned about it until very recently, and had naively assumed that first order logic sufficed for "everyday" usage of mathematics.

Let's begin with some simple examples of statements which *can* be expressed in first-order logic. If $B(x,y)$ is a binary relation on two objects x, y, then we can express the statement

For every x, there exists a y depending on x such that $B(x,y)$ is true

in first order logic as

$$\forall x \exists y : B(x,y)$$

and

For every x, there exists a y *independent* of x such that $B(x,y)$ is true

can be expressed as

$$\exists y \forall x : B(x,y).$$

Moving on to a more complicated example, if $Q(x,x',y,y')$ is a quaternary relation on four objects x, x', y, y', then we can express the statement

For every x and x', there exists a y depending only on

x and a y' depending on x,x' such that $Q(x,x',y,y')$ is true

as

$$\forall x \exists y \forall x' \exists y' : Q(x,x',y,y')$$

(note that this allows y' to depend on y also, but this turns out to be moot, because y depends only on x), and one can similarly express

For every x and x', there exists a y depending on

x and x' and a y' depending only on x' such that $Q(x,x',y,y')$ is true

as

$$\forall x' \exists y' \forall x \exists y : Q(x,x',y,y')$$

but it seems that one cannot express

For every x and x', there exists a y depending only on x and a y'
depending only on x' such that $Q(x,x',y,y')$ is true (2.7)

in first order logic! For instance, the statement

Theorem 2.36. *To every finitely generated real vector space V one can associate a unique non-negative integer* $\dim(V)$ *such that*

1. *V, W are isomorphic if and only if* $\dim(V) = \dim(W)$*;*
2. *an injection from V to W exists if and only if* $\dim(V) \leq \dim(W)$*;*
3. *a surjection from V to W exists if and only if* $\dim(V) \geq \dim(W)$*;*
4. $\dim(\mathbb{R}) = 1$*; and*
5. $\dim(V \oplus W) = \dim(V) + \dim(W)$ *for all V, W,*

which is part of the fundamental theorem of linear algebra, does not seem to be expressible as stated in first order set theory (though of course the concept of dimension can be explicitly constructed within this language), even if we drop the uniqueness and restrict ourselves to just the assertion that $\dim()$ obey, say, property 1, so that we get an assertion of the form (2.7). Note that the category of all finite-dimensional vector spaces is not a set (for reasons relating to Russell's paradox) and so we cannot view dim as a function. More generally, many statements in category theory dealing with large categories seem to not be expressible in first order logic.

I can't quite show that (2.7) is not expressible in first-order logic, but I can come very close, using non-standard analysis (see Section 2.5). The statement

Theorem 2.37. *For every real numbers x and x' there exists real numbers* $\mathrm{st}(x)$ *and* $\mathrm{st}(x')$ *depending only on x and x' respectively, such that* $\mathrm{st}(x+x') = \mathrm{st}(x) + \mathrm{st}(x')$, $\mathrm{st}(xx') = \mathrm{st}(x)\mathrm{st}(x')$, $\mathrm{st}(1) = 1$*, and* $\mathrm{st}(x)$ *is non-negative whenever x is non-negative, and also such that* $\mathrm{st}(x)$ *is not always equal to x.*

is true in the non-standard model of the real numbers, but false in the standard model (this is the classic algebra homework problem that the only order-preserving field homomorphism on the reals is the identity). Since the transfer principle ensures that all first-order statements that are true in the standard reals are also true in the non-standard reals, this means that the above statement cannot be expressed in first-order logic. If it weren't for the "$\mathrm{st}(x)$ is not always equal to x" part, this would basically be of the form (2.7).

It seems to me that first order logic is limited by the linear (and thus totally ordered) nature of its sentences; every new variable that is introduced must be allowed to depend on *all* the previous variables introduced to the left of that variable. This does not fully capture all of the dependency trees of variables which one deals with in mathematics. In analysis, we tend to get around this by using English phrasings such as

> ... assuming N is chosen sufficiently large depending on ε, and δ chosen sufficiently small depending on N...

and

> ... where C can depend on k and d, but is uniform with respect to n and f...,

or by using the tremendously convenient $O()$ and $o()$ notation of Landau. One then takes for granted that one can eventually unwind all these phrasings to get back to a sentence in formal, first-order logic. As far as analysis is concerned, this is a fairly safe assumption, since one usually deals with objects in very concrete sets such as the real numbers, and one can easily model all of these dependencies using functions from concrete sets to other concrete sets if necessary. (Also, the hierarchy of magnitudes in analysis does often tend to be rather linearly ordered.) But some subtleties may appear when one deals with large categories, such as the category of sets, groups, or vector spaces (though in most applications, one can cap the cardinality of these objects and then one can represent these categories up to equivalence by an actual set). It may be that a more diagrammatic language (perhaps analogous to the commutative diagrams in category theory, or one based on trees or partially ordered sets rather than linearly ordered ones) may be a closer fit to expressing the way one actually thinks about how variables interact with each other. Second-order logic¡/a¿ is, of course, an obvious candidate for such a language, but it may be overqualified for the task.

2.8.1 Notes

This article was originally posted on Aug 28, 2007 at

terrytao.wordpress.com/2007/08/27

Ori Gurel-Gurevich pointed out that if one used a first-order set theory such as NBG, which incorporates classes as well as sets, then statements such as Theorem 2.36 can be stated in first-order logic.

Andy D. gave the example of the quaternary relation $Q(x,x',y,y')$ defined as

$$(y \neq x) \wedge ((x = x') \implies (y = y')) \wedge ((x \neq x') \implies (y \neq y'))$$

for which (2.7) holds if and only if there is a perfect matching on the elements of the universe, or in other words if the universe is either infinite or finite of even order. But the parity of a finite universe is known to not be definable in first-order logic, thus establishing the claim in the article.

Suresh Venkat commented on connections between first and second-order logic and complexity theory, while Emmanuel Kowalski commented on connections between first-order definability and the structure of sets in arithmetic geometry. David Corfield pointed out the work of Hintikka on branching quantifiers, which can capture statements such as (2.7), and the work of Abramsky connecting these quantifiers to game theory.

Thanks to tom for corrections.

2.9 Amplification, arbitrage, and the tensor power trick

Today I would like discuss the *amplification trick* in harmonic analysis and combinatorics (and in particular, in the study of estimates); this trick takes an established estimate involving an arbitrary object (such as a function f), and obtains a stronger (or amplified) estimate by transforming the object in a well-chosen manner (often involving some new parameters) into a new object, applying the estimate to that new object, and seeing what that estimate says about the original object (after optimising the parameters or taking a limit). The amplification trick works particularly well for estimates which enjoy some sort of symmetry on one side of the estimate that is not represented on the other side; indeed, it can be viewed as a way to "arbitrage" differing amounts of symmetry between the left- and right-hand sides of an estimate. It can also be used in the contrapositive, amplifying a weak counterexample to an estimate into a strong counterexample. This trick also sheds some light as to why dimensional analysis works; an estimate which is not dimensionally consistent can often be amplified into a stronger estimate which is dimensionally consistent; in many cases, this new estimate is so strong that it cannot in fact be true, and thus dimensionally inconsistent inequalities tend to be either false or inefficient, which is why we rarely see them. (More generally, any inequality on which a group acts on either the left or right-hand side can often be "decomposed" into the "isotypic components" of the group action, either by the amplification trick or by other related tools, such as Fourier analysis.)

The amplification trick is a deceptively simple one, but it can become particularly powerful when one is arbitraging an unintuitive symmetry, such as symmetry under tensor powers. Indeed, the "tensor power trick", which can eliminate constants and even logarithms in an almost magical manner, can lead to some interesting proofs of sharp inequalities, which are difficult to establish by more direct means, as we shall see below.

The most familiar example of the amplification trick in action is probably the textbook proof of the Cauchy-Schwarz inequality

$$|\langle v, w\rangle| \leq \|v\|\|w\| \tag{2.8}$$

for vectors v, w in a complex Hilbert space. To prove this inequality, one might start by exploiting the obvious inequality

$$\|v - w\|^2 \geq 0 \tag{2.9}$$

but after expanding everything out, one only gets the weaker inequality

$$\mathrm{Re}\langle v, w\rangle \leq \frac{1}{2}\|v\|^2 + \frac{1}{2}\|w\|^2. \tag{2.10}$$

Now (2.10) is weaker than (2.8) for two reasons; the left-hand side is smaller, and the right-hand side is larger (thanks to the arithmetic mean-geometric mean inequality). However, we can amplify (2.10) by arbitraging some symmetry imbalances. Firstly, observe that the *phase rotation symmetry* $v \mapsto e^{i\theta}v$ preserves the RHS of (2.10) but not the LHS. We exploit this by replacing v by $e^{i\theta}v$ in (2.10) for some phase θ to be chosen later, to obtain

$$\mathrm{Re}(e^{i\theta}\langle v,w\rangle) \leq \frac{1}{2}\|v\|^2 + \frac{1}{2}\|w\|^2.$$

Now we are free to choose θ at will (as long as it is real, of course), so it is natural to choose θ to optimise the inequality, which in this case means to make the left-hand side as large as possible. This is achieved by choosing $e^{i\theta}$ to cancel the phase of $\langle v,w\rangle$, and we obtain

$$|\langle v,w\rangle| \leq \frac{1}{2}\|v\|^2 + \frac{1}{2}\|w\|^2. \tag{2.11}$$

This is closer to (2.8); we have fixed the left-hand side, but the right-hand side is still too weak. But we can amplify further, by exploiting an imbalance in a different symmetry, namely the *homogenisation symmetry* $(v,w) \mapsto (\lambda v, \frac{1}{\lambda} w)$ for a scalar $\lambda > 0$, which preserves the left-hand side but not the right. Inserting this transform into (2.11) we conclude that

$$|\langle v,w\rangle| \leq \frac{\lambda^2}{2}\|v\|^2 + \frac{1}{2\lambda^2}\|w\|^2$$

where $\lambda > 0$ is at our disposal to choose. We can optimise in λ by minimising the right-hand side, and indeed one easily sees that the minimum (or infimum, if one of v and w vanishes) is $\|v\|\|w\|$ (which is achieved when $\lambda = \sqrt{\|w\|/\|v\|}$ when v,w are non-zero, or in an asymptotic limit $\lambda \to 0$ or $\lambda \to \infty$ in the degenerate cases), and so we have amplified our way to the Cauchy-Schwarz inequality (2.8).

2.9.1 Amplification via phase, homogeneity, or dilation symmetry

Many similar examples of amplification are used routinely to prove the basic inequalities in harmonic analysis. For instance to deduce the complex-valued triangle inequality

$$|\int_X f(x)\,d\mu(x)| \leq \int_X |f(x)|\,d\mu(x) \tag{2.12}$$

(where (X,μ) is a measure space and f is absolutely integrable) from its real-valued counterpart, we first apply the latter inequality to $\mathrm{Re}(f)$ to obtain

$$|\mathrm{Re}\int_X f(x)\,d\mu(x)| \leq \int_X |\mathrm{Re}\, f(x)|\,d\mu(x).$$

To make the right-hand side phase-rotation-invariant, we crudely bound $|\mathrm{Re}\, f(x)|$ by $|f(x)|$, obtaining $|\mathrm{Re}\int_X f(x)\,d\mu(x)| \leq \int_X |f(x)|\,d\mu(x)$ and then one can arbitrage the imbalance in phase rotation symmetry to obtain (2.12). For another well-known example, to prove Hölder's inequality

$$\int_X f(x)g(x)\,d\mu(x) \leq \|f\|_{L^p(X,d\mu)}\|g\|_{L^q(X,d\mu)} \tag{2.13}$$

for non-negative measurable f,g and dual exponents $1 \leq p,q \leq \infty$, one can begin with the elementary (weighted) arithmetic mean-geometric mean inequality

$$ab \leq \frac{1}{p}a^p + \frac{1}{q}b^q \tag{2.14}$$

for non-negative a, b (which follows from the convexity of the function $\theta \mapsto a^\theta b^{1-\theta}$, which in turn follows from the convexity of the exponential function) to obtain the inequality

$$\int_X f(x)g(x)\,d\mu(x) \leq \frac{1}{p}\|f\|^p_{L^p(X,d\mu)} + \frac{1}{q}\|g\|^q_{L^q(X,d\mu)}.$$

This inequality is weaker than (2.13) (because of (2.14)); but if one amplifies by arbitraging the imbalance in the homogenisation symmetry $(f,g) \mapsto (\lambda f, \frac{1}{\lambda} g)$ one obtains (2.13). As a third example, the Sobolev embedding inequality

$$\|f\|_{L^q(\mathbf{R}^d)} \leq C_{p,q,d}(\|f\|_{L^p(\mathbf{R}^d)} + \|\nabla f\|_{L^p(\mathbf{R}^d)}), \tag{2.15}$$

which is valid for $1 < p < q < \infty$ and $\frac{1}{q} > \frac{1}{p} - \frac{1}{d}$ (and also valid in some endpoint cases) and all test functions (say) f on $\mathbf{R}^d$, can be amplified to obtain the *Gagliardo-Nirenberg inequality*

$$\|f\|_{L^q(\mathbf{R}^d)} \leq C_{p,q,d}\|f\|^{1-\theta}_{L^p(\mathbf{R}^d)}\|\nabla f\|^\theta_{L^p(\mathbf{R}^d)} \tag{2.16}$$

where $0 < \theta < 1$ is the number such that $\frac{1}{q} = \frac{1}{p} - \frac{\theta}{d}$, by arbitraging the action of the dilation group $f(x) \mapsto f(\lambda x)$. (In this case, the dilation action does not leave either the LHS or RHS of (2.15) invariant, but it affects the LHS in a well controlled manner, which can be normalised out by dividing by a suitable power of λ.) The same trick, incidentally, reveals why the Sobolev embedding inequality fails when $q < p$ or when $\frac{1}{q} < \frac{1}{p} - \frac{1}{d}$, because in these cases it leads to an absurd version of the Gagliardo-Nirenberg inequality. Observe also that the Gagliardo-Nirenberg inequality (2.16) is dimensionally consistent; the dilation action affects both sides of the inequality in the same way. (The weight of the representation of the dilation action on an expression is the same thing as the exponent of the length unit that one assigns to the dimension of that expression.) More generally, arbitraging a dilation symmetry allows a dimensionally consistent inequality to emerge from a dimensionally inconsistent (or dimensionally inefficient) one.

2.9.2 Amplification using linearity

Another powerful source of amplification is linearity (the principle of superposition). A simple example of this is *depolarisation*. Suppose one has a symmetric bilinear form $B(f,g) : X \times X \to \mathbf{R}$ from a normed vector space X to the real numbers, and one has already proven the polarised inequality

$$|B(f,f)| \leq A\|f\|^2_X$$

for all f in X. One can amplify this by replacing f with $f + cg$ for arbitrary $f, g \in X$ and a real parameter c, obtaining

$$|B(f,f) + 2cB(f,g) + c^2 B(g,g)| \leq A(\|f\|_X + |c|\|g\|_X)^2$$

Optimising this in c (e.g. taking $c := \|f\|_X/\|g\|_X$) and using the triangle inequality, one eventually obtains the amplified (depolarised) inequality

$$|B(f,g)| \leq CA\|f\|_X\|g\|_X$$

for some absolute constant $C > 0$.

For a slightly more sophisticated example, suppose for instance that one has a linear operator $T : L^p(X) \to L^p(Y)$ for some $0 < p < \infty$ and some measure spaces X,Y, and that one has established a scalar estimate of the form

$$\|Tf\|_{L^p(Y)} \leq A\|f\|_{L^p(X)} \tag{2.17}$$

for arbitrary scalar functions f. Then by replacing f by a signed sum $\sum_{n=1}^N \varepsilon_n f_n$, where $f_1, \ldots, f_N$ are arbitrary functions in $L^p(X)$ and $\varepsilon_n \in \{-1, +1\}$ are signs, and using linearity, we obtain

$$\|\sum_{n=1}^N \varepsilon_n T f_n\|_{L^p(Y)} \leq A\|\sum_{n=1}^N \varepsilon_n f_n\|_{L^p(X)}.$$

If we raise this to the p^{th} power, take the ε_n to be random (Bernoulli) signs (in order to avoid unexpectedly large cancellations in the series), and then take expectations of both sides, we obtain

$$\mathbf{E}\|\sum_{n=1}^N \varepsilon_n T f_n\|_{L^p(Y)}^p \leq A^p \mathbf{E}\|\sum_{n=1}^N \varepsilon_n f_n\|_{L^p(X)}^p.$$

If one then uses Khintchine's inequality to compute the expectations, one ends up with the vector valued estimate

$$\|(\sum_{n=1}^N |Tf_n|^2)^{1/2}\|_{L^p(Y)}^p \leq C_p A\|(\sum_{n=1}^N |f_n|^2)^{1/2}\|_{L^p(X)}$$

for some constant C_p depending only on p (in particular, it is independent of N). We can then use the monotone convergence theorem to amplify the finite sum to an infinite sum, thus

$$\|(\sum_{n=1}^\infty |Tf_n|^2)^{1/2}\|_{L^p(Y)}^p \leq C_p A\|(\sum_{n=1}^\infty |f_n|^2)^{1/2}\|_{L^p(X)}.$$

Comparing this to (2.17) we see that we have amplified a scalar inequality (in which the unknown function f takes values in the real or complex numbers) to a vector-valued inequality[20] (in which we have a sequence $f = (f_n)_{n=1}^\infty$ taking values in the Hilbert space $l^2(\mathbf{N})$).

If the estimate one is studying involves "localised" operators and "localisable" norms, then one can use linearity to amplify a global estimate into a more localised one. For instance, let us return to the Sobolev inequality (2.15). We can establish a partition of unity $1 = \sum_{n \in \mathbf{Z}^d} \psi(x-n)$ for some bump function ψ, then we see that

$$\|f\|_{L^q(\mathbf{R}^d)} \leq C_{d,q,\psi}(\sum_{n \in \mathbf{Z}^d} \|\psi(\cdot - n)f\|_{L^q(\mathbf{R}^d)}^q)^{1/q}.$$

Applying the Sobolev inequality (2.15) to each localised function $\psi(\cdot - n)f$ and then summing up, one obtains the localised Sobolev inequality $\|f\|_{L^q(\mathbf{R}^d)} \leq C'_{p,q,d}(\sum_{n \in \mathbf{Z}^d}(\|f\|_{L^p(Q_n)} + \|\nabla f\|_{L^p(Q_n)})^q)^{1/q}$, where Q_n is the cube of sidelength 1 centred at n. This estimate is a little stronger than (2.15), because the l^q summation norm is smaller than the l^p summation norm.

[20] This particular amplification was first observed by Marcinkiewicz and Zygmund[MaZy1939].

2.9.3 Amplification via translation invariance

If T is a translation-invariant operator on $\mathbf{R}^n$ which is not identically zero, one can automatically rule out a large variety of estimates concerning T due to their incompatibility with translation invariance (they would amplify themselves into an absurd estimate). For instance, it will not be possible to establish[21] any weighted estimate involving power weights such as $(1+|x|)^\alpha$ in which there is a higher exponent on the left. More precisely if $\alpha > \beta$ are real numbers and $0 < p,q < \infty$, then it is not possible for any estimate of the form

$$\|(1+|x|)^\alpha Tf\|_{L^q(\mathbf{R}^n)} \leq C_{p,q,\alpha,\beta,n}\|(1+|x|)^\beta f\|_{L^p(\mathbf{R}^n)}$$

to be true. Indeed, if such an estimate was true, then by using the translation invariance we can amplify the above estimate to

$$\|(1+|x-x_0|)^\alpha Tf\|_{L^q(\mathbf{R}^n)} \leq C_{p,q,\alpha,\beta,n}\|(1+|x-x_0|)^\beta f\|_{L^p(\mathbf{R}^n)}$$

for any $x_0 \in \mathbf{R}^n$. But if one fixes f and lets x_0 go to infinity, we see that the right-hand side grows like $|x_0|^\beta$ while the left-hand side grows like $|x_0|^\alpha$ (unless Tf vanishes entirely), leading to a contradiction.

One can obtain particularly powerful amplifications by combining translation-invariance with linearity, because one can now consider not just translates $f(x-x_0)$ of a single function f, but also consider superpositions $\sum_{n=1}^N c_n f(x-x_n)$ of such functions. For instance, we have the principle (which I believe was first articulated by Littlewood) that a non-trivial translation-invariant linear operator T can only map $L^p(\mathbf{R}^d)$ to $L^q(\mathbf{R}^d)$ when $q \geq p$. (Littlewood summarised this principle as "the higher exponents are always on the left".) To see this, suppose that we had an estimate of the form

$$\|Tf\|_{L^q(\mathbf{R}^d)} \leq A\|f\|_{L^p(\mathbf{R}^d)}. \tag{2.18}$$

We can amplify this estimate by replacing $f(x)$ by $\sum_{n=1}^N f(x-x_n)$, where N is some integer and $x_1,\ldots,x_N$ are widely separated points. If these points are sufficiently far apart, then the RHS of (2.18) is comparable to $AN^{1/p}\|f\|_{L^p(\mathbf{R}^d)}$, whereas the LHS is comparable to $N^{1/q}\|Tf\|_{L^q(\mathbf{R}^d)}$ (note how this uses both the translation-invariance and linearity of T). Thus in the limit we obtain

$$N^{1/q}\|Tf\|_{L^q(\mathbf{R}^d)} \leq AN^{1/p}\|f\|_{L^p(\mathbf{R}^d)}.$$

Letting N go to infinity, we obtain a contradiction unless $q \geq p$ (or unless T is identically zero).

The combination of translation invariance and linearity is so strong that it can amplify even a very qualitative estimate into a quantitative one. A good example of this is Stein's maximal principle[St1961]. Suppose we have some maximal operator $Mf := \sup_n |T_n f|$ on some compact group G with normalised Haar measure dm, where

[21] There is also a Fourier dual to this principle, familiar to experts in the analysis of PDE, which asserts that a function space norm with a low number of derivatives (i.e. a low-regularity norm) cannot control a norm with a high number of derivatives. Here, the underlying symmetry that drives this principle is modulation invariance rather than translation invariance.

the T_n are a sequence of translation-invariant operators which are uniformly bounded on some $L^p(G)$ space for some $1 < p \leq 2$. Suppose we are given the very weak information that Mf is finite almost everywhere for every $f \in L^p(G)$. (This is for instance the case if we know that $T_n f$ converge pointwise almost everywhere.) Miraculously, this qualitative hypothesis can be amplified into a much stronger quantitative one, namely that M is of weak type (p,p):

$$m(\{x \in G : Mf(x) \geq \lambda\}) \leq \frac{C}{\lambda^p} \|f\|_{L^p(G)}^p. \tag{2.19}$$

To see this, suppose for contradiction that (2.19) failed for any C; by homogeneity, it would also fail even when restricted to the case $\lambda = 1$. What this means (thanks to the axiom of choice) is that for any $\delta > 0$, there exists f_δ such that

$$\|f_\delta\|_{L^p(G)}^p < \delta m(E_\delta), \tag{2.20}$$

where E_δ is the set where $Mf_\delta > 1$.

At present, E_ε could be a very small subset of G, although we know that it has positive measure. But we can amplify this set to be very large by the following trick: pick an integer N comparable to $1/m(E_\delta)$, select N random shifts $g_1, \ldots, g_N \in G$ and random signs $\varepsilon_1, \ldots, \varepsilon_N \in \{-1, +1\}$ and replace f_δ by the randomised sum $F_\delta := \sum_{n=1}^N \varepsilon_n f_\delta(g_n^{-1} \cdot)$. This sum will tend to be large (greater than or comparable to 1) on most of the union $\bigcup_{n=1}^N g_n \cdot E_\varepsilon$; this can be made precise using Khintchine's inequality. On the other hand, another application of Khintchine's inequality using (2.20) shows that F_δ has an L^p norm of $O(\delta^{1/p})$ on the average. Thus we have constructed functions f of arbitrarily small L^p norm whose maximal function Mf is bounded away from zero on a set of measure bounded away from zero. From this and some minor additional tricks it is not difficult to then construct a function f in L^p whose maximal function is infinite on a set of positive measure, leading to the desired contradiction.

2.9.4 The tensor power trick

We now turn to a particularly cute source of amplification, namely the tensor power operation $f \mapsto f^{\otimes M}$ which takes a complex-valued function $f : X \to \mathbf{C}$ on some set X and replaces it with a tensor power $f^{\otimes M} : X^M \to \mathbf{C}$ defined by

$$f^{\otimes M}(x_1, \ldots, x_M) := f(x_1) \ldots f(x_M).$$

If one has an estimate for which only one of the sides behaves nicely under tensor powers, then there can be some opportunity for arbitrage. For instance, suppose we wanted to prove the *Hausdorff-Young inequality*

$$\|\hat{f}\|_{l^{p'}(\hat{G})} \leq \|f\|_{L^p(G)} \tag{2.21}$$

on arbitrary finite additive groups G and all $1 \leq p \leq 2$, where $p' = p/(p-1)$ is the dual exponent of p, $\hat{G}$ is the Pontryagin dual of G (i.e. the group of characters on G), we give G normalised counting measure, and $\hat{f}(\chi) := \frac{1}{|G|} \sum_{x \in G} f(x) \overline{\chi(x)}$ is the Fourier

transform on G. If we knew the Riesz-Thorin interpolation theorem, we could quickly deduce (2.21) from the trivial inequality

$$\|\hat{f}\|_{l^\infty(\hat{G})} \leq \|f\|_{L^1(G)} \tag{2.22}$$

and the Plancherel identity

$$\|\hat{f}\|_{l^2(\hat{G})} \leq \|f\|_{L^2(G)}; \tag{2.23}$$

indeed, this is one of the textbook applications of that theorem. But suppose for some reason one did not wish to use the Riesz-Thorin theorem (perhaps in a desire to avoid "non-elementary" methods, such as complex analysis), and instead wished to use the more elementary Marcinkiewicz interpolation theorem. Then, at first glance, it appears that one can only conclude the weaker estimate

$$\|\hat{f}\|_{l^{p'}(\hat{G})} \leq C_p \|f\|_{L^p(G)}$$

for some constant $C_p > 1$. However, we can exploit the fact that the Fourier transform commutes with tensor powers. Indeed, by applying the above inequality with f replaced by $f^{\otimes M}$ (and G replaced by G^M) we see that

$$\|\hat{f}\|^M_{l^{p'}(\hat{G})} \leq C_p \|f\|^M_{L^p(G)}$$

for every $M \geq 1$; taking M^{th} roots and then letting M go to infinity we obtain (2.21); the tensor power trick has "magically" deleted the constant C_p from the inequality. More generally, one can use the tensor power trick to deduce the Riesz-Thorin interpolation theorem from the Marcinkiewicz interpolation theorem (the key point being that the (L^p, L^q) operator norm of a tensor power $T^{\otimes M}$ of a linear operator T is just the M^{th} power of the operator norm of the original operator T). This gives a proof of the Riesz-Thorin theorem that does not require complex analysis.

Actually, the tensor power trick does not just make constants disappear; it can also get rid of logarithms. Because of this, we can make the above argument even more elementary by using a very crude form of the Marcinkiewicz interpolation argument. Indeed, suppose that f is a quasi-step function, or more precisely that it is supported on some set E in G and takes values between A and $2A$ for some $A > 0$. Then from (2.22), (2.23) we see that $\|\hat{f}\|_{l^\infty(\hat{G})} = O(A|E|/|G|)$ and $\|\hat{f}\|_{l^2(\hat{G})} = O(A(|E|/|G|)^{1/2})$, and hence $\|\hat{f}\|_{l^{p'}(\hat{G})} = O(A(|E|/|G|)^{1/p})$. Now if f is not a quasi-step function, one can decompose it into $O(1+\log|G|)$ such functions by the "wedding cake" decomposition (dividing the range of $|f|$ into dyadic intervals from $\|f\|_{L^\infty}$ to $\|f\|_{L^\infty}/|G|^{100}$; the portion of $|f|$ which is less than $\|f\|_{L^\infty}/|G|^{100}$ can be easily dealt with by crude methods). From the triangle inequality we then conclude the weak Hausdorff-Young inequality

$$\|\hat{f}\|_{l^{p'}(\hat{G})} \leq C_p(1+\log|G|)\|f\|_{L^p(G)}.$$

If one runs the tensor power trick again, one can eliminate both the constant factor C_p and the logarithmic factor $1+\log|G|$ and recover (2.21) (basically because $M^{1/M}$ converges to 1 as M goes to infinity). More generally, the tensor power trick can convert restricted or weak-type estimates into strong-type estimates whenever a tensor power symmetry is available.

The deletion of the constant C_p may seem minor, but there are some things one can do with a sharp[22] estimate that one cannot with a non-sharp one. For instance, by differentiating (2.21) at $p = 2$ (where equality holds) one can obtain the *entropy uncertainty principle*

$$\sum_{\chi \in \hat{G}} |\hat{f}(\xi)|^2 \log \frac{1}{|\hat{f}(\xi)|^2} + \frac{1}{|G|} \sum_{x \in G} |f(x)|^2 \log \frac{1}{|f(x)|^2} \geq 2 \log |G|$$

whenever we have the normalisation $\|f\|_{L^2(G)} = 1$. (More generally, estimates involving Shannon entropy tend to be rather amenable to the tensor power trick.)

The tensor power trick also allows one to *disprove* certain estimates. Observe that if two functions f, g on a finite additive group G such that $|f(x)| \leq g(x)$ for all x (i.e. g majorises f), then from Plancherel's identity we have

$$\|\hat{f}\|_{l^2(\hat{G})} \leq \|\hat{g}\|_{l^2(\hat{G})}$$

and more generally (by using the fact that the Fourier transform intertwines convolution and multiplication) that

$$\|\hat{f}\|_{l^p(\hat{G})} \leq \|\hat{g}\|_{l^p(\hat{G})}$$

for all even integers $p = 2, 4, 6, \ldots$. Hardy and Littlewood conjectured that a similar bound held for all $2 \leq p < \infty$, thus

$$\|\hat{f}\|_{l^p(\hat{G})} \leq C_p \|\hat{g}\|_{l^p(\hat{G})}.$$

But if such a bound held, then by the tensor power trick one could delete the constant C_p. But then a direct computation (for instance, inspecting what happens when f is infinitesimally close to g) shows that this amplified estimate fails, and so the Hardy-Littlewood majorant conjecture is false. (With a little more work, one can then transfer this failure from finite abelian groups G to other groups, such as the unit circle $\mathbf{R}/\mathbf{Z}$ or cyclic groups $\mathbf{Z}/N\mathbf{Z}$, which do not obviously admit tensor product structure; this was first done in [Ba1973], and with stronger quantitative estimates in [MoSh2002], [GrRu2004].)

The tensor product trick is also widely used in additive combinatorics (I myself learnt this trick from [Ru1996]). Here, one deals with sets A rather than functions f, but the idea is still the same: replace A by the Cartesian power A^M, see what estimate one gets, and let $M \to \infty$. There are many instances of this trick in the literature, but I'll just describe one representative one, due to Ruzsa[Ru1996]. An important inequality of Plünnecke[Pl1969] asserts, among other things, that for finite non-empty sets A, B of an additive group G, and any positive integer k, the iterated sumset $kB = B + \ldots + B$ obeys the bound

$$|kB| \leq \frac{|A+B|^k}{|A|^{k-1}}. \tag{2.24}$$

[22] I should remark that in Euclidean space, the constant in Hausdorff-Young can be improved to below 1, but this requires some particularly Euclidean devices, such as the use of Gaussians, although this is not too dissimilar as there are certainly many connections between Gaussians and tensor products (cf. the central limit theorem). All of the above discussion also has an analogue for Young's inequality. See [Be1975] for more details.

(This inequality, incidentally, is itself proven using a version of the tensor power trick, in conjunction with Hall's marriage theorem, but never mind that here.) This inequality can be amplified to the more general inequality

$$|B_1 + \ldots + B_k| \leq \frac{|A+B_1| \ldots |A+B_k|}{|A|^{k-1}}$$

via the tensor power trick as follows. Applying (2.24) with $B := B_1 \cup \ldots \cup B_k$, we obtain

$$|B_1 + \ldots + B_k| \leq \frac{(|A+B_1| + \ldots + |A+B_k|)^k}{|A|^{k-1}}.$$

The right-hand side looks a bit too big, but this is the same problem we encountered with the Cauchy-Schwarz or Holder inequalities, and we can resolve it in a similar way (i.e. by arbitraging homogeneity). If we replace G with the larger group $G \times \mathbf{Z}^k$ and replace each set B_i with the larger set $B_i \times \{e_i, 2e_i, \ldots, N_i e_i\}$, where $e_1, \ldots, e_k$ is the standard basis for $\mathbf{Z}^k$ and N_i are arbitrary positive integers (and replacing A with $A \times \{0\}$), we obtain

$$N_1 \ldots N_k |B_1 + \ldots + B_k| \leq \frac{(N_1|A+B_1| + \ldots + N_k|A+B_k|)^k}{|A|^{k-1}}.$$

Optimising this in $N_1, \ldots, N_k$ (basically, by making the $N_i|A+B_i|$ close to constant; this is a general principle in optimisation, namely that to optimise $X+Y$ it makes sense to make X and Y comparable in magnitude) we obtain the amplified estimate

$$|B_1 + \ldots + B_k| \leq C_k \frac{|A+B_1| \ldots |A+B_k|}{|A|^{k-1}}$$

for some constant C_k; but then if one replaces $A, B_1, \ldots, B_k$ with their Cartesian powers $A^M, B_1^M, \ldots, B_k^M$, takes M^{th} roots, and then sends M to infinity, we can delete the constant C_k and recover the inequality.

2.9.5 Notes

This article was originally posted on Sep 5, 2007 at

terrytao.wordpress.com/2007/09/05

A rather different perspective on the Cauchy-Schwarz inequality can be found at

www.dpmms.cam.ac.uk/\~{}wtg10/csineq.html

Emmanuel Kowalski pointed out that Deligne's proof [De1974] of the Weil conjectures also relies on the tensor power trick. Mike Steele pointed out Landau's proof of the maximum principle $|f(z)| \leq \sup_{w \in \gamma} |f(w)|$ for holomorphic functions f in an open domain, closed curves γ in that domain, and points z in the interior of that curve, also exploited the tensor power trick, by first using the Cauchy integral formula to establish a crude bound $|f(z)| \leq C_{z,\gamma} \sup_{w \in \gamma} |f(w)|$ and then deleting the constant $C_{z,\gamma}$ using the tensor power symmetry $f \mapsto f^n$.

Thanks to furia_kucha, Van Vu, and Andy Cotton-Clay for corrections.

2.10 The crossing number inequality

Today I'd like to discuss a beautiful inequality in graph theory, namely the *crossing number inequality*. This inequality gives a useful bound on how far a given graph is from being planar, and has a number of applications, for instance to sum-product estimates. Its proof is also an excellent example of the amplification trick (Section 2.9) in action; here the main source of amplification is the freedom to pass to subobjects, which is a freedom which I didn't touch upon in the previous section. The crossing number inequality (and its proof) is well known among graph theorists but perhaps not among the wider mathematical community, so I thought I would publicise it here.

In this article, when I talk about a graph, I mean an abstract collection of vertices V, together with some abstract edges E joining pairs of vertices together. We will assume that the graph is *undirected* (the edges do not have a preferred orientation), *loop-free* (an edge cannot begin and start at the same vertex), and *multiplicity-free* (any pair of vertices is joined by at most one edge). More formally, we can model all this by viewing E as a subset of $\binom{V}{2} := \{e \subset V : |e| = 2\}$, the set of 2-element subsets of V, and we view the graph G as an ordered pair G = (V,E). (The notation is set up so that $|\binom{V}{2}| = \binom{|V|}{2}$.)

Now one of the great features of graphs, as opposed to some other abstract maths concepts, is that they are easy to draw: the abstract vertices become dots on a plane, while the edges become line segments or curves connecting these dots[23]. Let us informally refer to such a concrete representation D of a graph G as a *drawing of that graph*. Clearly, any non-trivial graph is going to have an infinite number of possible drawings. In some of these drawings, a pair of edges might cross each other; in other drawings, all edges might be disjoint (except of course at the vertices, where edges with a common endpoint are obliged to meet). If G has a drawing D of the latter type, we say that the graph G is *planar*.

Given an abstract graph G, or a drawing thereof, it is not always obvious as to whether that graph is planar; just because the drawing that you currently possess of G contains crossings, does not necessarily mean that *all* drawings of G do. The wonderful little web game "Planarity" at `www.planarity.net` illustrates this point excellently. Nevertheless, there are definitely graphs which are not planar; in particular the complete graph K_5 on five vertices, and the complete bipartite graph $K_{3,3}$ on two sets of three vertices, are non-planar.

There is in fact a famous theorem of Kuratowski[Ku1930] that says that these two graphs are the only "source" of non-planarity, in the sense that any non-planar graph contains (a subdivision of) one of these graphs as a subgraph. (There is of course the even more famous *four-colour theorem* that asserts that every planar graphs is four-colourable, but this is not the topic of my article today.)

Intuitively, if we fix the number of vertices $|V|$, and increase the number of edges $|E|$, then the graph should become "increasingly non-planar"; conversely, if we keep the same number of edges $|E|$ but spread them amongst a greater number of vertices $|V|$, then the graph should become "increasingly planar". Is there a quantitative way to

[23] To avoid some technicalities we do not allow these curves to pass through the dots, except if the curve is terminating at that dot.

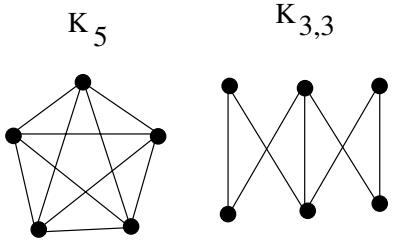


Figure 2.1: K_5 and $K_{3,3}$.

measure the "non-planarity" of a graph, and to formalise the above intuition as some sort of inequality?

It turns out that there is an elegant inequality that does precisely this, known as the *crossing number inequality* [AjChNeSz1982], [AjChNeSz1982]. Nowadays it can be proven by two elegant amplifications of Euler's formula, as we shall see.

If D is a drawing of a graph G, we define $\mathrm{cr}(D)$ to be the total number of crossings - where pairs of edges intersect at a point, for a reason other than sharing a common vertex. (If multiple edges intersect at the same point, each pair of edges counts once.) We then define the *crossing number* $\mathrm{cr}(G)$ of G to be the minimal value of $\mathrm{cr}(D)$ as D ranges over the drawings of G. Thus for instance $\mathrm{cr}(G) = 0$ if and only if G is planar. One can also verify that the two graphs K_5 and $K_{3,3}$ have a crossing number of 1. This quantity $\mathrm{cr}(G)$ will be the measure of how non-planar our graph G is. The problem is to relate this quantity in terms of the number of vertices —V— and the number of edges $|E|$. We of course do not expect an exact identity relating these three quantities (two graphs with the same number of vertices and edges may have a different number of crossing numbers), so will settle for good upper and lower bounds on $\mathrm{cr}(G)$ in terms of $|V|$ and $|E|$.

How big can the crossing number of a graph $G = (V,E)$ be? A trivial upper bound is $\mathrm{cr}(G) = O(|E|^2)$, because if we place the vertices in general position (or on a circle) and draw the edges as line segments, then every pair of edges crosses at most once. But this bound does not seem very tight; we expect to be able to find drawings in which most pairs of edges in fact do not intersect.

Let's turn our attention instead to lower bounds. We of course have the trivial lower bound $\mathrm{cr}(G) \geq 0$; can we do better? Let's first be extremely unambitious and see when one can get the minimal possible improvement on this bound, namely $\mathrm{cr}(G) > 0$. In other words, we want to find some conditions on $|V|$ and $|E|$ which will force G to be non-planar. We can turn this around by taking contrapositives: if G is planar, what does this tell us about $|V|$ and $|E|$?

Here, the natural tool is Euler's formula[24] $|V| - |E| + |F| = 2$, valid for any planar drawing, where $|F|$ is the number of faces (including the unbounded face). What do we know about $|F|$? Well, every face is adjacent to at least three edges, whereas every edge

[24]This is the one place where we shall really use the topological structure of the plane; the rest of the argument is combinatorial. There are some minor issues if the graph is disconnected, or if there are vertices of degree one or zero, but these are easily dealt with.

is adjacent to exactly two faces. By double counting the edge-face incidences, we conclude that $3|F| \leq 2|E|$. Eliminating $|F|$, we conclude that $|E| \leq 3|V| - 6$ for all planar graphs (and this bound is tight when the graph is triangular). Taking contrapositives, we conclude

$$\mathrm{cr}(G) > 0 \text{ whenever } |E| > 3|V| - 6. \tag{2.25}$$

Now, let us amplify this inequality by exploiting the freedom to delete edges. Indeed, observe that if a graph G can be drawn with only $\mathrm{cr}(G)$ crossings, then we can delete one of the crossings by removing an edge associated to that crossing, and so we can remove all the crossings by deleting at most $\mathrm{cr}(G)$ edges, leaving at least $|E| - \mathrm{cr}(G)$ edges (and $|V|$ vertices). Combining this with (2.25) we see that regardless of the number of crossings, we have

$$|E| - \mathrm{cr}(G) \leq 3|V| - 6$$

leading to the following amplification of (2.25):

$$\mathrm{cr}(G) \geq |E| - 3|V| + 6 \tag{2.26}$$

This is not the best bound, though, as one can already suspect by comparing (2.26) with the crude upper bound $\mathrm{cr}(G) = O(|E|^2)$. We can amplify (2.26) further by exploiting a second freedom, namely the ability to delete vertices. One could try the same sort of trick as before, deleting vertices which are associated to a crossing, but this turns out to be very inefficient (because deleting vertices also deletes an unknown number of edges, many of which had nothing to do with the crossing). Indeed, it would seem that one would have to be fiendishly clever to find an efficient way to delete a lot of crossings by deleting only very few vertices.

However, there is an amazing (and unintuitive) principle in combinatorics which states that when there is no obvious "best" choice for some combinatorial object (such as a set of vertices to delete), then often trying a *random* choice will give a reasonable answer, if the notion of "random" is chosen carefully. (See [Go2000] for some further discussion of this principle.) The application of this principle is known as the *probabilistic method*, first introduced by Erdős [Er1947].

Here is how it works in this current setting. Let $0 < p \leq 1$ be a parameter to be chosen later. We will randomly delete all but a fraction p of the vertices, by letting each vertex be deleted with an independent probability of $1 - p$ (and thus surviving with a probability of p). Let V' be the set of vertices that remain. Once one deletes vertices, one also has to delete the edges attached to these vertices; let E' denote the surviving edges (i.e. the edges connected to vertices in V'). Let $G' = (V', E')$ be the surviving graph (known as the subgraph of G induced by V'). Then from (2.26) we have

$$\mathrm{cr}(G') \geq |E'| - 3|V'| + 6.$$

Now, how do we get from this back to the original graph $G = (V, E)$? The quantities $|V'|$, $|E'|$, and $\mathrm{cr}(G')$ all fluctuate randomly, and are difficult to compute. However, their *expectations* are much easier to deal with. Accordingly, we take expectations of both sides (this is an example of the *first moment method*). Using linearity of expectation, we have

$$\mathbf{E}(\mathrm{cr}(G')) \geq \mathbf{E}(|E'|) - 3\mathbf{E}(|V'|) + 6.$$

These quantities are all relatively easy to compute. The easiest is $\mathbf{E}(|V'|)$. Each vertex in V has a probability p of ending up in V', and thus contributing 1 to —V'—. Summing up (using linearity of expectation again), we obtain $\mathbf{E}(|V'|) = p|V|$.

The quantity $\mathbf{E}(|E'|)$ is almost as easy to compute. Each edge e in E will have a probability p^2 of ending up in E', since both vertices have an independent probability of p of surviving. Summing up, we obtain $\mathbf{E}(|E'|) = p^2|E|$. (The events that each edge ends up in E' are not quite independent, but the great thing about linearity of expectation is that it works even without assuming any independence.)

Finally, we turn to $\mathbf{E}(\mathrm{cr}(G'))$. Let us draw G in the optimal way, with exactly $\mathrm{cr}(G)$ crossings. Observe that each crossing involves two edges and four vertices. (If the two edges involved in a crossing share a common vertex as well, thus forming an α shape, one can reduce the number of crossings by 1 by swapping the two halves of the loop in the α shape. So with the optimal drawing, the edges in a crossing do not share any vertices in common.) Passing to G', we see that the probability that the crossing survives in this drawing is only p^4. By one last application of linearity of expectation, the expected number of crossings of this diagram that survive for G' is $p^4\mathrm{cr}(G)$. This particular diagram may not be the optimal one for G', so we end up with an inequality $\mathbf{E}\mathrm{cr}(G') \leq p^4\mathrm{cr}(G)$. Fortunately for us, this inequality goes in the right direction, and we get a useful inequality:

$$p^4\mathrm{cr}(G) \geq p^2|E| - 3p|V| + 6.$$

In terms of $\mathrm{cr}(G)$, we have

$$\mathrm{cr}(G) \geq p^{-2}|E| - 3p^{-3}|V| + 6p^{-4}.$$

To finish the amplification, we need to optimise in p, subject of course to the restriction $0 < p \leq 1$, since p is a probability. To solve the optimisation problem exactly, one needs to solve a cubic; but we can perform a much cheaper computation by settling for a bound which is close to the optimal bound rather than exactly equal to it. A general principle is that optima are often obtained when two of the terms are roughly in balance. A bit of thought reveals that it might be particularly good to have $3p^{-3}|V|$ just barely smaller than $p^{-2}|E|$. (If it is a lot smaller, then p will be large, and we don't get a good bound on the right. If instead $3p^{-3}|V|$ is a lot bigger, then we are likely to have a negative right-hand side.) For instance, we could choose p so that $4p^{-3}|V| = p^{-2}|E|$; this is legal as long as $|E| \geq 4|V|$. Substituting this (and discarding the lower-order term $6p^{-4}$) we obtain the *crossing number inequality*

$$\mathrm{cr}(G) \geq \frac{|E|^3}{64|V|^2} \text{ whenever } |E| \geq 4|V|. \tag{2.27}$$

This is quite a strong amplification of (2.25) or (2.26) (except in the transition region in which $|E|$ is comparable to $|V|$). Is it sharp? We can compare it against the trivial bound $\mathrm{cr}(G) = O(|E|^2)$, and we observe that the two bounds match up to constants when $|E|$ is comparable to $|V|^2$. (Clearly, $|E|$ cannot be larger than $|V|^2$.) So the crossing number inequality is sharp (up to constants) for dense graphs, such as the complete graph K_n on n vertices.

Are there any other cases where it is sharp? We can answer this by appealing to the symmetries of (2.27). By the nature of its proof, the inequality is basically symmetric under passage to random induced subgraphs, but this symmetry does not give any further examples, because random induced subgraphs of dense graphs again tend to be dense graphs (cf. the computation of $\mathbf{E}|V'|$ and $\mathbf{E}|E'|$ above). But there is a second symmetry of (2.27) available, namely that of *replication*. If one takes k disjoint copies of a graph $G = (V,E)$, one gets a new graph with $k|V|$ vertices and $k|E|$ edges, and a moment's thought will reveal that the new graph has a crossing number of $k\mathrm{cr}(G)$. Thus replication is a symmetry of (2.27). Thus, (2.27) is also sharp up to constants for replicated dense graphs. It is not hard to see that these examples basically cover all possibilities of $|V|$ and $|E|$ for which $|E| \geq 4|V|$. Thus the crossing number inequality cannot be improved except for the constants. (The best constants known currently can be found in [PaRaTaTo2006].)

Remark 2.38. A general principle, by the way, is that one can roughly gauge the "strength" of an inequality by the number of independent symmetries (or approximate symmetries) it has. If for instance there is a three-parameter family of symmetries, then any example that demonstrates that sharpness of that inequality is immediately amplified to a three-parameter family of such examples (unless of course the example is fixed by a significant portion of these symmetries). The more examples that show an inequality is sharp, the more efficient it is - and the harder it is to prove, since one cannot afford to lose anything (other than perhaps some constants) in every one of the sharp example cases. This principle is of course consistent with the points in my previous article (Section 2.9) on arbitraging a weak asymmetric inequality into a strong symmetric one.

2.10.1 Application: the Szemerédi-Trotter theorem

It was noticed by Székely[Sz1997] that the crossing number is powerful enough to give easy proofs of several difficult inequalities in combinatorial incidence geometry. For instance, the Szemerédi-Trotter theorem concerns the number of incidences $I(P,L) := |\{(p,l) \in P \times L : p \in l\}|$ between a finite collection of points P and lines L in the plane. For instance, the three lines and three points of a triangle form six incidences; the five lines and ten points of a pentagram form 20 incidences; and so forth.

One can ask the question: given $|P|$ points and $|L|$ lines, what is the maximum number of incidences $I(P,L)$ one can form between these points and lines? (The minimum number is obviously 0, which is a boring answer.) The trivial bound is $I(P,L) \leq |P||L|$, but one can do better than this, because it is not possible for every point to lie on every line. Indeed, if we use nothing more than the axiom that every two points determine at most one line, combined with the Cauchy-Schwarz inequality, it is not hard to show (by double-counting the space of triples $(p,p',l) \in P \times P \times L$ such that $p,p' \in l$) that

$$|I(P,L)| \leq |P||L|^{1/2} + |L| \tag{2.28}$$

Dually, using the axiom that two lines intersect in at most one point, we obtain

$$|I(P,L)| \leq |L||P|^{1/2} + |P|. \tag{2.29}$$

(One can also deduce one inequality from the other by projective duality.)

Can one do better? The answer is yes, if we observe that a configuration of points and lines naturally determines a drawing of a graph, to which the crossing number can be applied. To see this, assume temporarily that every line in L is incident to at least two points in P. A line l in L which is incident to k points in P will thus contain $k-1$ line segments in P; $k-1$ is comparable to k. Since the sum of all the k is $I(P,L)$ by definition, we see that there are roughly $I(P,L)$ line segments of L connecting adjacent points in P; this is a diagram with $|P|$ vertices and roughly $|I(P,L)|$ edges. On the other hand, a crossing in this diagram can only occur when two lines in L intersect. Since two lines intersect in at most one point, the total number of crossings is $O(|L|^2)$. Applying the crossing number inequality (2.27), we obtain

$$|L|^2 \gg I(P,L)^3/|P|^2$$

if $I(P,L)$ is much larger than $|P|$, which leads to

$$I(P,L) = O(|L|^{2/3}|P|^{2/3} + |P|).$$

We can then remove our temporary assumption that lines in L are incident to at least two points, by observing that lines that are incident to at most one point will only contribute $O(|L|)$ incidences, leading to the *Szemerédi-Trotter theorem*

$$I(P,L) = O(|L|^{2/3}|P|^{2/3} + |P| + |L|).$$

This bound is somewhat stronger than the previous bounds, and is in fact surprisingly sharp; a typical example that demonstrates this is when P is the lattice $\{1,\ldots,N\} \times \{1,\ldots,N^2\}$ and L is the set of lines $\{(x,y) : y = mx+b\}$ with slope $m \in \{1,\ldots,N\}$ and intercept $b \in \{1,\ldots,N^2\}$; here $|P| = |L| = N^3$ and the number of incidences is roughly N^4.

The original proof of this theorem, by the way, proceeded by amplifying (2.28) using the method of *cell decomposition*; it is thus somewhat similar in spirit to Szekély's proof, but was a bit more complicated technically. In [Wo1999], Wolff conjectured a continuous version of this theorem for fractal sets, sometimes called the *Furstenberg set conjecture*, and related to the Kakeya conjecture; a small amount of progress beyond the analogue of (2.28) is known [KaTa2001], [Bo2003], but we are still far from the best possible result here.

2.10.2 Application: sum-product estimates

One striking application of the Szemerédi-Trotter theorem (and by extension, the crossing number inequality) is to the arena of *sum-product* estimates in additive combinatorics, which is currently a very active area of research, especially in finite fields, due to its connections with some long-standing problems in analytic number theory, as well as to some computer science problems concerning randomness extraction and expander graphs. However, our focus here will be on the more classical setting of sum-product estimates in the real line $\mathbf{R}$.

Let A be a finite non-empty set of non-zero real numbers. We can form the *sum set*

$$A+A := \{a+b : a,b \in A\}$$

and the *product set*

$$A \cdot A = \{ab : a,b \in A\}.$$

If A is in "general position", it is not hard to see that $A+A$ and $A \cdot A$ both have cardinality comparable to $|A|^2$. However, in certain cases one can make one or the other sets significantly smaller. For instance, if A is an arithmetic progression $\{a, a+r, \ldots, a+(k-1)r\}$, then the sum set $A+A$ has cardinality comparable to just $|A|$. Similarly, if A is a geometric progression $\{a, ar, \ldots, ar^{k-1}\}$, then the product set $A \cdot A$ has cardinality comparable to $|A|$. But clearly A cannot be an arithmetic progression and a geometric progression at the same time (unless it is very short). So one might conjecture that at least one of the sum set and product set should be significantly larger than A. Informally, this is saying that no finite set of reals can behave much like a subring of $\mathbf{R}$. This intuition was made precise by Erdős and Szemerédi [ErSz1983], who established the lower bound

$$\max(|A+A|, |A \cdot A|) \gg |A|^{1+c}$$

for some small $c > 0$ which they did not make explicit. They then conjectured that in fact c should be made arbitrary close to the optimal value of 1, and more precisely that

$$\max(|A+A|, |A \cdot A|) \gg |A|^2 \exp(-\delta \log |A| / \log\log |A|)$$

for large —A— and some absolute constant $\delta > 0$. (The exponential factor is sharp, as can be seen from setting $A = \{1, \ldots, N\}$, and using some analytic number theory to control the size of $A \cdot A$.)

The Erdős-Szemerédi conjecture remains open, however the value of c has been improved; currently, the best bound is due to Solymosi[So2005], who showed that c can be arbitrarily close to $3/11$. Solymosi's argument is based on an earlier argument of Elekes[El1997], who obtained $c = 1/4$ by a short and elegant argument based on the Szemerédi-Trotter theorem which we will now present. The basic connection between the two problems stems from the familiar formula $y = mx+b$ for a line, which clearly encodes a multiplicative and additive structure. We already used this connection implicitly in the example that demonstrated that the Szemerédi-Trotter theorem was sharp. For Elekes' argument, the challenge is to show that if $A+A$ and $A \cdot A$ are both small, then a suitable family of lines $y = mx+b$ associated to A will have a high number of incidences with some set of points associated to A, so that the Szemerédi-Trotter may then be profitably applied. It is not immediately obvious exactly how to do this, but Elekes settled upon the choice of letting $P := (A+A) \times (A \cdot A)$, and letting L be the space of lines $y = mx+b$ with slope in A^{-1} and intercept in A, thus $|P| = |A+A||A \cdot A|$ and $|L| = |A|^2$. One observes that each line in L is incident to —A— points in P, leading to $|A|^3$ incidences. Applying the Szemerédi-Trotter theorem and doing the algebra one eventually concludes that $\max(|A+A|, |A \cdot A|) \gg |A|^{5/4}$. (A more elementary proof of this inequality, not relying on the Szemerédi-Trotter theorem or crossing number bounds, and thus having the advantage on working on other archimedean fields such as $\mathbf{C}$, was subsequently found by Solymosi [So2008], but the best bounds on the sum-product problem in $\mathbf{R}$ still rely very much on the Szemerédi-Trotter inequality.)

2.11 Ratner's theorems

While working on a recent paper with Ben Green[GrTa2008f], I was introduced to the beautiful theorems of Marina Ratner[Ra1991] on unipotent flows on homogeneous spaces, and their application to questions in number theory, such as the *Oppenheim conjecture* (first solved by Margulis[Ma1989]). This is a subject that I am still only just beginning to learn, but hope to understand better in the future, especially given that quantitative analogues of Ratner's theorems should exist, and should have even more applications to number theory (see for instance the recent paper [EiMaVe2008]). In this post, I will try to describe some of the background for this theorem and its connection with the Oppenheim conjecture; I will not discuss the proof at all, largely because I have not fully understood it myself yet. For a nice introduction to these issues, I recommend Dave Morris' book[Mo2005] on the subject (and this article here is drawn in large part from that book).

Ratner's theorem takes place on a *homogeneous space*. Informally, a homogeneous space is a space X which looks "the same" when viewed from any point on that space. For instance, a sphere S^2 is a homogeneous space, but the surface of a cube is not (the cube looks different when viewed from a corner than from a point on an edge or on a face). More formally, a homogeneous space is a space X equipped with an action $(g,x) \mapsto gx$ of a group G of symmetries which is *transitive*: given any two points $x,y \in X$ on the space, there is at least one symmetry $g \in G$ that moves x to y, thus $y = gx$. (For instance the cube has several symmetries, but not enough to be transitive; in contrast, the sphere S^2 has the transitive action of the special orthogonal group $SO(3)$ as its symmetry group.) It is not hard to see that a homogeneous space X can always be identified (as a *G-set*, i.e. a set with an action of G) with a quotient $G/\Gamma := \{g\Gamma : g \in G\}$, where Γ is a subgroup of G; indeed, one can take Γ to be the stabiliser $\Gamma := \{g \in G : gx = x\}$ of an arbitrarily chosen point $x \in X$, and then identify $g\Gamma$ with $g\Gamma x = gx$. For instance, the sphere S^2 has an obvious action of the special orthogonal group $SO(3)$, and the stabiliser of (say) the north pole can be identified with $SO(2)$, so that the sphere can be identified with $SO(3)/SO(2)$. More generally, any Riemannian manifold of constant curvature is a homogeneous space; for instance, an m-dimensional torus can be identified with $\mathbf{R}^m/\mathbf{Z}^m$, while a surface X of constant negative curvature can be identified with $SL(2,\mathbf{R})/\Gamma$ for some subgroup Γ of $SL(2,\mathbf{R})$ (e.g. the hyperbolic plane $\mathbf{H}$ is isomorphic to $SL(2,\mathbf{R})/SO(2)$). Furthermore, the *cosphere bundle* S^*X of X - the space of unit (co)tangent vectors on X - is also a homogeneous space with structure group $SL(2,\mathbf{R})$. (For instance, the cosphere bundle $S^*\mathbf{H}$ of the hyperbolic plane $\mathbf{H}$ is isomorphic to $SL(2,\mathbf{R})/\{+1,-1\}$.)

For the purposes of Ratner's theorem, we only consider homogeneous spaces X in which the symmetry group G is a connected finite-dimensional Lie group, and X is *finite volume* (or more precisely, it has a finite non-trivial G-invariant measure). Every compact homogeneous space is finite volume, but not conversely; for instance the *modular curve* $SL(2,\mathbf{R})/SL(2,\mathbf{Z})$ is finite volume but not compact (it has a cusp). (The modular curve has two real dimensions, but just one complex dimension, hence the term "curve"; rather confusingly, it is also referred to as the "modular surface". As for the term "modular", observe that the moduli space of *unimodular lattices* in $\mathbf{R}^2$ has an obvious action of $SL(2,\mathbf{R})$, with the stabiliser of $\mathbf{Z}^2$ being $SL(2,\mathbf{Z})$, and so this moduli

space can be identified with the modular curve.)

Let $U \leq G$ be a subgroup of G. The group U then acts on X, creating an orbit $Ux := \{gx : g \in U\}$ inside X for every point x in X. Even though X "looks the same" from every point, the orbits of U need not all look alike, basically because we are not assuming U to be a normal subgroup (i.e. $Ug \neq gU$ in general). For instance on the surface of the earth, which we model as a sphere $S^2 = SO(3)/SO(2)$, if we let $U \cong SO(2)$ be the group of rotations around the Earth's axis, then the orbits Ux are nothing more than the circles of latitude, together with the north and south poles as singleton orbits.

In the above example, the orbits were closed subsets of the space X. But this is not always the case. Consider for instance the 2-torus $X := \mathbf{R}^2/\mathbf{Z}^2$, and let $U \leq \mathbf{R}^2$ be a line $U := \{(x, \alpha x) : x \in \mathbf{R}\}$. Then if the slope α of this line is irrational, the orbit Ux of a point x in the torus will be a dense one-dimensional subset of that two-dimensional torus, and thus definitely not closed. More generally, when considering the orbit of a subspace $U \leq \mathbf{R}^m$ on a torus $\mathbf{R}^m/\mathbf{Z}^m$, the orbit Ux of a point x will always be a dense subset of some subtorus of $\mathbf{R}^m/\mathbf{Z}^m$ (this is essentially Kronecker's theorem on simultaneous approximation by rationals).

From these examples we see that even if an orbit Ux is not closed, its closure $\overline{Ux}$ is fairly "nice" - indeed, in all of the above cases, the closure can be written as a closed orbit Hx of some other group $U \leq H \leq G$ intermediate between U and G.

Unfortunately, this nice state of affairs is not true for arbitrary flows on homogeneous spaces. A classic example is geodesic flow on surfaces M of constant negative curvature (such as the modular curve mentioned earlier). This flow can be viewed as an action of $\mathbf{R}$ (representing time) on the cosphere bundle S^*M (which represents the state space of a particle on M moving at unit speed), which is a homogeneous space with symmetry group $SL(2,\mathbf{R})$. In this example, the subgroup $U \leq SL(2,\mathbf{R})$ is given as

$$U := \{\begin{pmatrix} e^t & 0 \\ 0 & e^{-t} \end{pmatrix} : t \in \mathbf{R}\} \cong \mathbf{R}. \tag{2.30}$$

For certain surfaces, this flow is quite chaotic, for instance Morse [Mo1921] produced an example of a geodesic flow on a constant negative curvature surface whose closed orbit $\overline{Ux}$ had cross-sections that were homeomorphic to a Cantor set. (For the modular curve, there is an old result of Artin [Ar1929] that exhibits an orbit which is dense in the whole curve, but I don't know if one can obtain Cantor-like behaviour in this curve. There is also a connection between geodesic flow on this curve and continued fractions; see [KaUg2007].)

The reason for the "badness" of the above examples stems from the exponential instabilities present in the action of U, which can already be suspected from the presence of the exponential in (2.30). (Exponential instability is not a sufficient condition for chaos, but is often a necessary one.) Ratner's theorems assert, very roughly speaking, that if one eliminates all exponential behaviour from the group U, then the orbits Ux become nicely behaved again; they are either closed, or are dense in larger closed orbits Hx.

What does it mean to eliminate "all exponential behaviour"? Consider a one-dimensional matrix group

$$U = \{A^t : t \in \mathbf{R}\}$$

where A is a matrix with some designated logarithm $\log(A)$, and $A^t := \exp(t\log(A))$. Generally, we expect the coefficients of A^t to contain exponentials (as is the case in (2.30)), or sines and cosines (which are basically just a complex version of exponentials). However, if A is a *unipotent matrix* (i.e. the only eigenvalue is 1, or equivalently that $A = 1+N$ for some nilpotent matrix N), then A^t is a polynomial in t, rather than an exponential or sinusoidal function of t. More generally, we say that an element g of a Lie group G is unipotent if its adjoint action $x \mapsto gxg^{-1}$ on the Lie algebra $\mathfrak{g}$ is unipotent. Thus for instance any element in the centraliser of G is unipotent, and every element of a nilpotent group is unipotent.

We can now state one of Ratner's theorems.

Theorem 2.39 (Ratner's orbit closure theorem)**.** *Let $X = G/\Gamma$ be a homogeneous space of finite volume with a connected finite-dimensional Lie group G as symmetry group, and let U be a connected subgroup of G generated by unipotent elements. Let Ux be an orbit of U in X. Then the closure $\overline{Ux}$ is itself a homogeneous space of finite volume; in particular, there exists a closed subgroup $U \leq H \leq G$ such that $\overline{Ux} = Hx$.*

This theorem (first conjectured by Raghanuthan, I believe) asserts that the orbit of any unipotent flow is dense in some homogeneous space of finite volume. In the case of algebraic groups, it has a nice corollary: any unipotent orbit in an algebraic homogeneous space which is Zariski dense, is topologically dense as well.

In some applications, density is not enough; we also want *equidistribution*. Happily, we have this also:

Theorem 2.40 (Ratner's equidistribution theorem)**.** *Let X, G, U, x, H be as in the orbit closure theorem. Assume also that U is a one-parameter group, thus $U = \{g_t : t \in \mathbf{R}\}$ for some homomorphism $t \mapsto g_t$. Then Ux is equidistributed in Hx; thus for any continuous function $F : Hx \to \mathbf{R}$ we have*

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T F(g_t x)\, dt = \int_{Hx} F$$

where $\int_{Hx}$ represents integration on the normalised Haar measure on Hx.

One can also formulate this theorem (first conjectured by Dani[Da1986], I believe) for groups U that have more than one parameter, but it is a bit technical to do so and we shall omit it. My paper [GrTa2008f] with Ben Green concerns a quantitative version of this theorem in the special case when X is a nilmanifold, and where the continuous orbit Ux is replaced by a discrete polynomial sequence. (There is an extensive literature on generalising Ratner's theorems from continuous U to discrete U, which I will not discuss here.)

From the equidistribution theorem and a little bit of ergodic theory one has a measure-theoretic corollary, which describes ergodic measures of a group generated by unipotent elements:

Theorem 2.41 (Ratner's measure classification theorem)**.** *Let X be a finite volume homogeneous space for a connected Lie group G, and let U be a connected subgroup of G generated by unipotent elements. Let μ be a probability measure on X which is ergodic under the action of U. Then μ is the Haar measure of some closed finite volume orbit Hx for some $U \leq H \leq G$.*

2.11.1 The Oppenheim conjecture

To illustrate the power of Ratner's orbit closure theorem, we discuss the first major application of this theorem, namely to solve the Oppenheim conjecture. (Margulis' solution [Ma1989] of the Oppenheim conjecture actually predates Ratner's papers by a year or two, but Margulis solved the conjecture by establishing a special case of the orbit closure theorem.) I will not discuss applications of the other two theorems of Ratner here.

The Oppenheim conjecture concerns the possible value of quadratic forms in more than one variable, when all the variables are restricted to be integer. For instance, the famous four squares theorem of Lagrange asserts that the set of possible values of the quadratic form

$$Q(n_1,n_2,n_3,n_4) := n_1^2+n_2^2+n_3^2+n_4^2,$$

where n_1,n_2,n_3,n_4 range over the integers, are precisely the natural numbers $\{0,1,2,\ldots\}$.

More generally, if Q is a positive definite quadratic form in m variables, possibly with irrational coefficients, then the set $Q(\mathbf{Z}^m)$ of possible values of Q can be easily seen to be a discrete subset of the positive real axis. I can't resist mentioning here a beautiful theorem of Bhargava and Hanke: if a positive-definite quadratic form with integer coefficients represents all positive integers up to 290, then it in fact represents all positive integers. If the off-diagonal coefficients are even, one only needs to represent the integers up to 15; this was first done by Conway and Schneeberger[Co2000].

What about if Q is indefinite? Then a number of things can happen. If Q has integer coefficients, then clearly $Q(\mathbf{Z}^m)$ must take integer values, and can take arbitrarily large positive or negative such values, but can have interesting gaps in the representation. For instance, the question of which integers are represented by $Q(n_1,n_2) := n_1^2 - dn_2^2$ for some integer d already involves a little bit of class field theory of $\mathbf{Q}(\sqrt{-d})$, and was first worked out by Gauss.

Similar things can happen of course if Q has commensurate coefficients, i.e. Q has integer coefficients after dividing out by a constant. What if Q has incommensurate coefficients? In the two-variable case, we can still have some discreteness in the representation. For instance, if $\phi := \frac{1+\sqrt{5}}{2}$ is the golden ratio, then the quadratic form

$$Q(n_1,n_2) = n_1^2 - \phi^2 n_2^2 = (n_1 - \phi n_2)(n_1 + \phi n_2)$$

cannot get arbitrarily close to 0, basically because the golden ratio is very hard to approximate by a rational a/b (the best approximants being given, of course, by the Fibonacci numbers).

However, for indefinite quadratic forms Q of three or more variables $m \geq 3$ with incommensurate coefficients, Oppenheim[Op1929] conjectured that there was no discreteness whatsoever, and that the set $Q(\mathbf{Z}^m)$ was dense in $\mathbf{R}$. There was much partial progress on this problem in the case of many variables (in large part due to the power of the Hardy-Littlewood circle method in this setting), but the hardest case of just three variables was only solved in by Margulis[Ma1989] in 1989.

Nowadays, we can obtain Margulis' result as a quick consequence of Ratner's theorem as follows. It is not difficult to reduce to the most difficult case $m = 3$. We need to show that the image of $\mathbf{Z}^3$ under the quadratic form $Q : \mathbf{R}^3 \to \mathbf{R}$ is dense in

$\mathbf{R}$. Now, every quadratic form comes with a special orthogonal group $SO(Q)$, defined as the orientation-preserving linear transformations that preserve Q; for instance, the Euclidean form $x_1^2+x_2^2+x_3^2$ in $\mathbf{R}^3$ has the rotation group $SO(3)$, the Minkowski form $x_1^2+x_2^2+x_3^2-x_4^2$ has the Lorentz group $SO(3,1)$, and so forth. The image of $\mathbf{Z}^3$ under Q is the same as that of the larger set $SO(Q)\mathbf{Z}^3$. [We may as well make our domain as large as possible, as this can only make our job easier, in principle at least.] Since Q is indefinite, $Q(\mathbf{R}^3)=\mathbf{R}$, and so it will suffice to show that $SO(Q)\mathbf{Z}^3$ is dense in $\mathbf{R}^3$. Actually, for minor technical reasons it is convenient to just work with the identity component $SO(Q)^+$ of $SO(Q)$ (which has two connected components).

[An analogy with the Euclidean case $Q(x_1,x_2,x_3)=x_1^2+x_2^2+x_3^2$ might be enlightening here. If one spins around the lattice $\mathbf{Z}^3$ by the Euclidean orthogonal group $SO(Q)=SO(3)$, one traces out a union of spheres around the origin, where the radii of the spheres are precisely those numbers whose square can be expressed as the sum of three squares. In this case, $SO(Q)\mathbf{Z}^3$ is not dense, and this is reflected in the fact that not every number is the sum of three perfect squares. The Oppenheim conjecture asserts instead that if you spin a lattice by an irrational Lorentz group, one traces out a dense set.]

In order to apply Ratner's theorem, we will view $SO(Q)^+\mathbf{Z}^3$ as an orbit Ux in a symmetric space G/Γ. Clearly, U should be the group $SO(Q)^+$, but what to do about the set $\mathbf{Z}^3$? We have to turn it somehow into a point in a symmetric space. The obvious thing to do is to view $\mathbf{Z}^3$ as the zero coset (i.e. the origin) in the torus $\mathbf{R}^3/\mathbf{Z}^3$, but this doesn't work, because $SO(Q)^+$ does not act on this torus (it is not a subgroup of $\mathbf{R}^3$). So we need to lift up to a larger symmetric space G/Γ, with a symmetry group G which is large enough to accommodate $SO(Q)^+$.

The problem is that the torus is the moduli space for translations of the lattice $\mathbf{Z}^3$, but $SO(Q)^+$ is not a group of translations; it is instead a group of unimodular linear transformations, i.e. a subgroup of the special linear group $SL(3,\mathbf{R})$. This group acts on lattices, and the stabiliser of $\mathbf{Z}^3$ is $SL(3,\mathbf{Z})$. Thus the right homogeneous space to use here is $X := SL(3,\mathbf{R})/SL(3,\mathbf{Z})$, which has a geometric interpretation as the moduli space of unimodular lattices in $\mathbf{R}^3$ (i.e. a higher-dimensional version of the modular curve); X is not compact, but one can verify that X has finite volume, which is good enough for Ratner's theorem to apply. Since the group $G=SL(3,\mathbf{R})$ contains $U=SO(Q)^+$, U acts on X. Let $x=SL(3,\mathbf{Z})$ be the origin in X (under the moduli space interpretation, x is just the standard lattice $\mathbf{Z}^3$). If Ux is dense in X, this implies that the set of matrices $SO(Q)^+SL(3,\mathbf{Z})$ is dense in $SL(3,\mathbf{R})$; applying this to, say, the unit vector $(1,0,0)$, we conclude that $SO(Q)^+\mathbf{Z}^3$ is dense in $\mathbf{R}^3$ as required. (These reductions are due to Raghunathan. Note that the new claim is actually a bit stronger than the original Oppenheim conjecture; not only are we asserting now that $SO(Q)^+$ applied to the standard lattice $\mathbf{Z}^3$ sweeps out a dense subset of Euclidean space, we are saying the stronger statement that one can use $SO(Q)^+$ to bring the standard lattice "arbitrarily close" to any given unimodular lattice one pleases, using the topology induced from $SL(3,\mathbf{R})$.)

How do we show that Ux is dense in X? We use Ratner's orbit closure theorem! This theorem tells us that if Ux is *not* dense in X, it must be much smaller - it must be contained in a closed finite volume orbit Hx for some proper closed connected subgroup H of $SL(3,\mathbf{R})$ which still contains $SO(Q)^+$. [To apply this theorem, we need

to check that U is generated by unipotent elements, which can be done by hand; here is where we need to assume $m \geq 3$.] An inspection of the Lie algebras of $SL(3,\mathbf{R})$ and $SO(Q)^+$ shows in fact that the only such candidate for H is $SO(Q)^+$ itself (here is where we really use the hypothesis $m = 3$!). Thus $SO(Q)^+x$ is closed and finite volume in X, which implies that $SO(Q)^+ \cap SL(3,\mathbf{Z})$ is a lattice in $SO(Q)^+$. Some algebraic group theory (specifically, the *Borel density theorem*) then shows that $SO(Q)^+$ lies in the Zariski closure of $SO(Q)^+ \cap SL(3,\mathbf{Z})$, and in particular is definable over $\mathbf{Q}$. It is then not difficult to see that the only way this can happen is if Q has rational coefficients (up to scalar multiplication), and the Oppenheim conjecture follows.

2.11.2 Notes

This article was originally posted on September 29, 2007 at

terrytao.wordpress.com/2007/09/29

Thanks to Matheus for providing the reference [KaUg2007]. Thanks also to Elon Lindenstrauss for some corrections.

2.12 Unipotent elements of the Lorentz group, and conic sections

In my discussion of the Oppenheim conjecture[Op1929] in my recent article on Ratner's theorems (Section 2.11), I mentioned in passing the simple but crucial fact that the (orthochronous) special orthogonal group $SO(Q)^+$ of an indefinite quadratic form on $\mathbf{R}^3$ can be generated by unipotent elements. This is not a difficult fact to prove, as one can simply diagonalise Q and then explicitly write down some unipotent elements (the magic words here are "null rotations"). But this is a purely algebraic approach; I thought it would also be instructive to show the geometric (or dynamic) reason for why unipotent elements appear in the orthogonal group of indefinite quadratic forms in three dimensions. (I'll give away the punch line right away: it's because the parabola is a conic section.) This is not a particularly deep or significant observation, and will not be surprising to the experts, but I would like to record it anyway, as it allows me to review some useful bits and pieces of elementary linear algebra.

2.12.1 Unipotent matrices

Before we get to unipotent elements of a group, let us first understand geometrically what a unipotent matrix (or linear transformation) A is. Suppose we consider an orbit $x_n = A^n x$ of some initial vector x with respect to this transformation A (thus x_n is a linear recurrence sequence). How does x_n behave geometrically as $n \to \infty$?

Despite the simple and explicit description $x_n = A^n x$ of the orbit, the geometric behaviour can be rather complicated, depending crucially on the spectrum of A (and, to a lesser extent, on the choice of x). If for instance A has an eigenvalue λ with $\lambda > 1$, and x is an eigenvector of A with eigenvalue λ, then we will of course have $x_n = \lambda^n x_0$, thus this orbit will grow exponentially. Similarly, if one has an eigenvalue between 0 and 1, then it is possible for the orbit to decay exponentially.

If one has eigenvalues with a complex phase, one can have oscillation. If for instance A is the rotation matrix $A = R_\theta := \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$ corresponding to anticlockwise rotation around the origin by some non-trivial angle θ (and which has complex eigenvalues $e^{i\theta}$ and $e^{-i\theta}$), and (say) $x_0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, then the orbit $x_n = \begin{pmatrix} \cos n\theta \\ \sin n\theta \end{pmatrix}$ will oscillate around the unit circle indefinitely.

If an eigenvalue has non-trivial magnitude and non-trivial phase, one gets a combination of exponential growth or decay and oscillation, leading for instance to orbits which follow a logarithmic spiral (this will be the case for instance if $A = \lambda R_\theta$ for some rotation matrix R_θ and some dilation factor $\lambda > 0$).

One can have even more complicated behaviour if there are multiple eigenvalues in play. Consider for instance the matrix $A := \begin{pmatrix} \lambda & 0 \\ 0 & 1/\lambda \end{pmatrix}$ with $\lambda > 1$, with the initial vector $x := \begin{pmatrix} y \\ z \end{pmatrix}$ with both y and z non-zero (so that x has a non-trivial presence in

both the unstable and stable modes of A). Then the orbit $x_n = \begin{pmatrix} \lambda^n y \\ \lambda^{-n} z \end{pmatrix}$ will expand exponentially in the unstable mode and contract exponentially in the stable mode, and the orbit will lie along the rectangular hyperbola $\{\begin{pmatrix} a \\ b \end{pmatrix} : ab = yz\}$.

As the above examples show, orbits of linear transformations can exhibit a variety of behaviours, from exponential growth to exponential decay to oscillation to some combination of all three. But there is one special case in which the behaviour is much simpler, namely that the orbit remains polynomial. This occurs when A is a unipotent matrix, i.e. $A = I + N$ where N is nilpotent (i.e. $N^m = 0$ for some finite m). A typical example of a unipotent matrix is

$$A = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix} \tag{2.31}$$

(and indeed, by the Jordan normal form (see Section 2.13), all unipotent matrices are similar to direct sums of matrices of this type). For unipotent matrices, the binomial formula terminates after m terms to obtain a polynomial expansion for A^n:

$$\begin{aligned} A^n &= (I+N)^n \\ &= I + nN + \frac{n(n-1)}{2}N^2 + \ldots + \frac{n\ldots(n-m+2)}{(m-1)!}N^{m-1}. \end{aligned}$$

From this we easily see that, regardless of the choice of initial vector x, the coefficients of x_n are polynomial in n. (Conversely, if the coefficients of x_n are polynomial in n for every x, it is not hard to show that A is unipotent; I'll leave this as an exercise.) It is instructive to see what is going on at the coefficient level, using the matrix (2.31) as an example. If we express the orbit x_n in coordinates as $x_n = \begin{pmatrix} a_n \\ b_n \\ c_n \end{pmatrix}$, then the recurrence $x_{n+1} = Ax_n$ becomes

$$\begin{aligned} a_{n+1} &= a_n + b_n \\ b_{n+1} &= b_n + c_n \\ c_{n+1} &= c_n. \end{aligned}$$

We thus see that the sequence c_n is constant, the sequence b_n grows linearly, and a_n grows quadratically, so the whole orbit x_n has polynomial coefficients. If one views the recurrence $x_{n+1} = Ax_n$ as a dynamical system, the polynomial nature of the dynamics are caused by the absence of (both positive and negative) feedback loops: c affects b, and b affects a, but there is no loop in which a component ultimately affects itself, which is the source of exponential growth, exponential decay, and oscillation. Indeed, one can view this absence of feedback loops as a *definition* of unipotence.

For the purposes of proving a dynamical theorem such as Ratner's theorem, unipotence is important for several reasons. The lack of exponential growing modes means that the dynamics is not exponentially unstable going forward in time; similarly, the

lack of exponentially decaying modes means that the dynamics is not exponentially unstable going backward in time. The lack of oscillation does not improve the stability further, but it does have an important effect on the smoothness of the dynamics. Indeed, because of this lack of oscillation, orbits which are polynomial in nature obey an important dichotomy: either they go to infinity, or they are constant. There is a quantitative version of this statement, known as Bernstein's inequality: if a polynomial remains bounded over a long interval, then its derivative is necessarily small. (From a Fourier-analytic perspective, being polynomial with low degree is analogous to being "low frequency"; the Fourier-analytic counterpart of Bernstein's inequality is closely related to the Sobolev inequality, and is extremely useful in PDE. But I digress.) These facts seem to play a fundamental role in all arguments that yield Ratner-type theorems.

2.12.2 Unipotent actions

Now that we understand unipotent matrices, let us now understand what it means for the action $g : x \mapsto gx$ of a group element $g \in G$ on a homogeneous space G/Γ to be unipotent. By definition, this means that the adjoint action $g : X \mapsto gXg^{-1}$ on the Lie algebra $\mathfrak{g}$ of G is unipotent. By the above discussion, this is the same as saying that the orbit $(g^n X g^{-n})_{n\in\mathbf{Z}}$ always behaves polynomially in n.

This statement can be interpreted via the dynamics on the homogeneous space G/Γ. Consider a point $x \in G/\Gamma$, and look at the orbit $(g^n x)_{n\in\mathbf{Z}}$. Now let us perturb x infinitesimally in the direction of some Lie algebra element X to create a new point $x_\varepsilon := (1+\varepsilon X)x$, where one should think of ε as being infinitesimally small (or alternatively, one can insert errors of $O(\varepsilon^2)$ all over the place). Then the perturbed orbit $g^n x_\varepsilon$ at time n is located at

$$g^n(1+\varepsilon X)x = (1+\varepsilon g^n X g^{-n})g^n x.$$

If g is unipotent, we thus see that the two orbits $(g^n x)_{n\in\mathbf{Z}}$ and $(g^n x_\varepsilon)_{n\in\mathbf{Z}}$ only diverge polynomially in n, without any oscillation. In particular, we have the dichotomy that two orbits either diverge, or are translates of each other, together with Bernstein-like quantitative formulations of this dichotomy. This dichotomy is a crucial component in the proof of Ratner's theorem, and explains why we need the group action to be generated by unipotent elements.

2.12.3 Elliptic, parabolic, and hyperbolic elements of $SL_2(\mathbf{R})$

I have described the distinction between exponential growth/decay, oscillation, and unipotent (polynomial) behaviour. This distinction is particularly easy to visualise geometrically in the context of actions of $SL(2,\mathbf{R})$ on the (affine) plane. Specifically, let us consider an affine linear recurrence sequence

$$x_{n+1} := Ax_n + b; \quad x_0 := x \tag{2.32}$$

where $x \in \mathbf{R}^2$ is an element of the plane, $A \in SL(2,\mathbf{R})$ is a special linear transformation (i.e. a 2×2 matrix of determinant 1), and $b \in \mathbf{R}^2$ is a shift vector. If $A-I$ is

invertible, one can eliminate the shift b by translating the orbit x_n, or more specifically making the substitution

$$y_n := x_n + (A-I)^{-1}b$$

which simplifies (2.32) to

$$y_{n+1} := Ay_n; \quad y_0 := x + (A-I)^{-1}b$$

which allows us to solve for the orbit x_n explicitly as

$$x_n = A^n(x + (A-I)^{-1}b) - (A-I)^{-1}b.$$

Of course, we have to analyse things a little differently in the degenerate case that $A-I$ is not invertible, in particular the lower order term b plays a more significant role in this case. Leaving that case aside for the moment, we see from the above formula that the behaviour of the orbit x_n is going to be largely controlled by the spectrum of A. In this case, A will have two (generalised) eigenvalues $\lambda, 1/\lambda$ whose product is 1 (since $\det(A) = 1$) and whose sum is real (since A clearly has real trace). This gives three possibilities:

1. **Elliptic case.** Here $\lambda = e^{i\theta}$ is a non-trivial unit phase. Then A is similar (after a real linear transformation) to the rotation matrix R_θ described earlier, and so the orbit x_n lies along a linear transform of a circle, i.e. the orbit lies along an ellipse.

2. **Hyperbolic case.** Here λ is real with $|\lambda| > 1$ or $0 < |\lambda| < 1$. In this case A is similar to the diagonal matrix $\begin{pmatrix} \lambda & 0 \\ 0 & 1/\lambda \end{pmatrix}$, and so by previous discussion we see that the orbit x_n lies along a linear transform of a rectangular hyperbola, i.e. the orbit lies along a general hyperbola.

3. **Parabolic case.** This is the boundary case between the elliptic and hyperbolic cases, in which $\lambda = 1$. Then either A is the identity (in which case x_n travels along a line, or is constant), or else (by the Jordan normal form) A is similar to the matrix $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$. Applying a linear change of coordinates, we thus see that the affine recurrence $x_{n+1} = Ax_n + b$ is equivalent to the 2×2 system

$$y_{n+1} = y_n + z_n + c$$
$$z_{n+1} = z_n + d$$

for some real constants c, d and some real sequences y_n, z_n. If c, d are non-zero, we see that z_n varies linearly in n and y_n varies quadratically in n, and so (y_n, z_n) lives on a parabola. Undoing the linear change of coordinates, we thus see in this case that the original orbit x_n also lies along a parabola. (If c or d vanish, the orbit lies instead on a line.)

Thus we see that all elements of $SL(2, \mathbf{R})$ preserve some sort of conic section. The elliptic elements trap their orbits along ellipses, the hyperbolic elements trap their orbits along hyperbolae, and the parabolic elements trap their orbits along parabolae (or

along lines, in some degenerate cases). The elliptic elements thus generate oscillation, the hyperbolic elements generate exponential growth and decay, and the parabolic elements are unipotent and generate polynomial growth. (If one interprets elements of $SL(2,\mathbf{R})$ as area-preserving linear or affine transformations, then elliptic elements are rotations around some origin (and in some coordinate system), hyperbolic elements are compressions along one axis and dilations along another, and parabolic elements are shear transformations and translations.)

Remark 2.42. It is curious that every element of $SL(2,\mathbf{R})$ preserves at least one non-trivial quadratic form; this statement is highly false in higher dimensions (consider for instance what happens to diagonal matrices). I don't have a "natural" explanation of this fact - some sort of fixed point theorem at work, perhaps? I can cobble together a proof using the observations that (a) every matrix in $SL(2,\mathbf{R})$ is similar to its inverse, (b) the space of quadratic forms on $\mathbf{R}^2$ is odd-dimensional, (c) any linear transformation on an odd-dimensional vector space which is similar to its inverse has at least one eigenvalue equal to ± 1, (d) the action of a non-degenerate linear transformation on quadratic forms preserves positive definiteness, and thus cannot have negative eigenvalues, but this argument seems rather ad hoc to me.

One can view the parabolic elements of $SL(2,\mathbf{R})$ as the limit of elliptic or hyperbolic ones in a number of ways. For instance, the matrix $\begin{pmatrix} 1 & 1 \\ \varepsilon & 1 \end{pmatrix}$ is hyperbolic when $\varepsilon > 0$, parabolic when $\varepsilon = 0$, and elliptic when $\varepsilon < 0$. This is related to how the hyperbola, parabola, and ellipse emerge as sections of the light cone. Another way to obtain the parabola a limit is to view that parabola as an infinitely large ellipse (or hyperbola), with centre infinitely far away. For instance, the ellipse of vertical radius R and horizontal radius $\sqrt{R}$ centred at (0,R) is given by the equation $\frac{x^2}{R} + \frac{(y-R)^2}{R^2} = 1$, which can be rearranged as $y = \frac{1}{2}x^2 + \frac{1}{2R}y^2$. In the limit $R \to \infty$, this ellipse becomes the parabola $y = \frac{1}{2}x^2$, and rotations associated with those ellipses can converge to parabolic affine maps of the type described above. A similar construction allows one to view the parabola as a limit of hyperbolae; incidentally, one can use (the Fourier transform of) this limit to show (formally, at least) that the Schrdinger equation emerges as the non-relativistic limit of the Klein-Gordon equation.

2.12.4 The Lorentz group

Every non-degenerate quadratic form Q on d variables comes with its own symmetry group $SO(Q) \leq SL(d,\mathbf{R})$, defined as the group of special linear transformations which preserve Q. (Note that Q determines a translation-invariant pseudo-Riemannian metric, and thus a Haar measure; so any transformation which preserves Q must be volume-preserving and thus have a determinant of ± 1. So the requirement that the linear transformation be special is not terribly onerous.) Equivalently, $SO(Q)$ is the space of special linear transformations which preserve each of the level sets $\{x \in \mathbf{R}^d : Q(x) = \text{const}\}$ (which, by definition, is a quadric surface).

A non-degenerate quadratic form can always be diagonalised (e.g. by applying the Gram-Schmidt orthogonalisation process), and so after a linear change of coordinates

one can express Q as

$$Q(x_1,\ldots,x_d) = x_1^2 + \ldots + x_r^2 - x_{r+1}^2 - \ldots - x_d^2$$

for some $0 \leq r \leq d$. The pair $(r, d-r)$ is the signature of Q, and $SO(Q)$ is isomorphic to the group $SO(r,d-r)$. The signature is an invariant of Q; this is Sylvester's law of inertia.

In the Euclidean (i.e. definite) case $r = d$ (or $r = 0$), the level sets of Q are spheres (in diagonalised form) or ellipsoids (in general), and so the orbits of elements in $SO(Q) \cong SO(d)$ stay trapped on spheres or ellipsoids. Thus their orbits cannot exhibit exponential growth or decay, or polynomial behaviour; they must instead oscillate, much like the elliptic elements of $SL(2,\mathbf{R})$. In particular, $SO(Q)$ does not contain any non-trivial unipotent elements.

In the indefinite case $d = 2$, $r = 1$, the level sets of Q are hyperbolae (as well as the light cone $\{(x_1,x_2) : x_1^2 - x_2^2 = 0\}$, which in two dimensions is just a pair of intersecting lines). It is then geometrically clear that most elements of $SO(Q) \cong SO(1,1)$ are going to be hyperbolic, as their orbits will typically escape to infinity along hyperbolae. (The only exceptions are the identity and the negative identity.) Elements of $SO(1,1)$ are also known as Lorentz boosts. (More generally, $SO(d,1)$ (or $SO(1,d)$) is the structure group for special relativity in $d-1$ space and 1 time dimensions.)

Now we turn to the case of interest, namely $d = 3$ and Q indefinite, thus $r = 1$ or $r = 2$. By changing the sign of Q if necessary we may take $r = 1$, and after diagonalising we can write

$$Q(x_1,x_2,x_3) = x_1^2 + x_2^2 - x_3^2.$$

The level sets of Q are mostly hyperboloids, together with the light cone $\{(x_1,x_2,x_3) : x_1^2 + x_2^2 - x_3^2 = 0\}$. So a typical element of $SO(Q) \cong SO(2,1)$ will have orbits that are trapped inside light cones or on hyperboloids.

In general, these orbits will wander in some complicated fashion over such a cone or hyperboloid. But for some special elements of $SO(Q)$, the orbit is contained in a smaller variety. For instance, consider a Euclidean rotation around the x_3 axis by some angle θ. This clearly preserves Q, and the orbits of this rotation lie on horizontal circles, which are of course each contained in a hyperboloid or light cone. So we see that $SO(Q)$ contains elliptical elements, and this is "because" we can get ellipses as sections of hyperboloids and cones, by slicing them with spacelike planes.

Similarly, if one considers a Lorentz boost in the x_1, x_2 directions, we also preserve Q, and the orbits of this rotation lie on vertical hyperbolae (or on a one-dimensional light cone). So we see that $SO(Q)$ contains hyperbolic elements, which is "because" we can get hyperbolae as sections of hyperbolae and cones, by slicing them with timelike planes.

So, to get unipotent elements of $SO(Q)$, it is clear what we should do: we should exploit the fact that parabolae are also sections of hyperboloids and cones, obtained by slicing these surfaces along null planes. For instance, if we slice the hyperboloid $\{(x_1,x_2,x_3) : x_1^2 + x_2^2 - x_3^2 = 1\}$ with the null plane $\{(x_1,x_2,x_3) : x_3 = x_2 + 1\}$ we obtain the parabola $\{(x_1,x_3-1,x_3) : 2x_3 = x_1^2\}$. A small amount of calculation then lets us find a linear transformation which preserves both the hyperboloid and the null plane (and thus preserves Q and preserves the parabola); indeed, if we introduce null coordinates

$(y_1,y_2,y_3) := (x_1,x_3-x_2,x_3+x_2)$, then the hyperboloid and null plane are given by the equations $y_1^2 = y_2y_3+1$ and $y_2 = 1$ respectively; a little bit of algebra shows that the linear transformations $(y_1,y_2,y_3) \mapsto (y_1+ay_2,y_2,y_3+2ay_1+a^2y_2)$ will preserve both surfaces for any constant a. This provides a one-parameter family (a parabolic subgroup, in fact) of unipotent elements (known as *null rotations*) in $SO(Q)$. By rotating the null plane around we can get many such one-parameter families, whose orbits trace out all sorts of parabolae, and it is not too hard at this point to show that the unipotent elements can in fact be used to generate all of $SO(Q)$ (or $SO(Q)^+$).

Remark 2.43. Incidentally, the fact that the parabola is a section of a cone or hyperboloid of one higher dimension allows one (via the Fourier transform) to embed solutions to the free Schrödinger equation as solutions to the wave or Klein-Gordon equations of one higher dimension; this trick allows one, for instance, to derive the conservation laws of the former from those of the latter. See for instance Exercises 2.11, 3.2, and 3.30 of my book [Ta2006d].

2.12.5 Notes

This article was originally posted on Oct 5, 2007 at

terrytao.wordpress.com/2007/10/05

Thanks to Emmanuel Kowalski and Attila Smith for corrections.

2.13 The Jordan normal form and the Euclidean algorithm

In my recent article (Section 2.12), I used the Jordan normal form for a matrix in order to justify a couple of arguments. As a student, I learned the derivation of this form twice: firstly (as an undergraduate) by using the minimal polynomial, and secondly (as a graduate) by using the structure theorem for finitely generated modules over a principal ideal domain. I found though that the former proof was too concrete and the latter proof too abstract, and so I never really got a good intuition on how the theorem really worked. So I went back and tried to synthesise a proof that I was happy with, by taking the best bits of both arguments that I knew. I ended up with something which wasn't too different from the standard proofs (relying primarily on the (extended) Euclidean algorithm and the fundamental theorem of algebra), but seems to get at the heart of the matter fairly quickly, so I thought I'd put it up on this blog anyway.

Before we begin, though, let us recall what the Jordan normal form theorem is. For this post, I'll take the perspective of abstract linear transformations rather than of concrete matrices. Let $T : V \to V$ be a linear transformation on a finite dimensional complex vector space V, with no preferred coordinate system. We are interested in asking what possible "kinds" of linear transformations V can support (more technically, we want to classify the conjugacy classes of $\mathrm{End}(V)$, the ring of linear endomorphisms of V to itself). Here are some simple examples of linear transformations.

1. **The right shift**. Here, $V = \mathbf{R}^n$ is a standard vector space, and the *right shift* $U : V \to V$ is defined as $U(x_1,\ldots,x_n) = (0,x_1,\ldots,x_{n-1})$, thus all elements are shifted right by one position. (For instance, the 1-dimensional right shift is just the zero operator.)

2. **The right shift plus a constant**. Here we consider an operator $U + \lambda I$, where $U : V \to V$ is a right shift, I is the identity on V, and $\lambda \in \mathbf{C}$ is a complex number.

3. **Direct sums**. Given two linear transformations $T : V \to V$ and $S : W \to W$, we can form their direct sum $T \oplus S : V \oplus W \to V \oplus W$ by the formula $(T \oplus S)(v,w) := (Tv,Sw)$.

Our objective is then to prove the

Theorem 2.44 (Jordan normal form)**.** *Every linear transformation $T : V \to V$ on a finite dimensional complex vector space V is similar to a direct sum of transformations, each of which is a right shift plus a constant.*

(Of course, the same theorem also holds with left shifts instead of right shifts.)

2.13.1 Reduction to the nilpotent case

Recall that a linear transformation $T : V \to V$ is *nilpotent* if we have $T^m = 0$ for some positive integer m. For instance, every right shift operator is nilpotent, as is any direct sum of right shifts. In fact, these are essentially the only nilpotent transformations:

Theorem 2.45 (Nilpotent Jordan normal form)**.** *Every nilpotent linear transformation $T : V \to V$ on a finite dimensional vector space is similar to a direct sum of right shifts.*

We will prove this theorem later, but for now let us see how we can quickly deduce Theorem 2.44 from Theorem 2.45. The idea here is, of course, to split up the minimal polynomial, but it turns out that we don't actually need the minimal polynomial *per se*; any polynomial that annihilates the transformation will do.

More precisely, let $T : V \to V$ be a linear transformation on a finite-dimensional complex vector space V. Then the powers $I, T, T^2, T^3, \ldots$ are all linear transformations on V. On the other hand, the space of all linear transformations on V is a finite-dimensional vector space. Thus there must be a non-trivial linear dependence between these powers. In other words, we have $P(T) = 0$ (or equivalently, $V = \ker(P(T))$) for some polynomial P with complex coefficients.

Now suppose that we can factor this polynomial P into two coprime factors of lower degree, $P = QR$. Using the extended Euclidean algorithm (or more precisely, Bézout's identity), we can find more polynomials A, B such that $AQ + BR = 1$. In particular,

$$A(T)Q(T) + B(T)R(T) = I. \tag{2.33}$$

The formula (2.33) has two important consequences. Firstly, it shows that $\ker(Q(T)) \cap \ker(R(T)) = \{0\}$, since if a vector v was in the kernel of both $Q(T)$ and $R(T)$, then by applying (2.33) to v we obtain $v = 0$. Secondly, it shows that $\ker(Q(T)) + \ker(R(T)) = V$. Indeed, given any $v \in V$, we see from (2.33) that $v = R(T)B(T)v + Q(T)A(T)v$; since $Q(T)R(T) = R(T)Q(T) = P(T) = 0$ on V, we see that $R(T)B(T)v$ and $Q(T)A(T)$ lie in $\ker(Q(T))$ and $\ker(R(T))$ respectively. Finally, since all polynomials in T commute with each other, the spaces $\ker(Q(T))$ and $\ker(R(T))$ are T-invariant.

Putting all this together, we see that the linear transformation T on $\ker(P(T))$ is similar to the direct sum of the restrictions of T to $\ker(Q(T))$ and $\ker(R(T))$ respectively. We can iterate this observation, reducing the degree of the polynomial P which annihilates T, until we reduce to the case in which this polynomial P cannot be split into coprime factors of lesser degree. But by the fundamental theorem of algebra, this can only occur if P takes the form $P(t) = (t - \lambda)^m$ for some $\lambda \in \mathbf{C}$ and $m \geq 0$. In other words, we can reduce to the case when $(T - \lambda I)^m = 0$, or in other words T is equal to λI plus a nilpotent transformation. If we then subtract off the λI term, the claim now easily follows from Theorem 2.45.

Remark 2.46. From a modern algebraic geometry perspective, all we have done here is split the spectrum of T (or of the ring generated by T) into connected components.

It is interesting to see what happens when two eigenvalues get very close together. If one carefully inspects how the Euclidean algorithm works, one concludes that the coefficients of the polynomials $A(T)$ and $B(T)$ above become very large (one is trying to separate two polynomials $Q(T)$ and $R(T)$ that are only barely coprime to each other). Because of this, the Jordan decomposition becomes very unstable when eigenvalues begin to collide.

Because the fundamental theorem of algebra is used, it was necessary[25] to work in an algebraically closed field such as the complex numbers $\mathbf{C}$. Over the reals, one

[25] Indeed, one can in fact *deduce* the fundamental theorem of algebra from the Jordan normal form theorem.

picks up other "elliptic" components, such as 2×2 rotation matrices, which are not decomposable into translates of shift operators.

Thus far, the decompositions have been canonical - the spaces one is decomposing into can be defined uniquely in terms of T (they are the kernels of the primary factors of the minimal polynomial). However, the further splitting of the nilpotent (or shifted nilpotent) operators into smaller components will be non-canonical[26], depending on an arbitrary choice of basis.

2.13.2 Proof of the nilpotent case

To prove the nilpotent Jordan normal form theorem, I would like to take a dynamical perspective, looking at orbits $x, Tx, T^2x, \ldots$ of T. (These orbits will be a cheap substitute for the concept of a *Jordan chain*.) Since T is nilpotent, every such orbit terminates in some finite time m_x, which I will call the *lifespan* of the orbit (i.e. m_x is the least integer such that $T^{m_x}x = 0$). We will call x the *initial point* of the orbit, and $T^{m_x-1}x$ the *final point*.

We claim that the elements of a finite orbit $x, Tx, \ldots, T^{m_x-1}x$ are all linearly independent. This is best illustrated with an example. Suppose x has lifespan 3, thus $T^3x = 0$ and $T^2x \neq 0$. Suppose there linear dependence between x, Tx, T^2x, say $3x + 4Tx + 5T^2x = 0$. Applying T^2 we obtain $3T^2x = 0$, a contradiction. A similar argument works for any other linear dependence or for any other lifespan. (Note how we used the shift T here to eliminate all but the final point of the orbit; we shall use a similar trick with multiple orbits shortly.)

The vector space spanned by a finite orbit $x, Tx, \ldots, T^{m_x-1}x$ is clearly T-invariant, and if we use this finite orbit as a basis for this space then the restriction of T to this space is just the right shift. Thus to prove Theorem 2.45, it suffices to show

Theorem 2.47 (Nilpotent Jordan normal form, again)**.** *Let $T : V \to V$ be nilpotent. Then there exists a basis of V which is the concatenation of finite orbits $x, Tx, \ldots, T^{m_x-1}x$.*

We can prove this by the following argument (basically the same argument used to prove the *Steinitz exchange lemma*, but over the ring $\mathbf{C}[T]$ instead of $\mathbf{C}$). First observe that it is a triviality to obtain a concatenation of finite orbits which *span* V: just take any basis of V and look at all the finite orbits that they generate. Now all we need to do is to keep whittling down this over-determined set of vectors so that they span *and* are linearly independent, thus forming a basis.

Suppose instead that we had some collection of finite orbits $x_i, Tx_i, \ldots, T^{m_{x_i}-1}x_i$, for $i = 1, \ldots, r$ which spanned V, but which contained some non-trivial linear relation. To take a concrete example, suppose we had three orbits x, Tx, y, Ty, T^2y and z, Tz, T^2z, T^3z which had a non-trivial linear relation

$$3x + 4Tx + 5Ty + 6T^2y + 7T^2z = 0.$$

By applying some powers of T if necessary, and stopping just before everything vanishes, we may assume that our non-trivial linear relation only involves the final points

[26] On the other hand, the multiplicities of each type of shift-plus-constant factor will remain canonical; this is easiest to see by inspecting the dimensions of the kernels of $(T - \lambda I)^m$ for various λ, m using a Jordan normal form.

$T^{m_{x_i}-1}x_i$ of our orbits. For instance, in the above example we can apply T once to obtain the non-trivial linear relation

$$3Tx+5T^2y+7T^3z=0.$$

We then factor out as many powers of T as we can; in this case, we have

$$T(3x+5Ty+7T^2z)=0.$$

The expression in parentheses is a linear combination of various elements of our putative basis, in this case x, Ty, and T^2z, with each orbit being represented at most once. At least one of these elements is an initial point (in this case, x). We can then replace that element with the element in parentheses, thus shortening one of the orbits but keeping the span of the concatenated orbits unchanged. (In our example, we replace the orbit x, Tx of lifespan 2 with an orbit $3x+5Ty+7T^2z$ of lifespan 1.) Iterating this procedure until no further linear dependences remain, we obtain Theorem 2.47 and thus Theorem 2.44.

2.13.3 Notes

This article was originally posted on Oct 12, 2007 at

`terrytao.wordpress.com/2007/10/12`

Greg Kuperberg observed that the above proof also yields the classification of finitely generated PID modules in the case of finitely generated modules with non-trivial annihilator. In particular, replacing the polynomial ring $\mathbf{C}[X]$ by modules over $\mathbf{Z}$, one can obtain the classification of fintie abelian groups. Greg also pointed out that the Hahn-Hellinger theorem can be viewed as an infinite-dimensional analogue of the Jordan normal form for self-adjoint operators.

2.14 John's blowup theorem for the nonlinear wave equation

Today I'd like to discuss (part of) a cute and surprising theorem of Fritz John[Jo1979] in the area of non-linear wave equations, and specifically for the equation

$$\partial_{tt} u - \Delta u = |u|^p \tag{2.34}$$

where $u : \mathbf{R} \times \mathbf{R}^3 \to \mathbf{R}$ is a scalar function of one time and three spatial dimensions.

The evolution of this type of non-linear wave equation can be viewed as a "race" between the dispersive tendency of the linear wave equation

$$\partial_{tt} u - \Delta u = 0 \tag{2.35}$$

and the positive feedback tendencies of the nonlinear ODE

$$\partial_{tt} u = |u|^p. \tag{2.36}$$

More precisely, solutions to (2.35) tend to decay in time as $t \to +\infty$, as can be seen from the presence of the $\frac{1}{t}$ term in the explicit formula

$$u(t,x) = \frac{1}{4\pi t} \int_{|y-x|=t} \partial_t u(0,y)\, dS(y) + \partial_t [\frac{1}{4\pi t} \int_{|y-x|=t} u(0,y)\, dS(y)], \tag{2.37}$$

for such solutions in terms of the initial position $u(0,y)$ and initial velocity $\partial_t u(0,y)$, where $t > 0$, $x \in \mathbf{R}^3$, and dS is the area element of the sphere $\{y \in \mathbf{R}^3 : |y-x| = t\}$. (For this post I will ignore the technical issues regarding how smooth the solution has to be in order for the above formula to be valid.) On the other hand, solutions to (2.36) tend to blow up in finite time from data with positive initial position and initial velocity, even if this data is very small, as can be seen by the family of solutions

$$u_T(t,x) := c(T-t)^{-2/(p-1)}$$

for $T > 0$, $0 < t < T$, and $x \in \mathbf{R}^3$, where c is the positive constant $c := (\frac{2(p+1)}{(p-1)^2})^{1/(p-1)}$. For T large, this gives a family of solutions which starts out very small at time zero, but still manages to go to infinity in finite time.

The equation (2.34) can be viewed as a combination of equations (2.35) and (2.36) and should thus inherit a mix of the behaviours of both its "parents". As a general rule, when the initial data $u(0,\cdot), \partial_t u(0,\cdot)$ of solution is small, one expects the dispersion to "win" and send the solution to zero as $t \to \infty$, because the nonlinear effects are weak; conversely, when the initial data is large, one expects the nonlinear effects to "win" and cause blowup, or at least large amounts of instability. This division is particularly pronounced when p is large (since then the nonlinearity is very strong for large data and very weak for small data), but not so much for p small (for instance, when $p = 1$, the equation becomes essentially linear, and one can easily show that blowup does not occur from reasonable data.)

The theorem of John formalises this intuition, with a remarkable threshold value for p:

Theorem 2.48. *Let* $1 < p < \infty$.

1. *If* $p < 1+\sqrt{2}$, *then there exist solutions which are arbitrarily small (both in size and in support) and smooth at time zero, but which blow up in finite time.*

2. *If* $p > 1+\sqrt{2}$, *then for every initial data which is sufficiently small in size and support, and sufficiently smooth, one has a global solution (which goes to zero uniformly as* $t \to \infty$*).*

Remark 2.49. At the critical threshold $p = 1+\sqrt{2}$ one also has blowup from arbitrarily small data, as was shown subsequently by Schaeffer[Sc1985]

The ostensible purpose of this article is to try to explain why the curious exponent $1+\sqrt{2}$ should make an appearance here, by sketching out the proof of part 1 of John's theorem (I will not discuss part 2 here); but another reason I am writing this article is to illustrate how to make quick "back-of-the-envelope" calculations in harmonic analysis and PDE which can obtain the correct numerology for such a problem much faster than a fully rigorous approach. These calculations can be a little tricky to handle properly at first, but with practice they can be done very swiftly.

The first step, which is standard in nonlinear evolution equations, is to rewrite the differential equation (2.34) as an integral equation. Just as the basic ODE

$$\partial_t u = F$$

can be rewritten via the fundamental theorem of calculus in the integral form

$$u(t) = u(0) + \int_0^t F(s)\, ds,$$

it turns out that the inhomogeneous wave equation

$$\partial_{tt} u - \Delta u = F$$

can be rewritten via the fundamental solution (2.37) of the homogeneous equation (together with *Duhamel's principle*) in the integral form

$$u(t,x) = u_{\mathrm{lin}}(t,x) + \frac{1}{4\pi} \int_0^t \int_{|y-x|=|t-s|} \frac{1}{t-s} F(s,y)\, dS(y) ds$$

where u_{lin} is the solution to the homogeneous wave equation (2.35) with initial position $u(0,x)$ and initial velocity $\partial_t u(0,x)$ (and is given using (2.37)). [I plan to write more about this formula in a later article, but today I will just treat it as a miraculous identity. I will note however that the formula generalises Newton's formula $u(x) = \frac{1}{4\pi} \int_{\mathbf{R}^3} \frac{1}{|x-y|} F(y)\, dy$ for the standard solution to Poisson's equation $-\Delta u = F$.]

Using the fundamental solution, the nonlinear wave equation (2.34) can be rewritten in integral form as

$$u(t,x) = u_{\mathrm{lin}}(t,x) + \frac{1}{4\pi} \int_0^t \int_{|y-x|=|t-s|} \frac{1}{t-s} |u(s,y)|^p\, dS(y) ds. \tag{2.38}$$

Remark 2.50. Strictly speaking, one needs to first show that the solution exists and is sufficiently smooth before Duhamel's principle can be rigorously applied, but this turns out to be a routine technical detail and I will not discuss it here.

John's argument now exploits a remarkable feature of the fundamental solution of the three-dimensional wave equation, namely that it is non-negative; combining this with the non-negativity of the forcing term $|u|^p$, we see that the integral in (2.38), that represents the cumulative effect of the nonlinearity, is always non-negative. Thus we have the pointwise inequality

$$u(t,x) \geq u_{\mathrm{lin}}(t,x), \tag{2.39}$$

but also we see that any lower bound for u of the form $u(t,x) \geq v(t,x)$ can be immediately bootstrapped via (2.38) to a new lower bound

$$u(t,x) \geq u_{\mathrm{lin}}(t,x) + \frac{1}{4\pi} \int_0^t \int_{|y-x|=|t-s|} \frac{1}{t-s} |v(s,y)|^p \, dS(y) ds. \tag{2.40}$$

This gives a way to iteratively give lower bounds on a solution u, by starting with the lower bound (2.38) (and computing $u_{\mathrm{lin}}(t,x)$ explicitly using (2.37)) and then feeding this bound repeatedly into (2.40) to see what one gets. (This iteration procedure is closely related to the method of *Picard iteration* for constructing solutions to nonlinear ODE or PDE, which is still widely used today in the modern theory.)

What will transpire is that this iterative process will yield successively larger lower bounds when $p < 1+\sqrt{2}$, but will yield successively smaller lower bounds when $p > 1+\sqrt{2}$; this is the main driving force behind John's theorem. (To actually establish blowup in finite time when $p < 1+\sqrt{2}$, there is an auxiliary step that uses energy inequalities to show that once the solution gets sufficiently large, it will be guaranteed to develop singularities within a finite amount of additional time. To establish global solutions when $p > 1+\sqrt{2}$, one needs to show that the lower bounds constructed by this scheme in fact converge to the actual solution, and establish uniform control on all of these lower bounds.)

The remaining task is a computational one, to evaluate the various lower bounds for u arising from (2.39) and (2.40) from some given initial data. In principle, this is just an application of undergraduate several variable calculus, but if one sets about working out the relevant integrals exactly (using polar coordinates, etc.), the computations quickly become tediously complicated. But we don't actually need exact, closed-form expressions for these integrals; just knowing the order of magnitude of these integrals is enough. For that task, much faster (and looser) computational techniques are available.

Let's see how. We begin with the computation of the linear solution $u_{\mathrm{lin}}(t,x)$. This is given in terms of the initial data $u(0,x), \partial_t u(0,x)$ via the formula (2.37). Now, for the purpose of establishing John's theorem in the form stated above, we have the freedom to pick the initial data as we please, as long as it is smooth, small, and compactly supported. To make our life easier, we pick initial data with vanishing initial position and non-negative initial velocity, thus $u(0,x) = 0$ and $\partial_t u(0,x) \geq 0$; this eliminates the pesky partial derivative in (2.37) and makes u_{lin} non-negative. More concretely, let us take

$$\partial_t u(0,x) := \varepsilon \psi(x/\varepsilon)$$

for some fixed non-negative bump function ψ (the exact form is not relevant) and some small $\varepsilon > 0$, thus the initial velocity has very small amplitude and width. To simplify the notation we shall work with macroscopic values of ε, thus $\varepsilon \sim 1$, but it will be not hard to see that the arguments below also work for very small ε (though of course the smaller ε is, the longer it will take for blowup to occur).

As I said before, we only need an order of magnitude computation. Let us reflect this by describing the initial velocity $\partial_t u(0,x)$ in fuzzier notation:

$$\partial_t u(0,x) \sim 1 \text{ when } x = O(1).$$

Geometrically, $\partial_t u$ has "height" ~ 1 on a ball of radius $O(1)$ centred at the origin. We will retain this sort of fuzzy notation throughout the rest of the argument; it is not fully rigorous, but we can always go back and make the computations formal (and much lengthier) after we have performed the quick informal calculations to show the way ahead.

Thus we see from (2.37) that the linear solution $u_{\text{lin}}(t,x)$ can be expressed somewhat fuzzily in the form

$$u_{\text{lin}}(t,x) \sim \frac{1}{t} \int_{|y-x|=t; y=O(1)} 1 dS(y).$$

Note that the factor $\frac{1}{4\pi}$ can be discarded for the purposes of order of magnitude computation. Geometrically, the integral is measuring the area of the portion of the sphere $\{|y-x| = t\}$ which intersects the ball $\{y = O(1)\}$. A little bit of geometric visualisation will reveal that for large times $t \gg 1$, this portion of the sphere will vanish unless $|x| = t + O(1)$, in which case it is a spherical cap of diameter $O(1)$, and thus area $O(1)$. Thus we are led to the back-of-the-envelope computation

$$u_{\text{lin}}(t,x) \sim \frac{1}{t} \text{ when } |x| = t + O(1) \text{ and } t \gg 1$$

with $u_{\text{lin}}(t,x)$ zero when $|x| \neq t + O(1)$. (This vanishing outside of a neighbourhood of the light cone $\{|x| = t\}$ is a manifestation of the sharp Huygens principle.) In particular, from (2.39) we obtain the initial lower bound

$$u(t,x) \gg \frac{1}{t} \text{ when } |x| = t + O(1) \text{ and } t \gg 1.$$

If we then insert this bound into (2.40) and discard the linear term u_{lin} (which we already know to be positive, and which we have already "used up" in some sense) we obtain the lower bound

$$u(t,x) \gg \int_0^t \int_{|y-x|=|t-s|; |y|=s+O(1); s \gg 1} \frac{1}{t-s} \frac{1}{s^p} \, dS(y) ds.$$

This is a moderately scary looking integral. But we can get a handle on it by first looking at it geometrically. For a fixed point (t,x) in spacetime, the region of integration is the intersection of a backwards light cone $\{(s,y) : 0 \leq s \leq t; |y-x| = |t-s|\}$ with a thickened forwards light cone $\{(s,y) : |y| = s + O(1); s \gg 1\}$. If $|x|$ is much larger than

t, then these cones will not intersect. If $|x|$ is close to t, the intersection looks complicated, so let us consider the spacelike case when $|x|$ is much less than t, say $|x| \leq t/2$; we also continue working in the asymptotic regime $t \gg 1$. In this case, a bit of geometry or algebra shows that the intersection of the two light cones is a two-dimensional ellipsoid in spacetime of radii $\sim t$ (in particular, its surface area is $\sim t^2$), and living at times s in the interior of $[0,t]$, thus s and $t-s$ are both comparable to t. Thickening the forward cone, it is then geometrically intuitive that the intersection of the backwards light cone with the thickened forwards light cone is an angled strip around that ellipse of thickness ~ 1; thus the total measure of this strip is roughly $\sim t^2$. Meanwhile, since s and $t-s$ are both comparable to t, the integrand is of magnitude $\sim \frac{1}{t}\frac{1}{t^p}$. Putting all of this together, we conclude that

$$u(t,x) \gg t^2 \frac{1}{t}\frac{1}{t^p} = t^{1-p}$$

whenever we are in the interior cone region $\{(t,x) : t \gg 1; |x| \leq t/2\}$.

To summarise so far, the linear evolution filled out the light cone $\{(t,x) : t \gg 1; |x| = t + O(1)\}$ with a decay t^{-1}, and then the nonlinearity caused a secondary wave that filled out the interior region $\{(t,x) : t \gg 1; |x| < t/2\}$ with a decay t^{1-p}. We now compute the tertiary wave by inserting the secondary wave bound back into (2.40), to get

$$u(t,x) \gg \int_0^t \int_{|y-x|=|t-s|; |y|<s/2; s\gg 1} \frac{1}{t-s}\frac{1}{t^{p(1-p)}}\, dS(y)ds.$$

Let us continue working in an interior region, say $\{(t,x) : t \gg 1; |x| < t/4\}$. The region of integration is the intersection of the backwards light cone $\{(s,y) : 0 \leq s \leq t; |y-x| = t-s\}$ with an interior region $\{(s,t) : s \gg 1; |y| < s/2\}$. A brief sketch of the situation reveals that this intersection basically consists of the portion of the backwards light cone in which s is comparable in size to t. In particular, this intersection has a three-dimensional measure of $\sim t^3$, and on the bulk of this intersection, s and $t-s$ are both comparable to t. So we obtain a lower bound

$$u(t,x) \gg t^3 \frac{1}{t}\frac{1}{t^{p(1-p)}} = t^{1-p}t^{2-(p-1)^2}$$

whenever $t \gg 1$ and $|x| < t/4$.

Now we finally see where the condition $p < 1+\sqrt{2}$ will come in; if this condition is true, then $2-(p-1)^2$ is positive, and so the tertiary wave is stronger than the secondary wave, and also situated in essentially the same location of spacetime. This is the beginning of a positive feedback loop; the quaternary wave will be even stronger still, and so on and so forth. Indeed, it is not hard to show that if $p < 1+\sqrt{2}$, then for any constant A, one will have a lower bound of the form $u(t,x) \gg t^A$ in the interior of the light cone. This does not quite demonstrate blowup *per se* - merely superpolynomial growth instead - but actually one can amplify this growth into blowup with a little bit more effort (e.g. integrating (2.34) in space to eliminate the Laplacian term and investigating the dynamics of the spatial integral $\int_{\mathbf{R}^3} u(t,x)\, dx$, taking advantage of finite speed of propagation for this equation, which limits the support of u to the cone $\{|x| \leq t + O(1)\}$). A refinement of these arguments, taking into account more of the

components of the various waves in the iteration, also gives blowup for the endpoint $p = 1+\sqrt{2}$.

In the other direction, if $p > 1+\sqrt{2}$, the tertiary wave appears to be smaller than the secondary wave (though to fully check this, one has to compute a number of other components of these waves which we have discarded in the above computations). This sets up a negative feedback loop, with each new wave in the iteration scheme decaying faster than the previous, and thus suggests global existence of the solution, at least when the size of the initial data (which was represented by ε) was sufficiently small. This heuristic prediction can be made rigorous by controlling these iterates in various function space norms that capture these sorts of decay, but I will not detail them here.

Remark 2.51. More generally, any analysis of a semilinear equation that requires one to compute the tertiary wave tends to give conditions on the exponents which are quadratic in nature; if the quaternary wave was involved also, then cubic constraints might be involved, and so forth. In this particular case, an analysis of the primary and secondary waves alone (which would lead just to linear constraints on p) are not enough, because these waves live in very different regions of spacetime and so do not fully capture the feedback mechanism.

2.14.1 Notes

This article was originally posted on Oct 26, 2007 at
terrytao.wordpress.com/2007/10/26

2.15 Hilbert's nullstellensatz

I had occasion recently to look up the proof of *Hilbert's nullstellensatz*, which I haven't studied since cramming for my algebra qualifying exam as a graduate student. I was a little unsatisfied with the proofs I was able to locate - they were fairly abstract and used a certain amount of algebraic machinery, which I was terribly rusty on - so, as an exercise, I tried to find a more computational proof that avoided as much abstract machinery as possible. I found a proof which used only the extended Euclidean algorithm and high school algebra, together with an induction on dimension and the obvious observation that any non-zero polynomial of one variable on an algebraically closed field has at least one non-root. It probably isn't new (in particular, it might be related to the standard model-theoretic proof of the nullstellensatz, with the Euclidean algorithm and high school algebra taking the place of quantifier elimination), but I thought I'd share it here anyway.

Throughout this article, F is going to be a fixed algebraically closed field (e.g. the complex numbers $\mathbf{C}$). I'd like to phrase the nullstellensatz in a fairly concrete fashion, in terms of the problem of solving a set of simultaneous polynomial equations $P_1(x) = \ldots = P_m(x) = 0$ in several variables $x = (x_1, \ldots, x_d) \in F^d$ over F, thus $P_1, \ldots, P_m \in F[x]$ are polynomials in d variables. One obvious obstruction to solvability of this system is if the equations one is trying to solve are *inconsistent* in the sense that they can be used to imply $1 = 0$. In particular, if one can find polynomials $Q_1, \ldots, Q_m \in F[x]$ such that $P_1Q_1 + \ldots + P_mQ_m = 1$, then clearly one cannot solve $P_1(x) = \ldots = P_m(x) = 0$. The *weak nullstellensatz* asserts that this is, in fact, the only obstruction:

Theorem 2.52 (Weak nullstellensatz)**.** *Let $P_1, \ldots, P_m \in F[x]$ be polynomials. Then exactly one of the following statements holds:*

I. The system of equations $P_1(x) = \ldots = P_m(x) = 0$ has a solution $x \in F^d$.

II. There exist polynomials $Q_1, \ldots, Q_m \in F[x]$ such that $P_1Q_1 + \ldots + P_mQ_m = 1$.

Note that the hypothesis that F is algebraically closed is crucial; for instance, if F is the real line $\mathbf{R}$, then the equation $x^2 + 1 = 0$ has no solution, but there is no polynomial $Q(x)$ such that $(x^2 + 1)Q(x) = 1$.

Like many results of the "The only obstructions are the obvious obstructions" type, the power of the nullstellensatz lies in the ability to take a hypothesis about *non-existence* (in this case, non-existence of solutions to $P_1(x) = \ldots = P_m(x) = 0$) and deduce a conclusion about *existence* (in this case, existence of $Q_1, \ldots, Q_m$ such that $P_1Q_1 + \ldots + P_mQ_m = 1$). The ability to get "something from nothing" is clearly going to be both non-trivial and useful. In particular, the nullstellensatz offers an important duality between algebraic geometry (Conclusion I is an assertion that a certain algebraic variety is empty) and commutative algebra (Conclusion II is an assertion that a certain ideal is non-proper).

Now suppose one is trying to solve the more complicated system $P_1(x) = \ldots = P_d(x) = 0; R(x) \neq 0$ for some polynomials $P_1, \ldots, P_d, R$. Again, any identity of the form $P_1Q_1 + \ldots + P_mQ_m = 1$ will be an obstruction to solvability, but now more obstructions are possible: any identity of the form $P_1Q_1 + \ldots + P_mQ_m = R^r$ for some non-negative

integer r will also obstruct solvability. The *strong nullstellensatz* asserts that this is the only obstruction:

Theorem 2.53 (Strong nullstellensatz)**.** *Let $P_1,\ldots,P_m,R \in F[x]$ be polynomials. Then exactly one of the following statements holds:*

I. The system of equations $P_1(x) = \ldots = P_m(x) = 0$, $R(x) \neq 0$ has a solution $x \in F^d$.

II. There exist polynomials $Q_1,\ldots,Q_m \in F[x]$ and a non-negative integer r such that $P_1Q_1 + \ldots + P_mQ_m = R^r$.

Of course, the weak nullstellensatz corresponds to the special case in which $R = 1$. The strong nullstellensatz is usually phrased instead in terms of ideals and radicals, but the above formulation is easily shown to be equivalent to the usual version (modulo Hilbert's basis theorem).

One could consider generalising the nullstellensatz a little further by considering systems of the form $P_1(x) = \ldots = P_m(x) = 0, R_1(x),\ldots,R_n(x) \neq 0$, but this is not a significant generalisation, since all the inequations $R_1(x) \neq 0,\ldots,R_n(x) \neq 0$ can be concatenated into a single inequation $R_1(x)\ldots R_n(x) \neq 0$. The presence of the exponent r in Conclusion II is a little annoying; to get rid of it, one needs to generalise the notion of an algebraic variety to that of a scheme (which is worth doing for several other reasons too, in particular one can now work over much more general objects than just algebraically closed fields), but that is a whole story in itself (and one that I am not really qualified to tell).

2.15.1 The base case $d = 1$

In an earlier draft of this article, I had attempted to prove the weak nullstellensatz by induction on dimension, and then deduced the strong nullstellensatz from the weak via a lifting trick of Zariski (the key observation being that the inequation $x \neq 0$ was equivalent to the solvability of the equation $xy - 1 = 0$ for some y). But I realised that my proof of the weak nullstellensatz was incorrect (it required the strong nullstellensatz in one lower dimension as the inductive hypothesis), and so now I am proceeding by establishing the strong nullstellensatz directly.

We shall induct on the dimension d (i.e. the number of variables in the system of equations).

The case $d = 0$ is trivial, so we use the $d = 1$ case as the base case, thus $P_1,\ldots,P_m,R$ are all polynomials of one variable. This case follows easily from the fundamental theorem of algebra, but it will be important for later purposes to instead use a more *algorithmic* proof in which the coefficients of the polynomials $Q_1,\ldots,Q_m$ required for Conclusion II are obtained from the coefficients of $P_1,\ldots,P_m,R$ in an explicitly computable fashion (using only the operations of addition, subtraction, multiplication, division, and branching on whether a given field element is zero or non-zero). In particular, one does not need to locate roots of polynomials in order to construct $Q_1,\ldots,Q_m$ (although one will of course need to do so to locate a solution x for Conclusion I). [It is likely that one could get these sorts of computability properties on $Q_1,\ldots,Q_m$ "for free" from Galois theory, but I have not attempted to do so.] It is however instructive to

secretly apply the fundamental theorem of algebra throughout the proof which follows, to clarify what is going on.

Let us say that a collection $(P_1,\ldots,P_m;R)$ of polynomials *obeys the nullstellensatz* if at least one of Conclusions I and II is true. It is clear that Conclusions I and II cannot both be true, so to prove the nullstellensatz it suffices to show that every collection $(P_1,\ldots,P_m;R)$ obeys the nullstellensatz.

We can of course throw away any of the P_i that are identically zero, as this does not affect whether $(P_1,\ldots,P_m;R)$ obeys the nullstellensatz. If none of the P_i remain, then we have Conclusion I, because the polynomial R has at most finitely many zeroes, and because an algebraically closed field must be infinite. So suppose that we have some non-zero P_i. We then repeatedly use the extended Euclidean algorithm to locate the greatest common divisor $D(x)$ of the remaining P_i. Note that this algorithm automatically supplies for us some polynomials $Q_1(x),\ldots,Q_m(x)$ such that

$$P_1(x)Q_1(x)+\ldots+P_m(x)Q_m(x)=D(x).$$

Because of this, we see that $(P_1,\ldots,P_m;R)$ obeys the nullstellensatz if and only if $(D;R)$ obeys the nullstellensatz. So we have effectively reduced to the case $m=1$.

Now we apply the extended Euclidean algorithm again, this time to D and R, to express the gcd D' of D and R as a combination $D'=DA+RB$, and also to factor $D=D'S$ and $R=D'T$ for some polynomials A,B,S,T with $AS+BT=1$. A little algebra then shows that one has a solution to the problem

$$D(x)=0;R(x)\neq 0$$

whenever one has a solution to the problem

$$S(x)=0;D'(x)\neq 0.$$

Also, if some power of D' is a multiple of S, then some power of R is a multiple of D. Thus we see that if $(S;D')$ obeys the nullstellensatz, then $(D;R)$ does also. But we see that the net degree of S and D' is less than the net degree of D and R unless R is constant, so by infinite descent we may reduce to that case. If R is zero then we clearly have Conclusion II, so we may assume R is non-zero. If D is constant then we again have Conclusion II, so assume that D is non-constant. But then as the field is algebraically closed, D has at least one root, and so we are in case Conclusion I. This completes the proof of the $d=1$ case.

For the inductive step, it is important to remark that the above proof is *algorithmic* in the sense that a computer which was given the coefficients for $P_1,\ldots,P_m,R$ as inputs could apply a finite number of arithmetic operations (addition, subtraction, multiplication, division), as well as a finite number of branching operations based on whether a given variable was zero or non-zero, in order to output either

1. the coefficients of a non-constant polynomial D with the property that any root x of D would give us Conclusion I;

2. the coefficients of a non-zero polynomial R with the property that any non-root x of R would give us Conclusion I; or

3. the coefficients of polynomials $Q_1,\ldots,Q_m$ which gave Conclusion II for some specific r.

In most cases, the number of branching operations is rather large (see for instance the example of solving two linear equations below). There is however one simple case in which only one branching is involved, namely when $m=2$, $R=1$, and P_1,P_2 are monic. In this case, we have an identity of the form

$$P_1S_1+P_2S_2=\mathrm{Res}(P_1,P_2)$$

where S_1,S_2 are polynomials (whose coefficients are polynomial combinations of the coefficients of P_1 and P_2 and $\mathrm{Res}(P_1,P_2)\in F$ is the *resultant* of P_1 and P_2, which is another polynomial combination of the coefficients of P_1 and P_2. If the resultant is non-zero then we have

$$\frac{1}{\mathrm{Res}(P_1,P_2)}S_1P_1+\frac{1}{\mathrm{Res}(P_1,P_2)}S_2P_2=1$$

and so the system is unsolvable (we have Conclusion II); otherwise, the system is solvable.

2.15.2 The inductive case $d>1$

Now we do the inductive case, when $d\geq 2$ and the claim has already been proven for $d-1$. The basic idea is to view Conclusion I not as a system of equations in d unknowns, but as a $d-1$-dimensional family of systems in one unknown. We will then apply the $d=1$ theory to each system in that family and use the algorithmic nature of that theory to glue everything together properly.

We write the variable $x\in F^d$ as $x=(y,t)$ for $y\in F^{d-1}$ and $t\in F$. The ring $F[x]$ of polynomials in d variables can thus be viewed as a ring $F[y][t]$ of polynomials in one variable t, in which the coefficients lie in the ring $F[y]$.

Let I be the ideal in $F[x]$ generated by $P_1,\ldots,P_m$. We either need to solve the system

$$P_1(y,t)=\ldots=P_m(y,t)=0;R(y,t)\neq 0 \tag{2.41}$$

or show that

$$R^r=0 \bmod I \text{ for some } r. \tag{2.42}$$

We assume that no solution to (2.41) exists, and use this to synthesise a relation of the form (2.42). Let $y\in F^{d-1}$ be arbitrary. We can view the polynomials $P_1(y,t),\ldots,P_m(y,t),R(y,t)$ as polynomials in F[t], whose coefficients lie in F but happen to depend in a polynomial fashion on y. To emphasise this, we write $P_{j,y}(t)$ for $P_j(y,t)$ and $R_y(t)$ for $R(y,t)$. Then by hypothesis, there is no t for which

$$P_{1,y}(t)=\ldots=P_{m,y}(t)=0;\quad R_y(t)\neq 0.$$

To motivate the strategy, let us consider the easy case when $R=1$, $m=2$, and P_1, P_2 are monic polynomials in t. Then by our previous discussion, the above system is solvable for any fixed y precisely when $\mathrm{Res}(P_{1,y},P_{2,y})$ is zero. So either the equation

$\mathrm{Res}(P_{1,y}, P_{2,y}) = 0$ has a solution, in which case we have (2.41), or it does not. But in the latter case, by applying the nullstellensatz at one lower dimension we see that $\mathrm{Res}(P_{1,y}, P_{2,y})$ must be constant in y. But recall that the resultant is a linear combination $P_{1,y}S_{1,y} + P_{2,y}S_{2,y}$ of $P_{1,y}$ and $P_{2,y}$, where the polynomials $S_{1,y}$ and $S_{2,y}$ depend polynomially on $P_{1,y}$ and $P_{2,y}$ and thus on y itself. Thus we end up with (2.42), and the induction closes in this case.

Now we turn to the general case. Applying the $d = 1$ analysis, we conclude that there exist polynomials $Q_{1,y}, \ldots, Q_{m,y} \in F[t]$ of t, and an $r = r_y \geq 0$, such that

$$P_{1,y}(t)Q_{1,y}(t) + \ldots + P_{m,y}(t)Q_{m,y}(t) = R_y^{r_y}(t). \tag{2.43}$$

Now, if the exponent r_y was constant in y, and the coefficients of $Q_{1,y}, \ldots, Q_{m,y}$ depended polynomially on y, we would be in case (2.42) and therefore done.

It is not difficult to make r_y constant in y. Indeed, we observe that the degrees of $P_{1,y}(t), \ldots, P_{m,y}(t)$ are bounded uniformly in y. Inspecting the $d = 1$ analysis, we conclude that the exponent r_y returned by that algorithm is then also bounded uniformly in y. We can always raise the value of r_y by multiplying both sides of (2.43) by R_y, and so we can make $r = r_y$ independent of y, thus

$$P_1(y,t)Q_{1,y}(t) + \ldots + P_m(y,t)Q_{m,y}(t) = R^r(y,t). \tag{2.44}$$

Now we need to work on the Q's. Unfortunately, the coefficients on Q are not polynomial in y; instead, they are *piecewise rational* in y. Indeed, by inspecting the algorithm used to prove the $d = 1$ case, we see that the algorithm makes a finite number of branches, depending on whether certain polynomial expressions $T(y)$ of y are zero or non-zero. At the end of each branching path, the algorithm returns polynomials $Q_{1,y}, \ldots, Q_{m,y}$ whose coefficients were rational combinations of the coefficients of $P_{1,y}, \ldots, P_{m,y}$ and are thus rational functions of x. Furthermore, all the division operations are by polynomials $T(y)$ which were guaranteed to be non-zero by some stage of the branching process, and so the net denominator of any of these coefficients is some product of the $T(y)$ that are guaranteed non-zero.

An example might help illustrate what's going on here. Suppose that $m = 2$ and $R = 1$, and that $P_1(y,t), P_2(y,t)$ are linear in t, thus

$$P_{1,y}(t) = a(y) + tb(y); P_{2,y}(t) = c(y) + td(y)$$

for some polynomials $a, b, c, d \in F[y]$. To find the gcd of $P_{1,y}$ and $P_{2,y}$ for a given y, which determines the solvability of the system $P_{1,y}(t) = P_{2,y}(t) = 0$, the Euclidean algorithm branches as follows:

1. If $b(y)$ is zero, then
   (a) If $a(y)$ is zero, then
      i. If $d(y)$ is non-zero, then $0P_{1,y} + \frac{1}{d(y)}P_{2,y}$ is the gcd (and the system is solvable).
      ii. Otherwise, if $d(y)$ is zero and $c(y)$ is non-zero, then $0P_{1,y} + \frac{1}{c(y)}P_{2,y} = 1$ is the gcd (and the system is unsolvable).

iii. Otherwise, if $d(y)$ and $c(y)$ are both zero, then $0P_{1,y} + 0P_{2,y}$ is the gcd (and the system is solvable).

(b) Otherwise, if $a(y)$ is non-zero, then $\frac{1}{a(y)}P_{1,y} + 0P_{2,y} = 1$ is the gcd (and the system is unsolvable).

2. Otherwise, if $b(y)$ is non-zero, then

(a) If $a(y)d(y) - b(y)c(y)$ is non-zero, then $\frac{d(y)}{a(y)d(y)-b(y)c(y)}P_{1,y} - \frac{b(y)}{a(y)d(y)-b(y)c(y)}P_{2,y} = 1$ is the gcd (and the system is unsolvable).

(b) Otherwise, if $a(y)d(y) - b(y)c(y)$ is zero, then $\frac{1}{b(y)}P_{1,y} + 0P_{2,y}$ is the gcd (and the system is solvable).

So we see that even in the rather simple case of solving two linear equations in one unknown, there is a moderately complicated branching tree involved. Nevertheless, there are only finitely many branching paths. Some of these paths may be *infeasible*, in the sense that there do not exist any $y \in F^{d-1}$ which can follow these paths. But given any feasible path, say one in which the polynomials $S_1(y), \ldots, S_a(y)$ are observed to be zero, and $T_1(y), \ldots, T_b(y)$ are observed to be non-zero, we know (since we are assuming no solution to (2.41)) that the algorithm creates an identity of the form (2.44) in which the coefficients of $Q_{1,y}, \ldots, Q_{m,y}$ are rational polynomials in y, whose denominators are products of $T_1, \ldots, T_b$. We may thus clear denominators (enlarging r if necessary) and obtain an identity of the form

$$P_1(y,t)U_1(y,t) + \ldots + P_m(y,t)U_m(y,t) = (T_1(y)\ldots T_b(y)R(y))^r \tag{2.45}$$

for some polynomials $U_1, \ldots, U_m$. This identity holds whenever y is such that $S_1(y), \ldots, S_a(y)$ are zero and $T_1(y), \ldots, T_b(y)$ are non-zero. But an inspection of the algorithm shows that the only reason we needed $T_1(y), \ldots, T_b(y)$ to be non-zero was in order to divide by these numbers; if we clear denominators throughout, we thus see that we can remove these constraints and deduce that (2.45) holds whenever $S_1(y), \ldots, S_a(y)$ are zero. Further inspection of the algorithm then shows that even if $S_1(y), \ldots, S_a(y)$ are non-zero, this only introduces additional terms to (2.45) which are combinations (over $F[y,t]$) of $S_1, \ldots, S_a$. Thus, for any feasible path, we obtain an identity in $F[y,t]$of the form

$$P_1U_1 + \ldots + P_mU_m = (T_1 \ldots T_bR)^r + S_1V_1 + \ldots + S_aV_a$$

for some polynomials $U_1, \ldots, U_m, V_1, \ldots, V_a \in F[y,t]$. In other words, we see that

$$(T_1 \ldots T_bR)^r = 0 \bmod I, S_1, \ldots, S_a \tag{2.46}$$

for any feasible path.

Now what we need to do is fold up the branching tree and simplify the relations (2.46) until we obtain (2.42). More precisely, we claim that (2.46) holds (for some r) not only for complete feasible paths (in which we follow the branching tree all the way to the end), but for *partial* feasible paths, in which we branch some of the way and then stop in a place where at least one $y \in F^{d-1}$ can solve all the constraints branched on so far. In particular, the empty feasible path will then give (2.42).

To prove this claim, we induct backwards on the length of the partial path. So suppose we have some partial feasible path, which required $S_1(y),\ldots,S_a(y)$ to be zero and $T_1(y),\ldots,T_b(y)$ to be non-zero in order to get here. If this path is complete, then we are already done, so suppose there is a further branching, say on a polynomial $W(y)$. At least one of the cases $W(y)=0$ and $W(y)\neq 0$ must be feasible; and so we now divide into three cases.

Case 1: $W(y)=0$ is feasible and $W(y)\neq 0$ is infeasible. If we follow the $W(y)=0$ path and use the inductive hypothesis, we obtain a constraint

$$(T_1\ldots T_bR)^r = 0 \bmod I, S_1, \ldots, S_a, W \tag{2.47}$$

for some r. On the other hand, since $W(y)\neq 0$ is infeasible, we see that the problem

$$S_1(y) = \ldots = S_a(y) = 0; T_1\ldots T_bW(y) \neq 0$$

has no solution. Since the nullstellensatz is assumed to hold for dimension $d-1$, we conclude that

$$(T_1\ldots T_bW)^{r'} = 0 \bmod S_1,\ldots,S_a.$$

for some r'. If we then multiply (2.47) by $(T_1\ldots T_bR)^{r'}$ to eliminate the role of W, we conclude (2.46) (for $r+r'$) as required.

Case 2: $W(y)=0$ is infeasible and $W(y)\neq 0$ is feasible. If we follow the $W(y)\neq 0$ path, we obtain a constraint

$$(T_1\ldots T_bWR)^{r''} = 0 \bmod I, S_1, \ldots, S_a \tag{2.48}$$

for some r'', while the infeasibility of the $W(y)=0$ path means that there is no solution to

$$S_1(y) = \ldots = S_a(y) = W(y) = 0; T_1\ldots T_b(y) \neq 0$$

and so by the nullstellensatz in dimension $d-1$ we have

$$(T_1\ldots T_b)^{r'''} = WZ \bmod S_1,\ldots,S_a$$

for some polynomial Z and some r'''. If we then multiply (2.48) by $Z^{r''}$ to eliminate W, we obtain (2.46) as desired (for $r''+r'''$).

Case 3: $W(y)=0$ and $W(y)\neq 0$ are both feasible. In this case we obtain the constraints (2.47) and (2.48). We rewrite (2.47) in the form

$$(T_1\ldots T_bR)^r = WZ \bmod S_1,\ldots,S_a$$

for some Z, and then multiply (2.48) by Z^r to eliminate W and obtain (2.46) as desired (for $r+r''$).

This inductively establishes (2.46) for all partial branching paths, leading eventually to (2.42) as desired.

2.15.3 Notes

This article was originally posted on Nov 26, 2007 at

terrytao.wordpress.com/2007/11/26

An anonymous reader pointed out that a simpler version of the above proof was obtained by Arrondo[Ar2006] (and indepedently by Manetti). The main new idea is to first apply a generic linear change of variables to ensure some additional non-degeneracy in the coefficients of the polynomials, which reduces the number of possibilities when one then turns to the induction on dimension.

2.16 The Hahn-Banach theorem, Menger's theorem, and Helly's theorem

In Section 2.15, I discussed how an induction on dimension approach could establish Hilbert's nullstellensatz, which I interpreted as a result describing all the obstructions to solving a system of polynomial equations and inequations over an algebraically closed field. Today, I want to point out that exactly the same approach also gives the Hahn-Banach theorem (at least in finite dimensions), which we interpret as a result describing all the obstructions to solving a system of linear inequalities over the reals (or in other words, a linear programming problem); this formulation of the Hahn-Banach theorem is sometimes known as Farkas' lemma. Then I would like to discuss some standard applications of the Hahn-Banach theorem, such as the separation theorem of Dieudonné, the minimax theorem of von Neumann, Menger's theorem, and Helly's theorem.

To simplify the exposition we shall only work in finite dimensions and with finite complexity objects, such as finite systems of linear inequalities, or convex polytopes with only finitely many sides. The results can be extended to the infinite complexity setting but this requires a bit of care, and can distract from the main ideas, so I am ignoring all of these extensions here.

Let us first phrase a formulation of the Hahn-Banach theorem - namely, Farkas' lemma - which is deliberately chosen to mimic that of the nullstellensatz in Section 2.15. We consider systems of linear inequalities of the form

$$P_1(x), \ldots, P_m(x) \geq 0$$

where $x \in \mathbf{R}^d$ lies in a finite-dimensional real vector space, and $P_1, \ldots, P_m : \mathbf{R}^d \to \mathbf{R}$ are affine-linear functionals. We are interested in the classic linear programming problem of whether such a system admits a solution. One obvious obstruction would be if the above system of inequalities are inconsistent in the sense that they imply $-1 \geq 0$. More precisely, if we can find non-negative reals $q_1, \ldots, q_m \geq 0$ such that $q_1 P_1 + \ldots + q_m P_m = -1$, then the above system is not solvable. Farkas' lemma asserts that this is in fact the only obstruction:

Lemma 2.54 (Farkas' lemma)**.** *Let $P_1, \ldots, P_m : \mathbf{R}^d \to \mathbf{R}$ be affine-linear functionals. Then exactly one of the following statements holds:*

I. The system of inequalities $P_1(x), \ldots, P_m(x) \geq 0$ has a solution $x \in \mathbf{R}^d$.

II. There exist non-negative reals $q_1, \ldots, q_m \geq 0$ such that $q_1 P_1 + \ldots + q_m P_m = -1$.

As in Section 2.15, we prove this by induction on d. The trivial case $d = 0$ could be used as the base case, but again it is instructive to look at the $d = 1$ case first before starting the induction.

If $d = 1$, then each inequality $P_j(x) \geq 0$ can be rescaled into one of three forms: $x - a_j \geq 0$, $b_j - x \geq 0$, or $c_j \geq 0$, where a_j, b_j, or c_j is a real number. The latter inequalities are either trivially true or trivially false, and can be discarded in either case. As for the inequalities of the first and second type, they can be solved so long as all of the a_j which appear here are less than equal to all of the b_j which appear. If this

is not the case, then we have $b_j < a_k$ for some j,k, which allows us to fashion -1 as a non-negative linear combination of $(x-a_k)$ and (b_j-x), and the claim follows.

Now suppose that $d \geq 2$ and the claim has already been proven for $d-1$. As in the previous post, we now split $x = (x',t)$ for $x' \in \mathbf{R}^{d-1}$ and $t \in \mathbf{R}$. Each linear inequality $P_j(x',t) \geq 0$ can now be rescaled into one of three forms: $t - a_j(x') \geq 0$, $b_j(x') - t \geq 0$, and $c_j(x') \geq 0$.

We fix x' and ask what properties x' must obey in order for the above system to be solvable in t. By the one-dimensional analysis, we know that the necessary and sufficient conditions are that $c_j(x') \geq 0$ for all c_j, and that $a_j(x') \leq b_k(x')$ for all a_j and b_k. If we can find an x' obeying these inequalities, then we are in conclusion I and we are done. Otherwise, we apply the induction hypothesis and conclude that we can fashion -1 as a non-negative linear combination of the $c_j(x')$ and of the $b_k(x') - a_j(x')$. But the $b_k(x') - a_j(x')$ can in turn be expressed as a non-negative linear combination of $t - a_j(x')$ and $b_k(x') - t$, and so we are in conclusion II as desired.

Exercise 2.1. Use Farkas' lemma to derive the duality theorem in linear programming.

2.16.1 Applications

Now we connect the above lemma to results which are closer to the Hahn-Banach theorem in its traditional form. We begin with

Theorem 2.55 (Separation theorem)**.** *Let A, B be disjoint convex polytopes in $\mathbf{R}^d$. Then there exists an affine-linear functional $P : \mathbf{R}^d \to \mathbf{R}$ such that $P(x) \geq 1$ for $x \in A$ and $P(x) \leq -1$ for $x \in B$.*

Proof. We can view the system of inequalities $P(x) - 1 \geq 0$ for $x \in A$ and $-1 - P(x) \geq 0$ for $x \in B$ as a system of linear equations on P (or, if you wish, on the coefficients of P). If this system is solvable then we are done, so suppose the system is not solvable. Applying Farkas' lemma, we conclude that there exists $x_1, \ldots, x_m \in A$ and $y_1, \ldots, y_n \in B$ and non-negative constants $q_1, \ldots, q_m, r_1, \ldots, r_n$ such that

$$\sum_{i=1}^{m} q_i (P(x_i) - 1) + \sum_{j=1}^{n} r_j (-1 - P(y_j)) = -1$$

for all P. Setting P to be an arbitrary constant, we conclude

$$\sum_{i=1}^{m} q_i = \sum_{j=1}^{n} r_j = 1/2,$$

and so by the affine nature of P, we can rearrange the above identity as

$$P(\sum_{i=1}^{m} 2q_i x_i) = P(\sum_{j=1}^{n} 2r_i y_j)$$

for all P; since affine functions separate points, we conclude that

$$\sum_{i=1}^{m} 2q_i x_i = \sum_{j=1}^{n} 2r_i y_j.$$

But by convexity the left-hand side is in A and the right-hand side in B, a contradiction. □

The above theorem asserts that any two disjoint convex polytopes can be separated by a hyperplane. One can establish more generally that any two disjoint convex bodies can be separated by a hyperplane; in particular, this implies that if a convex function always exceeds a concave function, then there is an affine linear function separating the two. From this it is a short step to the Hahn-Banach theorem, at least in the setting of finite-dimensional spaces; if one wants to find a linear functional $\lambda : \mathbf{R}^n \to \mathbf{R}$ which has prescribed values on some subspace W, and lies between some convex and concave sets (e.g. $-\|x\| \leq \lambda(x) \leq \|x\|$ for some semi-norm $\|x\|$), then by quotienting out W we can reduce to the previous problem.

We turn now to the minimax theorem. Consider a zero-sum game between two players, Alice and Bob. Alice can pick any one of n strategies using a probability distribution $p = (p_1, \ldots, p_n)$ of her choosing; simultaneously, Bob can pick any one of m strategies using a probability distribution $q = (q_1, \ldots, q_m)$ of his choosing. Alice's expected payoff F then takes the form $F(p,q) := \sum_{i=1}^n \sum_{j=1}^m c_{i,j} p_i q_j$ for some fixed real coefficients $c_{i,j}$; Bob's expected payoff in this zero-sum game is then $-F(p,q)$.

Theorem 2.56 (Minimax theorem)**.** *Given any coefficients $c_{i,j}$, there exists a unique optimal payoff α such that*

I. *(Alice can expect to win at least α) There exists an* optimal strategy *p_* for Alice such that $F(p_*, q) \geq \alpha$ for all q;*

II. *(Bob can expect to lose at most α) There exists an* optimal strategy *q_* for Bob such that $F(p, q_*) \leq \alpha$ for all p.*

Proof. By playing Alice's optimal strategy off against Bob's, we see that the supremum of the set of α which obeys conclusion I is clearly finite, and less than or equal to the infimum of the set of α which obey conclusion II, which is also finite. To finish the proof, it suffices to show that these two numbers are equal. If they were not equal, then we could find an α for which *neither* of conclusions I and II were true.

If conclusion I failed, this means that the system of linear equations

$$p_1, \ldots, p_n \geq 0; p_1 + \ldots + p_n \leq 1; F(p,q) \geq \alpha \text{ for all } q$$

has no solution. From the convexity of F(p,q) in q, we can replace this system with

$$p_1, \ldots, p_n \geq 0; p_1 + \ldots + p_n \leq 1; \sum_{i=1}^n c_{i,j} p_i \geq \alpha \text{ for all } j.$$

Applying Farkas' lemma, we conclude that some non-negative combination of $p_1, \ldots, p_n$, $1 - p_1 - \ldots - p_n$, and $\sum_{i=1}^n c_{i,j} p_i - \alpha$ for $1 \leq j \leq m$ are identically -1. A little bit of algebra shows that there must therefore exist a strategy q for Bob such that -1 can be expressed as a non-negative combination of $p_1, \ldots, p_n$, $1 - p_1 - \ldots - p_n$, and $F(p,q) - \alpha$. In particular, $F(p,q) - \alpha$ must be negative for all strategies p for Alice, and so conclusion II is true, a contradiction. □

The minimax theorem can be used to give game-theoretic proofs of various theorems of Hahn-Banach type. Here is one example:

Theorem 2.57 (Menger's theorem)**.** *Let G be a directed graph, and let v and w be non-adjacent vertices in G. Then the* max-flow *from v to w in G (the largest number of disjoint paths one can find in G from v to w) is equal to the* min-cut *(the least number of vertices (other than v or w) one needs to delete from G to disconnect v from w).*

The proof we give here is definitely not the shortest proof of Menger's theorem, but it does illustrate how game-theoretic techniques can be used to prove combinatorial theorems.

Proof. Consider the following zero-sum game. Bob picks a path from v to w, and Alice picks a vertex (other than v or w). If Bob's path hits Alice's vertex, then Alice wins 1 (and Bob wins -1); otherwise Alice wins 0 (and Bob wins 0 as well). Let α be Alice's optimal payoff. Observe that we can prove $\alpha \leq 1/\text{maxflow}$ by letting Bob picks one of some maximal collection of disjoint paths from v to w at random as his strategy; conversely, we can prove $\alpha \geq 1/\text{mincut}$ by letting Alice pick a vertex from some minimal cut set at random as her strategy. To finish the proof we need to show that in fact $1/\alpha \geq \text{mincut}$ and $1/\alpha \leq \text{maxflow}$.

Let's first show that $1/\alpha \geq \text{mincut}$. Let's assume that Alice is playing an optimal strategies, to get the optimal payoff α for Alice. Then it is not hard to use the optimality to show that every path that Bob might play must be hit by Alice with probability exactly α (and any other path will be hit by Alice with probability at least α), and conversely every vertex that Alice might pick will be hit by Bob with probability α (and any other vertex will be hit by Bob with probability at most α).

Now suppose that two of Bob's paths intersect at some intermediate vertex u. One can show that the two resulting sub-paths from v to u must have an equal chance of being hit by Alice, otherwise by swapping those two sub-paths one can create a path which Alice hits with probability strictly less than α, a contradiction. Similarly, the two sub-paths from u to w must have equal chance of being hit by Alice.

Now consider all the vertices u that Alice can pick for which there exists a path of Bob which hits u before it hits any other vertex of Alice. Let U be the set of such u. Every path from v to w must hit U, because if it is possible to avoid U and instead hit another vertex of Alice, it will again be possible to crate a path that Alice hits with probability strictly less than α, by the above discussion. Thus, U is a cut set. Any given path of Bob hits exactly one vertex in U (again by the above discussion). Since each u in U has a probability α of being hit by Bob, we thus see that this cut set has size exactly $1/\alpha$. Thus $1/\alpha \geq \text{mincut}$ as desired.

Now we show that $\alpha \leq \text{maxflow}$. Define an *$\alpha$-flow* to be a collection of non-negative weights on the directed edges of G such that

1. the net flow at v (the total inflow minus total outflow) is -1, and the net flow at w is $+1$;

2. for any other vertex u, the net flow at u is zero, and total inflow or outflow at u is at most α.

We first observe that at least one α-flow exists. Indeed, if we pick one of Bob's optimal strategies, and weight each edge by the probability that Bob's path passes through that edge, one easily verifies that this gives a α-flow.

Given an α-flow, consider the undirected graph consisting of undirected versions of directed edges on which the weight of the α-flow is positive. If this undirected graph contains a oriented cycle, then one can modify the α-flow on this cycle by an epsilon, increasing the flow weight by ε on edges on the cycle that go with the flow, and reducing them by the same ε on edges that go against the flow; note that this preserves the property of being an α-flow. Increasing ε, we eventually reduce one of the weights to zero, thus reducing the number of edges on which the flow is supported. We can repeat this procedure indefinitely until one arrives at an α-flow whose undirected graph contains no cycles. Now, as v has outflow at least $+1$ and every vertex adjacent to v can have inflow at most α (recall that v is not adjacent to w), the flow must propagate from v to at least $1/\alpha$ other vertices. Each of these vertices must eventually flow to w (which is the only vertex with positive flow) by at least one path; by the above discussion, these paths need to be disjoint. Thus we have $1/\alpha \leq maxflow$ as desired. □

Remark 2.58. The above argument can also be generalised to prove the max-flow min-cut theorem, but we will not do so here.

Exercise 2.2. Use Menger's theorem to prove Hall's marriage theorem.

Now we turn to Helly's theorem. One formulation of this theorem is the following:

Theorem 2.59 (Helly's theorem)**.** *Let $B_1, \ldots, B_m$ be a collection of convex bodies in $\mathbf{R}^d$ with $m > d$. If every $d+1$ of these convex bodies have a point in common, then all m of these convex bodies have a point in common.*

Remark 2.60. The reader is invited to verify Helly's theorem in the $d = 1$ case, to get a flavour as to what is going on.

For simplicity we shall just prove Helly's theorem in the model case when each of the $B_1, \ldots, B_m$ are convex polytopes.

Proof. A convex polytope is the intersection of finitely many half-spaces. From this, one quickly sees that to prove Helly's theorem for convex polytopes, it suffices to do so for half-spaces. By translating things a bit, we may assume that none of the half-spaces go through the origin. Then each half-space can be expressed as the form $\{x : P(x) \geq 1\}$ for some linear functional $P : \mathbf{R}^n \to \mathbf{R}$. Note that by duality, one can view P as living in an d-dimensional vector space $(\mathbf{R}^d)^*$.

Let say that there are m half-spaces involved, and let $P_1, \ldots, P_m$ be the corresponding linear functionals. It suffices to show that if the system $P_1(x), \ldots, P_m(x) \geq 1$ has no solution, then there is some sub-collection $P_{i_1}, \ldots, P_{i_j}$ with $j \leq d+1$ such that $P_{i_1}(x), \ldots, P_{i_j}(x) \geq 1$ also has no solution.

By Farkas' lemma, we know that the system $P_1(x), \ldots, P_m(x) \geq 1$ has no solution if and only if 0 is a convex combination of the $P_1, \ldots, P_m$. So we have everything reduces to establishing

Theorem 2.61 (Dual Helly theorem)**.** *Suppose that* 0 *can be expressed as a convex combination of a collection of vectors* $v_1, \ldots, v_m$ *in a* d*-dimensional vector space. Then* 0 *is also a convex combination of at most* $d+1$ *vectors from that collection.*

To prove this theorem, we use an argument a little reminiscent to that used to prove $1/\alpha \leq$ maxflow in the proof of Menger's theorem. Suppose we can express 0 as a convex combination of some of the v_i. If at most $d+1$ of the vectors have non-zero coefficients attached to them then we are done. Now suppose instead that at least $d+2$ vectors have non-zero coefficients, say $v_1, \ldots, v_{d+2}$ have non-zero coefficients. Then there are at least two linear dependencies among these vectors, which allows us to find coefficients $c_1, \ldots, c_{d+2}$ summing to zero, but not all zero, such that $c_1 v_1 + \ldots + c_{d+2} v_{d+2} = 0$. We can then perturb our preceding convex combination by an ε multiple of this equation to obtain a new representation of 0 as a convex combination of vectors. If we increase ε, we must eventually send one of the coefficients to zero, decreasing the total number of vectors with non-zero coefficients. Iterating this procedure we eventually obtain the dual Helly theorem and hence the original version of Helly's theorem. □

2.16.2 Notes

This article was originally posted on Nov 30, 2007 at

`terrytao.wordpress.com/2007/11/30`

Francois and Mattias Aschenbrenner pointed out that a variant of Farkas lemma, in which positivity is replaced by integrality, was obtained by Kronecker in 1884, and generalised to certain ordered rings in [Sc2006].

2.17 Einstein's derivation of $E = mc^2$

Einstein's equation $E = mc^2$ describing the equivalence of mass and energy is arguably the most famous equation in physics. But his beautifully elegant *derivation* of this formula[Ei1905] from previously understood laws of physics is considerably less famous. (There is an amusing Far Side cartoon in this regard, with the punchline "squared away", which you can find on-line by searching hard enough.)

In this article I would like to present Einstein's original derivation here. Actually, to be precise, in the paper mentioned above, Einstein uses the postulates of special relativity and other known laws of physics to show the following:

Proposition 2.62 (Mass-energy equivalence)**.** *If a body at rest emits a total energy of E while remaining at rest, then the mass of that body decreases by E/c^2.*

Assuming that bodies at rest with zero mass necessarily have zero energy, this implies the famous formula $E = mc^2$ - but only for bodies which are at rest. For moving bodies, there is a similar formula, but one has to first decide what the correct definition of mass is for moving bodies; I will not discuss this issue here, though it can be found in any textbook on relativity.

Broadly speaking, the derivation of the above proposition proceeds via the following five steps:

1. Using the postulates of special relativity, determine how space and time coordinates transform under changes of reference frame (i.e. derive the Lorentz transformations).

2. Using 1., determine how the temporal frequency ν (and wave number k) of photons transform under changes of reference frame (i.e. derive the formulae for relativistic Doppler shift).

3. Using Planck's law $E = h\nu$ (and de Broglie's law $p = \hbar k$) and 2., determine how the energy E (and momentum p) of photons transform under changes of reference frame.

4. Using the law of conservation of energy (and momentum) and 3., determine how the energy (and momentum) of bodies transform under changes of reference frame.

5. Comparing the results of 4. with the classical Newtonian approximations $KE \approx \frac{1}{2}m|v|^2$ (and $p \approx mv$), deduce the relativistic relationship between mass and energy for bodies at rest (and more generally between mass, velocity, energy, and momentum for moving bodies).

Actually, as it turns out, Einstein's analysis for bodies at rest only needs to understand changes of reference frame at *infinitesimally low* velocity, $|v| \ll c$. However, in order to see enough relativistic effects to deduce the mass-energy equivalence, one needs to obtain formulae which are accurate to second order in v (or more precisely, v/c), as opposed to those in Newtonian physics which are accurate to first order in

v (or v/c). Also, to understand the relationship between mass, velocity, energy, and momentum for moving bodies rather than bodies at rest, one needs to consider non-infinitesimal changes of reference frame.

Remark 2.63. Einstein's argument is, of course, a physical argument rather than a mathematical one. While I will use the language and formalism of pure mathematics here, it should be emphasised that I am not exactly giving a formal proof of the above Proposition in the sense of modern mathematics; these arguments are instead more like the classical proofs of Euclid, in that numerous "self evident" assumptions about space, time, velocity, etc. will be made along the way. (Indeed, there is a very strong analogy between Euclidean geometry and the Minkowskian geometry of special relativity.) One can of course make these assumptions more explicit, and this has been done in many other places, but I will avoid doing so here in order not to overly obscure Einstein's original argument.

2.17.1 Lorentz transforms to first order

To simplify the notation, we shall assume that the ambient spacetime S has only one spatial dimension rather than three, although the analysis here works perfectly well in three spatial dimensions (as was done in Einstein's original paper). Thus, in any inertial reference frame F, the spacetime S is parameterised by two real numbers (t,x). Mathematically, we can describe each frame F as a bijection between S and $\mathbf{R} \times \mathbf{R}$. To normalise these coordinates, let us suppose that all reference frames agree to use a single event O in S as their origin $(0,0)$; thus

$$F(O) = (0,0) \tag{2.49}$$

for all frames F.

Given an inertial reference frame $F : S \to \mathbf{R} \times \mathbf{R}$, one can generate new inertial reference frames in two different ways. One is by reflection: one takes the same frame, with the same time coordinate, but reverses the space coordinates to obtain a new frame $\overline{F} : S \to \mathbf{R} \times \mathbf{R}$, thus reversing the orientation of the frame. In equations, we have

$$F(E) = (t,x) \implies \overline{F}(E) = (t,-x) \tag{2.50}$$

for any spacetime event E. Another way is by replacing the observer which is stationary in F with an observer which is moving at a constant velocity v in F, to create a new inertial reference frame $F_v : S \to \mathbf{R} \times \mathbf{R}$ with the same orientation as F. In our analysis, we will only need to understand infinitesimally small velocities v; there will be no need to consider observers traveling at speeds close to the speed of light.

The new frame $F_v : S \to \mathbf{R} \times \mathbf{R}$ and the original frame $F : S \to \mathbf{R} \times \mathbf{R}$ must be related by some transformation law

$$F_v = L_v \circ F \tag{2.51}$$

for some bijection $L_v : \mathbf{R} \times \mathbf{R} \to \mathbf{R} \times \mathbf{R}$. A priori, this bijection L_v could depend on the original frame F as well as on the velocity v, but the principle of relativity implies that L_v is in fact the same in all reference frames F, and so only depends on v.

It is thus of interest to determine what the bijections $L_v : \mathbf{R} \times \mathbf{R} \to \mathbf{R} \times \mathbf{R}$ are. From our normalisation (2.49) we have

$$L_v(0,0) = (0,0) \tag{2.52}$$

but this is of course not enough information to fully specify L_v. To proceed further, we recall *Newton's first law*, which states that an object with no external forces applied to it moves at constant velocity, and thus traverses a straight line in spacetime as measured in any inertial reference frame. (We are assuming here that the property of "having no external forces applied to it" is not affected by changes of inertial reference frame. For non-inertial reference frames, the situation is more complicated due to the appearance of fictitious forces.) This implies that L_v transforms straight lines to straight lines. (To be pedantic, we have only shown this for straight lines corresponding to velocities that are physically attainable, but let us ignore this minor technicality here.) Combining this with (2.52), we conclude that L_v is a linear transformation. (It is a cute exercise to verify this claim formally, under reasonable assumptions such as smoothness of L_v.) Thus we can view L_v now as a 2×2 matrix.

When $v = 0$, it is clear that L_v should be the identity matrix I. Making the plausible assumption that L_v varies smoothly with v, we thus have the Taylor expansion

$$L_v = I + L'_0 v + O(v^2) \tag{2.53}$$

for some matrix L'_0 and for infinitesimally small velocities v. (Mathematically, what we are doing here is analysing the Lie group of transformations L_v via its Lie algebra.) Expanding everything out in coordinates, we obtain

$$L_v(t,x) = ((1 + \alpha v + O(v^2))t + (\beta v + O(v^2))x, (\gamma v + O(v^2))t + (1 + \delta v + O(v^2))x) \tag{2.54}$$

for some absolute constants $\alpha, \beta, \gamma, \delta \in \mathbf{R}$ (not depending on t, x, or v).

The next step, of course, is to pin down what these four constants are. We can use the reflection symmetry (2.50) to eliminate two of these constants. Indeed, if an observer is moving at velocity v in frame F, it is moving in velocity $-v$ in frame $\overline{F}$, and hence $\overline{F_v} = \overline{F}_{-v}$. Combining this with (2.50), (2.51), (2.54) one eventually obtains

$$\alpha = 0 \text{ and } \delta = 0. \tag{2.55}$$

Next, if a particle moves at velocity v in frame F, and more specifically moves along the worldline $\{(t, vt) : t \in \mathbf{R}\}$, then it will be at rest in frame F_v, and (since it passes through the universally agreed upon origin O) must then lie on the worldline $\{(t', 0) : t' \in \mathbf{R}\}$. From (2.51), we conclude

$$L_v(t, vt) \in \{(t', 0) : t' \in \mathbf{R}\} \text{ for all } t. \tag{2.56}$$

Inserting this into (2.54) (and using (2.55)) we conclude that $\gamma = -1$. We have thus pinned down L_v to first order almost completely:

$$L_v(t,x) = (t + \beta vx, x - vt) + O(v^2(|t| + |x|)). \tag{2.57}$$

Thus, rather remarkably, using nothing more than the principle of relativity and Newton's first law, we have almost entirely determined the reference frame transformation laws, save for the question of determining the real number β. [In mathematical terms, what we have done is classify the one-dimensional Lie subalgebras of $\mathfrak{gl}_2(\mathbf{R})$ which are invariant under spatial reflection, and coordinatised using (2.56).] If this number vanished, we would eventually recover classical Galilean relativity. If this number was positive, we would eventually end up with the (rather unphysical) situation of Euclidean relativity, in which spacetime had a geometry isomorphic to that of the Euclidean plane. As it turns out, though, in special relativity this number is negative. This follows from the second postulate of special relativity, which asserts that the speed of light c is the same in all inertial reference frames. In equations (and because F_v has the same orientation as F), this is asserting that

$$L_v(t,ct) \in \{(t',ct') : t' \in \mathbf{R}\} \text{ for all } t \tag{2.58}$$

and

$$L_v(t,-ct) \in \{(t',-ct') : t' \in \mathbf{R}\} \text{ for all } t. \tag{2.59}$$

Inserting either of (2.58), (2.59) into (2.57) we conclude that $\beta = -1/c^2$, and thus we have obtained a full description of L_v to first order:

$$L_v(t,x) = (t - \frac{vx}{c^2}, x - vt) + O(v^2(|t| + |x|)). \tag{2.60}$$

2.17.2 Lorentz transforms to second order

It turns out that to get the mass-energy equivalence, first-order expansion of the Lorentz transformations L_v is not sufficient; we need to expand to second order. From Taylor expansion we know that

$$L_v = I + L_0' v + \frac{1}{2} L_0'' v^2 + O(v^3) \tag{2.61}$$

for some matrix L_0''. To compute this matrix, let us make the plausible assumption that if the frame F_v is moving at velocity v with respect to F, then F is moving at velocity $-v$ with respect to F_v. (One can justify this by considering two frames receding at equal and opposite directions from a single reference frame, and using reflection symmetry to consider how these two frames move with respect to each other.) Applying (2.51) we conclude that $L_{-v} \circ L_v = I$. Inserting this into (2.61) and comparing coefficients we conclude that $L_0'' = (L_0')^2$. Since L_0' is determined from (2.60), we can compute everything explicitly, eventually ending up at the second order expansion

$$L_v(t,x) = (t - \frac{vx}{c^2} + \frac{tv^2}{2c^2}, x - vt + \frac{xv^2}{2c^2}) + O(v^3(|t| + |x|)). \tag{2.62}$$

One can continue in this fashion (exploiting the fact that the L_v must form a Lie group (with the Lie algebra already determined), and using (2.56) to fix the parameterisation $v \mapsto L_v$ of that group) to eventually get the full expansion of L_v, namely

$$L_v(t,x) = \left(\frac{t - vx/c^2}{\sqrt{1 - v^2/c^2}}, \frac{x - vt}{\sqrt{1 - v^2/c^2}} \right),$$

but we will not need to do so here.

2.17.3 Doppler shift

The formula (2.62) is already enough to recover the relativistic Doppler shift formula (to second order in v) for radiation moving at speed c with some wave number k. Mathematically, such radiation moving to the right in an inertial reference frame F can be modeled by the function

$$A\cos(k(x-ct)+\theta)$$

for some amplitude A and phase shift θ. If we move to the coordinates $(t',x') = L_v(t,x)$ provided by an inertial reference frame F', a computation then shows that the function becomes

$$A\cos(k_+(x'-ct')+\theta)$$

where $k_+ = (1-v/c+v^2/2c^2+O(v^3))k$. (actually, if the radiation is tensor-valued, the amplitude A might also transform in some manner, but this transformation will not be of relevance to us.) Similarly, radiation moving at speed c to the left will transform from

$$A\cos(k(x+ct)+\theta)$$

to

$$A\cos(k_-(x+ct)+\theta)$$

where $k_- = (1+v/c+v^2/2c^2+O(v^3))k$. This describes how the wave number k transforms under changes of reference frame by small velocities v. The temporal frequency ν is linearly related to the wave number k by the formula

$$\nu = \frac{c}{2\pi}k, \tag{2.63}$$

and so this frequency transforms by the (red-shift) formula

$$\nu_+ = (1-v/c+v^2/2c^2+O(v^3))\nu \tag{2.64}$$

for right-ward moving radiation and by the (blue-shift) formula

$$\nu_- = (1+v/c+v^2/2c^2+O(v^3))\nu \tag{2.65}$$

for left-ward moving radiation. (As before, one can give an exact formula here, but the above asymptotic will suffice for us.)

2.17.4 Energy and momentum of photons

From the work of Planck, and of Einstein himself on the photoelectric effect, it was known that light could be viewed both as a form of radiation (moving at speed c), and also made up of particles (photons). From Planck's law, each photon has an energy of $E = h\nu$ and (from de Broglie's law) a momentum of $p = \pm\hbar k = \pm\frac{h}{2\pi}k$, where h

is Planck's constant, and the sign depends on whether one is moving rightward or leftward. In particular, from (2.63) we have the pleasant relationship

$$E = |p|c \tag{2.66}$$

for photons. [More generally, it turns out that for arbitrary bodies, momentum, velocity, and energy are related by the formula $p = \frac{1}{c^2}Ev$, though we will not derive this fact here.] Applying (2.64), (2.65), we see that if we view a photon in a new reference frame F_v, then the observed energy E and momentum p now become

$$E_+ = (1 - v/c + v^2/2c^2 + O(v^3))E; \quad p_+ = (1 - v/c + v^2/2c^2 + O(v^3))p \tag{2.67}$$

for right-ward moving photons, and

$$E_- = (1 + v/c + v^2/2c^2 + O(v^3))E; \quad p_- = (1 + v/c + v^2/2c^2 + O(v^3))p \tag{2.68}$$

for left-ward moving photons.

These two formulae (2.67), (2.68) can be unified using (2.66) into a single formula

$$(E'/c^2, p') = L_v(E/c^2, p) + O(v^3) \tag{2.69}$$

for any photon (moving either leftward or rightward) with energy E and momentum p as measured in frame F, and energy E' and momentum p' as measured in frame F_v.

Remark 2.64. Actually, the error term $O(v^3)$ can be deleted entirely by working a little harder. From the linearity of L_v and the conservation of energy and momentum, it is then natural to conclude that (2.69) should also be valid not only for photons, but for any object that can exchange energy and momentum with photons. This can be used to derive the formula $E = mc^2$ fairly quickly, but let us instead give the original argument of Einstein, which is only slightly different.

2.17.5 Einstein's argument

We are now ready to give Einstein's argument. Consider a body at rest in a reference frame F with some mass m and some rest energy E. (We do not yet know that E is equal to mc^2.) Now let us view this same mass in some new reference frame F_v, where v is a small velocity. From Newtonian mechanics, we know that a body of mass m moving at velocity v acquires a kinetic energy of $\frac{1}{2}mv^2$. Thus, assuming that Newtonian physics is valid at low velocities to top order, the net energy E' of this body as viewed in this frame F_v should be

$$E' = E + \frac{1}{2}mv^2 + O(v^3). \tag{2.70}$$

Remark 2.65. If assumes that the transformation law (2.69) applies for this body, one can already deduce the formula $E = mc^2$ for this body at rest from (2.70) (and the assumption that bodies at rest have zero momentum), but let us instead give Einstein's original argument.

We return to frame F, and assume that our body emits two photons of equal energy $\Delta E/2$, one moving left-ward and one moving right-ward. By (2.66) and conservation of momentum, we see that the body remains at rest after this emission. By conservation of energy, the remaining energy in the body is $E - \Delta E$. Let's say that the new mass in the body is $m - \Delta m$. Our task is to show that $\Delta E = \Delta m c^2$.

To do this, we return to frame F_v. By (2.67), the rightward moving photon has energy

$$(1 - v/c + v^2/2c^2 + O(v^3))\frac{\Delta E}{2}; \tag{2.71}$$

in this frame; similarly, the leftward moving photon has energy

$$(1 + v/c + v^2/2c^2 + O(v^3))\frac{\Delta E}{2}. \tag{2.72}$$

What about the body? By repeating the derivation of (2.69), it must have energy

$$(E - \Delta E) + \frac{1}{2}(m - \Delta m)v^2 + O(v^3). \tag{2.73}$$

By the principle of relativity, the law of conservation of energy has to hold in the frame F_v as well as in the frame F. Thus, the energy (2.71) + (2.72) + (2.73) in frame F_v after the emission must equal the energy $E' =$ (2.70) in frame F_v before emission. Adding everything together and comparing coefficients we obtain the desired relationship $\Delta E = \Delta m c^2$.

Remark 2.66. One might quibble that Einstein's argument only applies to emissions of energy that consist of equal and opposite pairs of photons. But one can easily generalise the argument to handle arbitrary photon emissions, especially if one takes advantage of (2.69); for instance, another well-known (and somewhat simpler) variant of the argument works by considering a photon emitted from one side of a box and absorbed on the other. More generally, any other energy emission which could potentially in the future decompose entirely into photons would also be handled by this argument, thanks to conservation of energy. Now, it is possible that other conservation laws prevent decomposition into photons; for instance, the law of conservation of charge prevents an electron (say) from decomposing entirely into photons, thus leaving open the possibility of having to add a linearly charge-dependent correction term to the formula $E = mc^2$. But then one can renormalise away this term by redefining the energy to subtract such a term; note that this does not affect conservation of energy, thanks to conservation of charge.

2.17.6 Notes

This article was originally posted on Dec 28, 2007 at

`terrytao.wordpress.com/2007/12/28`

Laurens Gunnarsen pointed out that Einstein's argument required the use of quantum mechanics to derive the equation $E = mc^2$, but that this equation can also be derived within the framework of classical mechanics by relying more heavily on the representation theory of the Lorentz group.

Thanks to Blake Stacey for corrections.

Chapter 3

Lectures

3.1 Simons Lecture Series: Structure and randomness

On Apr 5-7, 2007, I gave one of the Simons Lecture Series at MIT (the other lecture series was given by David Donoho). I gave three lectures, each expounding on some aspects of the theme "the dichotomy between structure and randomness" (see also my ICM talk [Ta2006], [Ta2006a] on this topic). This theme seems to pervade many of the areas of mathematics that I work in, and my lectures aim to explore how this theme manifests itself in several of these. In the first lecture, I describe the dichotomy as it appears in Fourier analysis and in number theory. In the second, I discuss the dichotomy in ergodic theory and graph theory, while in the third, I discuss PDE.)

3.1.1 Structure and randomness in Fourier analysis and number theory

The "dichotomy between structure and randomness" seems to apply in circumstances in which one is considering a "high-dimensional" class of objects (e.g. sets of integers, functions on a space, dynamical systems, graphs, solutions to PDE, etc.). For sake of concreteness, let us focus today on sets of integers (later lectures will focus on other classes of objects). There are many different types of objects in these classes, however one can broadly classify them into three categories:

- *Structured* objects - objects with a high degree of predictability and algebraic structure. A typical example are the odd integers $A := \{\ldots,-3,-1,1,3,5,\ldots\}$. Note that if some large number n is known to lie in A, this reveals a lot of information about whether $n+1$, $n+2$, etc. will also lie in A. Structured objects are best studied using the tools of *algebra* and *geometry*.

- *Pseudorandom* objects - the opposite of structured; these are highly unpredictable and totally lack any algebraic structure. A good example is a randomly chosen set B of integers, in which each element n lies in B with an independent probability of $1/2$. (One can imagine flipping a coin for each integer n, and defining B to be the set of n for which the coin flip resulted in heads.) Note that if some integer n is known to lie in B, this conveys no information whatsoever about the relationship of $n+1$, $n+2$, etc. with respect to B. Pseudorandom objects are best studied using the tools of analysis and probability.

- *Hybrid* sets - sets which exhibit some features of structure and some features of pseudorandomness. A good example is the primes $P := 2,3,5,7,\ldots$. The primes have some obvious structure in them: for instance, the prime numbers are all positive, they are all odd (with one exception), they are all adjacent to a multiple of six (with two exceptions), and their last digit is always 1, 3, 7, or 9 (with two exceptions). On the other hand, there is evidence that the primes, despite being a deterministic set, behave in a very "pseudorandom" or "uniformly distributed" manner. For instance, from the prime number theorem in arithmetic progressions we know that the last digits of large prime numbers are uniformly distributed in the set $\{1,3,7,9\}$; thus, if N is a large integer, the number of primes less than N ending in (say) 3, divided by the total number of primes less than N, is known

to converge to $1/4$ in the limit as N goes to infinity. In order to study hybrid objects, one needs a large variety of tools: one needs tools such as algebra and geometry to understand the structured component, one needs tools such as analysis and probability to understand the pseudorandom component, and one needs tools such as decompositions, algorithms, and evolution equations to separate the structure from the pseudorandomness.

A recurring question in many areas of analysis is the following: given a specific object (such as the prime numbers), can one determine precisely what the structured components are within the object, and how pseudorandom the remaining components of the object are? One reason for asking this question is that it often helps one compute various *statistics* (averages, sums, integrals, correlations, norms, etc.) of the object being studied. For instance, one can ask for how many twin pairs $\{n, n+2\}$, with n between 1 and N, one can find within a given set. In the structured set A given above, the answer is roughly $N/2$. For the random set B given above, the answer is roughly $N/4$; thus one sees that while A and B have exactly the same density (namely, $1/2$), their statistics are rather different due to the fact that one is structured and one is random. As for the prime numbers, nobody knows for certain what the answer is (although the Hardy-Littlewood prime tuples conjecture [HaLi1923] predicts the answer to be roughly $1.32\frac{N}{\log^2 N}$), because we do not know enough yet about the pseudorandomness of the primes. On the other hand, the parity structure of the prime numbers is enough to show that the number of *adjacent* pairs $\{n, n+1\}$ in the primes is exactly one: $\{2,3\}$.

The problem of determining exactly what the structured and pseudorandom components are of any given object is still largely intractable. However, what we have learnt in many cases is that we can at least show that an arbitrary object can be *decomposed* into some structured component and some pseudorandom component. Also there is often an *orthogonality* property (or *dichotomy*): if an object is orthogonal (or has small correlation with) all structured objects, then it is necessarily pseudorandom, and vice versa. Finally, we are sometimes lucky enough to be able to *classify* all the structured objects which are relevant for any given problem (e.g. computing a particular statistic). In such cases, one merely needs (in principle) to compute how the given object correlates with each member in one's list of structured objects in order to determine what the desired statistic is. This is often simpler (though still non-trivial) than computing the statistic directly.

To illustrate these general principles, let us focus now on a specific area in analytic number theory, namely that of finding additive patterns in the prime numbers $\{2,3,5,7,\ldots\}$. Despite centuries of progress on these problems, many questions are still unsolved, for instance:

- (Twin prime conjecture) There are infinitely many positive integers n such that $n, n+2$ are both prime.
- (Sophie Germain prime conjecture) There are infinitely many positive integers n such that $n, 2n+1$ are both prime.
- (Even Goldbach conjecture) For every even number $N \geq 4$, there is a natural number n such that $n, N-n$ are both prime.

On the other hand, we do have some positive results:

- (Vinogradov's theorem)[Vi1937] For every sufficiently large odd number N, there are positive integers n, n' such that $n, n', N-n-n'$ are all prime. (The best explicit bound currently known for "sufficiently large" is $N \geq 10^{1346}$ [LiWa2002]; the result has also been verified¡ for $7 \leq N \leq 10^{20}$ [Sa1998].

- (van der Corput's theorem)[vdC1939] There are infinitely many positive integers n, n' such that $n, n+n', n+2n'$ are all prime.

- (Green-Tao theorem)[GrTa2008] For any positive integer k, there are infinitely many positive integers n, n' such that $n, n+n', \ldots, n+(k-1)n'$ are all prime.

- (A polynomial generalisation) For any integer-valued polynomials $P_1(n), \ldots, P_k(n)$ with $P_1(0) = \ldots = P_k(0) = 0$, there are infinitely many positive integers n, n' such that $n+P_1(n'), \ldots, n+P_k(n')$ are all prime.

As a general rule, it appears that it is feasible (after non-trivial effort) to find patterns in the primes involving two or more degrees of freedom (as described by the parameters n, n' in above examples), but we still do not have the proper technology for finding patterns in the primes involving only one degree of freedom n. (This is of course an oversimplification; for instance, the pattern n, $n+2$, n', $n'+2$ has two degrees of freedom, but finding infinitely many of these patterns in the primes is equivalent to the twin prime conjecture, and thus presumably beyond current technology. If however one makes a non-degeneracy assumption, one can make the above claim more precise; see [GrTa2008b].)

One useful tool for establishing some (but not all) of the above positive results is Fourier analysis (which in this context is also known as the *Hardy-Littlewood circle method*). Rather than give the textbook presentation of that method here, let us try to motivate why Fourier analysis is an essential feature of many of these problems from the perspective of the dichotomy between structure and randomness, and in particular viewing structure as an obstruction to computing statistics which needs to be understood before the statistic can be accurately computed.

To treat many of the above questions concerning the primes in a unified manner, let us consider the following general setting. We consider k affine-linear forms $\psi_1, \ldots, \psi_k : \mathbf{Z}^r \to \mathbf{Z}$ on r integer unknowns, and ask

Question 3.1. *Does there there exist infinitely many r-tuples $\vec{n} = (n_1, \ldots, n_r) \in \mathbf{Z}_+^r$ of positive integers such that $\psi_1(\vec{n}), \ldots, \psi_k(\vec{n})$ are simulatenously prime?*

For instance, the twin prime conjecture is the case when $k = 2$, $r = 1$, $\psi_1(n) = n$, and $\psi_2(n) = n+2$; van der Corput's theorem is the case when $k = 3$, $r = 2$, and $\psi_j(n, n') = n + (j-1)n'$ for $j = 0, 1, 2$; and so forth.

Because of the "obvious" structures in the primes, the answer to the above question can be "no". For instance, since all but one of the primes are odd, we know that there are not infinitely many patterns of the form $n, n+1$ in the primes, because it is not possible for $n, n+1$ to both be odd. More generally, given any prime q, we know that all but one of the primes is coprime to q. Hence, if it is not possible for

$\psi_1(\vec{n}),\ldots,\psi_k(\vec{n})$ to all be coprime to q, the answer to the above question is basically no (modulo some technicalities which I wish to gloss over) and we say that there is an *obstruction at q*. For instance, the pattern $n, n+1$ has an obstruction at 2. The pattern $n, n+2, n+4$ has no obstruction at 2, but has an obstruction at 3, because it is not possible for $n, n+2, n+4$ to all be coprime to 3. And so forth.

Another obstruction comes from the trivial observation that the primes are all positive. Hence, if it is not possible for $\psi_1(\vec{n}),\ldots,\psi_k(\vec{n})$ to all be positive for infinitely many values of $\vec{n}$, then we say that there is an *obstruction at infinity*, and the answer to the question is again "no" in this case. For instance, for any fixed N, the pattern $n, N-n$ can only occur finitely often in the primes, because there are only finitely many n for which $n, N-n$ are both positive.

It is conjectured that these "local" obstructions are the only *obstructions* to solvability of the above question. More precisely, we have

Conjecture 3.2. *(Dickson's conjecture)[Di1904] If there are no obstructions at any prime q, and there are no obstructions at infinity, then the answer to the above question is "yes".*

This conjecture would imply the twin prime and Sophie Germain conjectures, as well as the Green-Tao theorem; it also implies the Hardy-Littlewood prime tuples conjecture[HaLi1923] as a special case. There is a quantitative version of this conjecture which predicts a more precise count as to how many solutions there are in a given range, and which would then also imply Vinogradov's theorem, as well as Goldbach's conjecture (for sufficiently large N); see [GrTa2008b] for further discussion. As one can imagine, this conjecture is still largely unsolved, however there are many important special cases that have now been established - several of which were achieved via the Hardy-Littlewood circle method.

One can view Dickson's conjecture as an *impossibility statement*: that it is impossible to find any other obstructions to solvability for linear patterns in the primes than the obvious local obstructions at primes q and at infinity. (It is also a good example of a *local-to-global principle*, that local solvability implies global solvability.) Impossibility statements have always been very difficult to prove - one has to locate all possible obstructions to solvability, and eliminate each one of them in turn. In particular, one has to exclude various exotic "conspiracies" between the primes to behave in an unusually structured manner that somehow manages to always avoid all the patterns that one is seeking within the primes. How can one disprove a conspiracy?

To give an example of what such a "conspiracy" might look like, consider the twin prime conjecture, that of finding infinitely many pairs $n, n+2$ which are both prime. This pattern encounters no obstructions at primes q or at infinity and so Dickson's conjecture predicts that there should be infinitely many such patterns. In particular, there are no obstructions at 3 because prime numbers can equal 1 or 2 mod 3, and it is possible to find pairs n, $n+2$ which also have this property. But suppose that it transpired that all but finitely many of the primes ended up being 2 mod 3. From looking at tables of primes this seems to be unlikely, but it is not immediately obvious how to disprove it; it could well be that once one reaches, say, 10^{100}, there are no more primes equal to 1 mod 3. If this unlikely "conspiracy" in the primes was true, then there would be only finitely many twin primes. Fortunately, we have *Dirichlet's*

theorem, which guarantees infinitely many primes equal to a mod q whenever a, q are coprime, and so we can rule out this particular type of conspiracy. (This does strongly suggest, though, that knowledge of Dirichlet's theorem is a necessary but not sufficient condition in order to solve the twin prime conjecture.) But perhaps there are other conspiracies that one needs to rule out also?

To look for other conspiracies that one needs to eliminate, let us rewrite the conspiracy "all but finitely many of the primes are 2 mod 3" in the more convoluted format

$$0.6 < \{\frac{1}{3}p\} < 0.7 \text{ for all but finitely many primes } p$$

where $\{x\}$ is the fractional part of x. This type of conspiracy can now be generalised; for instance consider the statement

$$0 < \{\sqrt{2}p\} < 0.01 \text{ for all but finitely many primes } p \tag{3.1}$$

Again, such a conspiracy seems very unlikely - one would expect these fractional parts to be uniformly distributed between 0 and 1, rather than concentrate all in the interval $[0,0.01]$ - but it is hard to rule this conspiracy out *a priori*. And if this conspiracy (3.1) was in fact true, then the twin prime conjecture would be false, as can be quickly seen by considering the identity

$$\{\sqrt{2}(n+2)\} - \{\sqrt{2}n\} = 2\sqrt{2} \bmod 1,$$

which forbids the two fractional parts on the left-hand side to simultaneously fall in the interval $[0,0.01]$. Thus, in order to solve the twin prime conjecture, one must rule out (3.1). Fortunately, it has been known since the work of Vinogradov[Vi1937] that $\{\sqrt{2}p\}$ is in fact uniformly distributed in the interval $[0,1]$, and more generally that $\{\alpha p\}$ is uniformly distributed in [0,1] whenever α is irrational. Indeed, by Weyl's famous equidistribution theorem (see e.g. [KuNe1974]), this uniform distribution, this is equivalent to the exponential sum estimate

$$\sum_{p<N} e^{2\pi i\alpha p} = o(\sum_{p<N} 1),$$

and we now see the appearance of Fourier analysis in this subject.

One can rather easily concoct an endless stream of further conspiracies, each of which could contradict the twin prime conjecture; this is one of the reasons why this conjecture is considered so difficult. Let us thus leave this conjecture for now and consider some two-parameter problems. Consider for instance the problem of finding infinitely many patterns of the form n, $n+n'$, $n+2n'+2$ (i.e. arithmetic progressions of length 3, but with the last element shifted by 2). Once again, the conspiracy (3.1), if true, would obstruct solvability for this pattern, due to the easily verified identity

$$\{\sqrt{2}n\} - 2\{\sqrt{2}(n+n')\} + \{\sqrt{2}(n+2n'+2)\} = 2\sqrt{2} \bmod 1$$

which is related to the fact that the function $\sqrt{2}n$ has a vanishing second derivative. (Note however that the same conspiracy does not obstruct solvability of an unmodified arithmetic progression n, $n+n'$, $n+2n'$. This highlights a special property of arithmetic

progressions, which most other patterns do not have, namely that arithmetic progressions tend to exist both in structured objects and in pseudorandom objects (and also in hybrids of the two). This is why results about arithmetic progressions have tended to be easier to establish than those about more general patterns, as one does not need to know as much about the structured and random components of the set in which one is looking for progressions.)

More generally, we can see that if the primes correlate in some unusual way with a linear character $e^{2\pi i\alpha p}$, then this is likely to bias or distort the number of patterns $\{n, n+n', n+2n'+2\}$ in a significant manner. However, thanks to Fourier analysis, we can show that these "Fourier conspiracies" are in fact the *only* obstructions to counting this type of pattern. Very roughly, one can sketch the reason for this as follows. Firstly, it is helpful to create a counting function for the primes, namely the *von Mangoldt function* $\Lambda(n)$, defined as $\log p$ whenever n is a power of a prime p, and 0 otherwise. This rather strange-looking function is actually rather natural, because of the identity

$$\sum_{d|n} \Lambda(d) = \log n$$

for all positive integers n, where the sum is over all positive integers d which divide n; this identity is a restatement of the fundamental theorem of arithmetic, and in fact defines the von Mangoldt function uniquely. The problem of counting patterns $\{n, n+n', n+2n'+2\}$ is then roughly equivalent to the task of computing sums such as

$$\sum_n \sum_{n'} \Lambda(n)\Lambda(n+n')\Lambda(n+2n'+2) \tag{3.2}$$

where we shall be intentionally vague as to what range the variables n, n' will be summed over. We have the *Fourier inversion formula*

$$\Lambda(n) = \int_0^1 e^{2\pi i n\theta} \hat{\Lambda}(\theta) d\theta$$

where

$$\hat{\Lambda}(\theta) := \sum_n \Lambda(n) e^{-2\pi i n\theta}$$

is a sum very similar in nature to the sums $\sum_{p<N} e^{2\pi i p\alpha}$ mentioned earlier. Substituting this formula into (3.2), we essentially get an expression of the form

$$\int_0^1 \hat{\Lambda}(\theta)^2 \hat{\Lambda}(-2\theta) e^{4\pi i\theta}\, d\theta$$

(again ignoring issues related to the ranges that n, n' are being summed over). Thus, if one gets good enough control on the Fourier coefficients $\hat{\Lambda}(\theta)$, which can be viewed as a measure of how much the primes "conspire" with a linear phase oscillation with frequency θ, then one can (in principle, at least) count the solutions to the pattern $\{n, n+n', n+2n'+2\}$ in the primes. This is the Hardy-Littlewood circle method in a nutshell, and this is for instance how van der Corput's theorem and Vinogradov theorem were first proven.

I have glossed over the question of how one actually *computes* the Fourier coefficients $\hat{\Lambda}(\theta)$. It turns out that there are two cases. In the "major arc" case when θ is rational, or close to rational (with a reasonably small denominator), then the problem turns out to be essentially equivalent to counting primes in arithmetic progressions, and so one uses tools related to Dirichlet's theorem (i.e. L-functions, the Siegel-Walfisz theorem [Wa1936], etc.). In the "minor arc" case when θ is far from rational, one instead uses identities such as

$$\Lambda(n) = \sum_{d|n} \mu(d) \log \frac{n}{d},$$

where μ is the *Möbius function* (i.e. $\mu(n) := (-1)^k$ when n is the product of k distinct prime factors for some $k \geq 0$, and $\mu(n) = 0$ otherwise), to split the Fourier coefficient as

$$\hat{\Lambda}(\theta) = \sum_d \sum_m \mu(d) \log(m) e^{2\pi i \alpha dm}$$

and then one uses the irrationality of α to exhibit some significant oscillation in the phase $e^{2\pi i\alpha dm}$, which cannot be fully canceled out by the oscillation in the $\mu(d)$ factor. (In practice, the above strategy does not work directly, and one has to work with various truncated or smoothed out versions of the above identities; this is technical and will not be discussed here.)

Now suppose we look at progressions of length 4: n, $n+n'$, $n+2n'$, $n+3n'$. As with progressions of length 3, "linear" or "Fourier" conspiracies such as (3.1) will bias or distort the total count of such progressions in the primes less than a given number N. But, in contrast to the length 3 case where these are the only conspiracies that actually influence things, for length 4 progressions there are now "quadratic" conspiracies which can cause trouble. Consider for instance the conspiracy

$$0 < \{\sqrt{2}p^2\} < 0.01 \text{ for all but finitely many primes } p. \tag{3.3}$$

This conspiracy, which can exist even when all linear conspiracies are eliminated, will significantly bias the number of progressions of length 4, due to the identity

$$\{\sqrt{2}n^2\} - 3\{\sqrt{2}(n+n')^2\} + 3\{\sqrt{2}(n+2n')^2\} - \{\sqrt{2}(n+3n')^2\} = 0 \bmod 1$$

which is related to the fact that the function $\sqrt{2}n^2$ has a vanishing third derivative. In this case, the conspiracy works in one's favour, increasing the total number of progressions of length 4 beyond what one would have naively expected; as mentioned before, this is related to a remarkable "indestructability" property of progressions, which can be used to establish things like the Green-Tao theorem without having to deal directly with these obstructions. Thus, in order to count progressions of length 4 in the primes accurately (and not just to establish the qualitative result that there are infinitely many of them), one needs to eliminate conspiracies such as (3.3), which necessitates understanding exponential sums such as $\sum_{p<N} e^{2\pi i\alpha p^2}$ for various rational or irrational numbers α. What's worse, there are several further "generalised quadratic" conspiracies which can also bias this count, for instance the conspiracy

$$0 < \{\lfloor\sqrt{2}p\rfloor\sqrt{3}p\} < 0.01 \text{ for all but finitely many primes } p,$$

where $x \mapsto \lfloor x \rfloor$ is the greatest integer function. The point here is that the function $\lfloor \sqrt{2}x \rfloor \sqrt{3}x$ has a third divided difference which does not entirely vanish (as with the genuine quadratic $\sqrt{2}x^2$), but does vanish a significant portion of the time (because the greatest integer function obeys the linearity property $\lfloor x+y \rfloor = \lfloor x \rfloor + \lfloor y \rfloor$ a significant fraction of the time), which does lead ultimately to a non-trivial bias effect. Because of this, one is also faced with estimating exponential sums such as $\sum_{p<N} e^{2\pi i \lfloor \sqrt{2}p \rfloor \sqrt{3}p}$. It turns out that the correct way to phrase all of these obstructions is via the machinery of *2-step nilsequences*: details can be found in [GrTa2008b, GrTa2008c, GrTa2008d]. As a consequence, we can in fact give a precise count as to how many arithmetic progressions of primes of length 4 with all primes less than N; it turns out to be

$$(\frac{3}{4}\prod_{p\geq 5}(1-\frac{3p-1}{(p-1)^3})+o(1))\frac{N^2}{\log^4 N} \approx 0.4764\frac{N^2}{\log^4 N}.$$

The method also works for other linear patterns of comparable "complexity" to progressions of length 4. We are currently working on the problem of longer progressions, in which cubic and higher order obstructions appear (which should be modeled by 3-step and higher nilsequences); some work related to this should appear here shortly.

3.1.2 Structure and randomness in ergodic theory and graph theory

In this second lecture, I wish to talk about the dichotomy between structure and randomness as it manifests itself in four closely related areas of mathematics:

- *Combinatorial number theory*, which seeks to find patterns in unstructured dense sets (or colourings) of integers;
- *Ergodic theory* (or more specifically, multiple recurrence theory), which seeks to find patterns in positive-measure sets under the action of a discrete dynamical system on probability spaces (or more specifically, measure-preserving actions of the integers $\mathbf{Z}$);
- *Graph theory*, or more specifically the portion of this theory concerned with finding patterns in large unstructured dense graphs; and
- *Ergodic graph theory*, which is a very new and undeveloped subject, which roughly speaking seems to be concerned with the patterns within a measure-preserving action of the infinite permutation group S_∞, which is one of several models we have available to study infinite "limits" of graphs.

The two "discrete" (or "finitary", or "quantitative") fields of combinatorial number theory and graph theory happen to be related to each other, basically by using the Cayley graph construction; I will give an example of this shortly. The two "continuous" (or "infinitary", or "qualitative") fields of ergodic theory and ergodic graph theory are at present only related on the level of analogy and informal intuition, but hopefully some more systematic connections between them will appear soon.

On the other hand, we have some very rigorous connections between combinatorial number theory and ergodic theory, and also (more recently) between graph theory and ergodic graph theory, basically by the procedure of viewing the infinitary continuous setting as a limit of the finitary discrete setting. These two connections go by the names of the *Furstenberg correspondence principle* and the *graph correspondence principle* respectively. These principles allow one to tap the power of the infinitary world (for instance, the ability to take limits and perform completions or closures of objects) in order to establish results in the finitary world, or at least to take the *intuition* gained in the infinitary world and transfer it to a finitary setting. Conversely, the finitary world provides an excellent model setting to refine one's understanding of infinitary objects, for instance by establishing quantitative analogues of "soft" results obtained in an infinitary manner. I will remark here that this best-of-both-worlds approach, borrowing from both the finitary and infinitary traditions of mathematics, was absolutely necessary for Ben Green and I in order to establish our result on long arithmetic progressions in the primes. In particular, the infinitary setting is excellent for being able to rigorously define and study concepts (such as structure or randomness) which are much "fuzzier" and harder to pin down exactly in the finitary world.

Let me first discuss the connection between combinatorial number theory and graph theory. We can illustrate this connection with two classical results from the former and latter field respectively:

- **Schur's theorem**[Sc1916]: If the positive integers are coloured using finitely many colours, then one can find positive integers x, y such that x, y, $x+y$ all have the same colour.

- **Ramsey's theorem**[Ra1930]: If an infinite complete graph is edge-coloured using finitely many colours, then one can find a triangle all of whose edges have the same colour.

(In fact, both of these theorems can be generalised to say much stronger statements, but we will content ourselves with just these special cases). It is in fact easy to see that Schur's theorem is deducible from Ramsey's theorem. Indeed, given a colouring of the positive integers, one can create an infinite coloured complete graph (the *Cayley graph* associated to that colouring) whose vertex set is the integers $\mathbf{Z}$, and such that an edge $\{a,b\}$ with $a < b$ is coloured using the colour assigned to $b-a$. Applying Ramsey's theorem, together with the elementary identity $(c-a) = (b-a)+(c-b)$, we then quickly deduce Schur's theorem.

Let us now turn to ergodic theory. The basic object of study here is a *measure-preserving system* (or *probability-preserving system*), which is a probability space $(X, \mathscr{B}, \mu)$ (i.e. a set X equipped with a sigma-algebra $\mathscr{B}$ of measurable sets and a probability measure μ on that sigma-algebra), together with a shift map $T : X \to X$, which for simplicity we shall take to be invertible and bi-measurable (so its inverse is also measurable); in particular we have iterated shift maps $T^n : X \to X$ for any integer n, giving rise to an action of the integers $\mathbf{Z}$. The important property we need is that the shift map is *measure-preserving*, thus $\mu(T(E)) = \mu(E)$ for all measurable sets E.

In the previous lecture we saw that sets of integers could be divided (rather informally) into structured sets, pseudorandom sets, and hybrids between the two. The same

is true in ergodic theory - and this time, one can in fact make these notions extremely precise. Let us first start with some examples:

- The *circle shift*, in which $X := \mathbf{R}/\mathbf{Z}$ is the standard unit circle with normalised Haar measure, and $T(x) := x+\alpha$ for some fixed real number α. If we identify X with the unit circle in the complex plane via the standard identification $x \mapsto e^{2\pi i x}$, then the shift corresponds to an anti-clockwise rotation by α. This is a very structured system, and corresponds in combinatorial number theory to *Bohr sets* such as $\{n \in \mathbf{Z} : 0 < \{n\alpha\} < 0.01\}$, which implicitly made an appearance in the previous lecture.

- The *two-point shift*, in which $X := \{0,1\}$ with uniform probability measure, and T simply interchanges 0 and 1. This very structured system corresponds to the set A of odd numbers (or of even numbers) mentioned in the previous lecture. More generally, any permutation on a finite set gives rise to a simple measure-preserving system.

- The *skew shift*, in which $X := (\mathbf{R}/\mathbf{Z})^2$ is the 2-torus with normalised Haar measure, and $T(x,y) := (x+\alpha, y+x)$ for some fixed real number α. If we just look at the behaviour of the x-component of this torus we see that the skew shift contains the circle shift as a *factor*, or equivalently that the skew shift is an *extension* of the circle shift (in this particular case, since the fibres are circles and the action on the fibres is rotation, we call this a *circle extension* of the circle shift). This system is also structured (but in a more complicated way than the previous two shifts), and corresponds to quadratically structured sets such as the *quadratic Bohr set* $\{n \in \mathbf{Z} : 0 < \{\sqrt{2}n^2\} < 0.01\}$, which made an appearance in the previous lecture.

- The *Bernoulli shift*, in which $X := \{0,1\}^{\mathbf{Z}} \equiv 2^{\mathbf{Z}}$ is the space of infinite $0-1$ sequences (or equivalently, the space of all sets of integers), equipped with uniform product probability measure, and T is the left shift $T(x_n)_{n\in\mathbb{Z}} := (x_{n+1})_{n\in\mathbb{Z}}$. This is a very random system, corresponding to the random sets B discussed in the previous lecture.

- Hybrid systems, e.g. products of a circle shift and a Bernoulli shift, or extensions of a circle shift by a Bernoulli system, a doubly skew shift (a circle extension of a circle extension of a circle shift), etc.

One can classify these systems in precise terms according to how the shift action T^n moves sets E around. On the one hand, we have some well-defined notions which represent structure:

- *Trivial* systems are such that $T^n E = E$ for all E and all n.

- *Periodic* systems are such that for every E, there exists a positive n such that $T^n E = E$. The two-point shift is an example, as is the circle shift when α is rational.

- *Almost periodic* or *compact* systems are such that for every E and every $\varepsilon > 0$, there exists a positive n such that $T^n E$ and E differ by a set of measure at most ε. The circle shift is a good example of this (thanks to Weyl's equidistribution theorem). The term "compact" is used because there is an equivalent characterisation of compact systems, namely that the orbits of the shift in $L^2(X)$ are always precompact in the strong topology.

On the other hand, we have some well-defined terms which represent pseudorandomness:

- *Strongly mixing* systems are such that for every E, F, we have $\mu(T^n E \cap F) \to \mu(E)\mu(F)$ as n tends to infinity; the Bernoulli shift is a good example. Informally, this is saying that shifted sets become asymptotically independent of unshifted sets.

- *Weakly mixing* systems are such that for every E, F, we have $\mu(T^n E \cap F) \to \mu(E)\mu(F)$ as n tends to infinity after excluding a set of exceptional values of n of asymptotic density zero. For technical reasons, weak mixing is a better notion to use in the structure-randomness dichotomy than strong mixing (for much the same reason that one always wants to allow negligible sets of measure zero in measure theory).

There are also more complicated (but well-defined) hybrid notions of structure and randomness which we will not give here. We will however briefly discuss the situation for the skew shift. This shift is not almost periodic: most sets A will become increasingly "skewed" as it gets shifted, and will never return to resemble itself again. However, if one restricts attention to the underlying circle shift factor (i.e. restricting attention only to those sets which are unions of vertical fibres), then one recovers almost periodicity. Furthermore, the skew shift is almost periodic *relative* to the underlying circle shift, in the sense that while the shifts $T^n A$ of a given set A do not return to resemble A globally, they do return to resemble A when restricted to any fixed vertical fibre (this can be shown using the method of *Weyl sums* from Fourier analysis and analytic number theory). Because of this, we say that the skew shift is a *compact extension* of a compact system.

As discussed in the above examples, every dynamical system is capable of generating some interesting sets of integers, specifically *recurrence sets* $\{n \in \mathbf{Z} : T^n x_0 \in E\}$ where E is a set in X and x_0 is a point in X. This set actally captures much of the dynamics of E in the system (especially if X is ergodic and x_0 is *generic*). The *Furstenberg correspondence principle* reverses this procedure, starting with a set of integers A and using that to generate a dynamical system which "models" that set in a certain way. Modulo some minor technicalities, it works as follows.

1. As with the Bernoulli shift, we work in the space $X := \{0,1\}^{\mathbf{Z}} \equiv 2^{\mathbf{Z}}$, with the product sigma-algebra and the left shift; but we leave the probability measure μ (which can be interpreted as the distribution of a certain random subset of the integers) undefined for now. The original set A can now be interpreted as a single point inside X.

2. Now pick a large number N, and shift A backwards and forwards up to N times, giving rise to $2N+1$ sets $T^{-N}A,\ldots,T^{N}A$, which can be thought of as $2N+1$ points inside X. We consider the uniform distribution on these points, i.e. we shift A by a random amount between $-N$ and N. This gives rise to a discrete probability measure μ_N on X (which is only supported on $2N+1$ points inside X). Each of these measures is approximately invariant under the shift T.

3. We now let N go to infinity. We apply the (sequential form of the) Banach-Alaoglu theorem, which among other things shows that the space of Borel probability measures on a compact Hausdorff space (which X is) is sequentially compact in the weak-* topology. (This particular version of Banach-Alaoglu can in fact be established by a diagonalisation argument which completely avoids the axiom of choice.) Thus we can find a subsequence of the measures μ_N which converge in the weak-* topology to a limit μ (this subsequence and limit may not be unique, but this will not concern us). Since the μ_N are approximately invariant under T, with the degree of approximation improving with N, one can easily show that the limit measure μ is shift-invariant.

By using this recipe to construct a measure-preserving system from a set of integers, it is possible to deduce theorems in combinatorial number theory from those in ergodic theory (similarly to how the Cayley graph construction allowed one to deduce theorems in combinatorial number theory from those in graph theory). The most famous example of this concerns the following two deep theorems:

- **Szemerédi's theorem**[Sz1975]: If A is a set of integers of positive upper density, and k is a positive integer, then A contains infinitely many arithmetic progressions $x,x+n,\ldots,x+(k-1)n$ of length k. (Note that the case $k=2$ is trivial.)

- **Furstenberg's recurrence theorem**[Fu1977]: If E is a set of positive measure in a measure-preserving system, and k is a positive integer, then there are infinitely many integers n for which $\mu(A\cap T^{n}A\cap\ldots\cap T^{(k-1)n}A)>0$. (Note that the case $k=2$ is the more classical *Poincaré recurrence theorem*).

Using the above correspondence principle (or a slight variation thereof), it is not difficult to show that the two theorems are in fact equivalent; see for instance Furstenberg's book[Fu1981]. The power of these two theorems derives from the fact that the former works for *arbitrary* sets of positive density, and the latter works for *arbitrary* measure-preserving systems - there are essentially no structural assumptions on the basic object of study in either, and it is therefore quite remarkable that one can still conclude such a non-trivial result.

The story of Szemerédi's theorem is quite a long one, which I have discussed in many other places [TaVu2006], [Ta2006e], [Ta2007b], [Ta2007c], though I will note here that all the proofs of this theorem exploit the dichotomy between structure and randomness (and there are some good reasons for this - the underlying cause of arithmetic progressions is totally different in the structured and pseudorandom cases). I will however briefly describe how Furstenberg's recurrence theorem is proven (following the approach of Furstenberg, Katznelson, and Ornstein[FuKaOr1982]; there are a couple other ergodic theoretic proofs, including of course Furstenberg's original proof).

The first major step is to establish the *Furstenberg structure theorem*, which takes an arbitrary measure-preserving system and describes it as a suitable hybrid of a compact system and a weakly mixing system (or more precisely, a weakly mixing extension of a transfinite tower of compact extensions). This theorem relies on Zorn's lemma, although it is possible to give a proof of the recurrence theorem without recourse to the axiom of choice. The proof requires various tools from infinitary analysis (e.g. the compactness of integral operators) but is relatively straightforward. Next, one makes the rather simple observation that the Furstenberg recurrence theorem is easy to show both for compact systems and for weakly mixing systems. In the former case, the almost periodicity shows that there are lots of integers n for which $T^n A$ is almost identical with A (in the sense that they differ by a set of small measure) - which, after shifting by n again, implies that $T^{2n} A$ is almost identical with $T^n A$, and so forth - which soon makes it easy to arrange matters so that $A \cap T^n A \cap \ldots \cap T^{(k-1)n} A$ is non-empty. In the latter case, the weak mixing shows that for most n, the sets (or "events") A and $T^n A$ are almost uncorrelated (or "independent"); similarly, for any fixed m, we have $A \cap T^m A$ and $T^n(A \cap T^m A) = T^n A \cap T^{n+m} A$ almost uncorrelated for n large enough. By using the Cauchy-Schwarz inequality (in the form of a useful lemma of van der Corput) repeatedly, we can eventually show that $A, T^n A, \ldots, T^{(k-1)n} A$ are almost *jointly* independent (as opposed to being merely almost pairwise independent) for many n, at which point the recurrence theorem is easy to show. It is somewhat more tricky to show that one can also combine these arguments with each other to show that the recurrence property also holds for the transfinite combinations of compact and weakly mixing systems that come out of the Furstenberg structure theorem, but it can be done with a certain amount of effort, and this concludes the proof of the recurrence theorem. This same method of proof turns out, with several additional technical twists, to establish many further varieties of recurrence theorems, which in turn (via the correspondence principle) gives several powerful results in combinatorial number theory, several of which continue to have no non-ergodic proof even today.

(There has also been a significant amount of progress more recently by several ergodic theorists [CoLe1988], [FuWe1996], [HoKr2005], [Zi2007] in understanding the "structured" side of the Furstenberg structure theorem, in which dynamical notions of structure, such as compactness, have been converted into algebraic and topological notions of structure, in particular into the actions of nilpotent Lie groups on their homogeneous spaces. This is an important development, and is closely related to the polynomial and generalised polynomial sequences appearing in the previous talk, but it would be beyond the scope of this talk to discuss it here.)

Let us now leave ergodic theory and return to graph theory. Given the power of the Furstenberg correspondence principle, it is natural to look for something similar in graph theory, which would connect up results in finitary graph theory with some infinitary variant. A typical candidate for a finitary graph theory result that one would hope to do this for is the triangle removal lemma, which was discussed in Section 1.6. That lemma is in fact closely connected with Szemerédi's theorem, indeed it implies the $k = 3$ case of that theorem (i.e. Roth's theorem[Ro1953]) in much the same way that Ramsey's theorem implies Schur's theorem. It does turn out that it is possible to obtain such a correspondence, although the infinitary analogues of things like the triangle removal lemma are a little strange-looking (see e.g. [Ta2007e] or [LoSz2008]). But

it is easier to proceed by instead working with the concept of a *graph limit*. There are several equivalent formulations of this limit, including the notion of a "graphon" introduced by Lovász and Szegedy[LoSz2006], the flag algebra construction introduced by Razborov[Ra2008c], and the notion of a permutation-invariant measure space introduced by myself[Ta2007e]. I will discuss my own construction here, which is closely modelled on the Furstenberg correspondence principle. What it does is starts with a sequence G_n of graphs (which one should think of as getting increasingly large, while remaining dense) and extracts a limit object, which is a probability space $(X, \mathscr{B}, \mu)$ together with an action of the permutation group S_∞ on the integers, as follows.

1. We let $X = 2^{\binom{\mathbf{Z}}{2}}$ be the space of all graphs on the integers, with the standard product (i.e. weak) topology, and hence product sigma-algebra. This space has an obvious action of the permutation group S_∞, formed by permuting the vertices.

2. Each graph G_n generates a random graph on the integers - or equivalently, a probability measure μ_n in X - as follows. We randomly and independently sample the vertices of the graph G_n infinitely often, creating a sequence $(v_{n,i})_{i \in \mathbf{Z}}$ of vertices in the graph G_n. (Of course, many of these vertices will collide, but this will be not be important for us.) This then creates a random graph on the integers, with any two integers i and j connected by an edge if their associated vertices $v_{n,i}, v_{n,j}$ are distinct and are connected by an edge in G_n. By construction, the probability measure μ_n associated to this graph is already S_∞-invariant.

3. We then let n go to infinity, and extract a weak limit μ just as with the Furstenberg correspondence principle.

It is then possible to prove results somewhat analogous to the Furstenberg structure theorem and Furstenberg recurrence theorem in this setting, and use this to prove several results in graph theory (as well as its more complicated generalisation, hypergraph theory). I myself am optimistic that by transferring more ideas from traditional ergodic theory into this new setting of "ergodic graph theory", that one could obtain a new tool for systematically establishing a number of other qualitative results in graph theory, particularly those which are traditionally reliant on the Szemerédi regularity lemma[Sz1978] (which is almost a qualitative result itself, given how poor the bounds are). This is however still a work in progress.

3.1.3 Structure and randomness in PDE

In this third lecture, I will talk about how the dichotomy between structure and randomness pervades the study of two different types of partial differential equations (PDE):

- *Parabolic PDE*, such as the heat equation $u_t = \Delta u$, which turn out to play an important role in the modern study of geometric topology; and

- *Hamiltonian PDE*, such as the Schrödinger equation $u_t = i\Delta u$, which are heuristically related Liouville's theorem to measure-preserving actions of the real line (or *time axis*) $\mathbf{R}$, somewhat in analogy to how combinatorial number theory and graph theory were related to measure-preserving actions of $\mathbf{Z}$ and S_∞ respectively, as discussed in the previous lecture.

(In physics, one would also insert some physical constants, such as Planck's constant $\hbar$, but for the discussion here it is convenient to normalise away all of these constants.)

Observe that the form of the heat equation and Schrödinger equation differ only by a constant factor of i (cf. "Wick rotation"). This makes the *algebraic* structure of the heat and Schrödinger equations very similar (for instance, their fundamental solutions also only differ by a couple factors of i), but the *analytic* behaviour of the two equations turns out to be very different. For instance, in the category of Schwartz functions, the heat equation can be continued forward in time indefinitely, but not backwards in time; in contrast, the Schrödinger equation is time reversible and can be continued indefinitely in both directions. Furthermore, as we shall shortly discuss, parabolic equations tend to *dissipate* or destroy the pseudorandom components of a state, leaving only the structured components, whereas Hamiltonian equations instead tend to *disperse* or radiate away the pseudorandom components from the structured components, without destroying them.

Let us now discuss parabolic PDE in more detail. We begin with a simple example, namely how the heat equation can be used to solve the Dirichlet problem, of constructing a harmonic function $\Delta u_\infty = 0$ in a nice domain Ω with some prescribed boundary data. As this is only an informal discussion I will not write down the precise regularity and boundedness hypotheses needed on the domain or data. The harmonic function will play the role here of the "structured" or "geometric" object. From calculus of variations we know that a smooth function $u_\infty : \Omega \to \mathbf{R}$ is harmonic with the specified boundary data if and only if it minimises the *Dirichlet energy* $E(u) := \frac{1}{2}\int_\Omega |\nabla u|^2$, which is a convex functional on u, with the prescribed boundary data. One way to locate the harmonic minimum u_∞ is to start with an arbitrary smooth initial function $u_0 : \Omega \to \mathbf{R}$, and then perform gradient flow $\partial_t u = -\frac{\delta E}{\delta u}(u) = \Delta u$ on this functional, i.e. solve the heat equation with initial data $u(0) = u_0$. One can then show (e.g. by spectral theory of the Laplacian) that regardless of what (smooth) data u_0 one starts with, the solution $u(t)$ to the heat equation exists for all positive time, and converges to the (unique) harmonic function u_∞ on Ω with the prescribed boundary data in the limit $t \to \infty$. Thus we see that the heat flow removes the "random" component $u_0 - u_\infty$ of the initial data over time, leaving only the "structured" component u_∞.

There are many other settings in geometric topology in which one wants to locate a geometrically structured object (e.g. a harmonic map, a constant-curvature manifold, a minimal surface, etc.) within a certain class (e.g. a homotopy class) by minimising an energy-like functional. In some cases one can achieve this by brute force, creating a minimising sequence and then extracting a limiting object by some sort of compactness argument (as is for instance done in the Sacks-Uhlenbeck theory[SaUh1981] of minimal 2-spheres), but then one often has little control over the resulting structured object that one obtains in this manner. By using a parabolic flow (as for instance done in the work of Eells-Sampson[EeSa1964] to obtain harmonic maps in a given homotopy class via harmonic map heat flow) one can often obtain much better estimates and other control on the limit object, especially if certain curvatures in the underlying geometry have a favourable sign.

The most famous recent example of the use of parabolic flows to establish geometric structure from topological objects is, of course, Perelman's use of the Ricci flow applied to compact 3-manifolds with arbitrary Riemannian metrics, in order to establish

the Poincaré conjecture (for the special case of simply connected manifolds) and more generally the geometrisation conjecture (for arbitrary manifolds). Perelman's work showed that Ricci flow, when applied to an arbitrary manifold, will eventually create either extremely geometrically structured, symmetric manifolds (e.g. spheres, hyperbolic spaces, etc.), or singularities which are themselves very geometrically structured (and in particular, their asymptotic behaviour is extremely rigid and can be classified completely). By removing all of the geometric structures that are generated by the flow (via surgery, if necessary) and continuing the flow indefinitely, one can eventually remove all the "pseudorandom" elements of the initial geometry and describe the original manifold in terms of a short list of very special geometric manifolds, precisely as predicted by the geometrisation conjecture. It should be noted that Hamilton[Ha1982] had earlier carried out exactly this program assuming some additional curvature hypotheses on the initial geometry; also, when Ricci flow is instead applied to two-dimensional manifolds (surfaces) rather than three, Hamilton [Ha1988] observed that Ricci flow extracts a constant-curvature metric as its structured component of the original metric, giving an independent proof of the uniformisation theorem (see [ChLuTi2006] for full details).

Let us now leave parabolic PDE and geometric topology and now discuss Hamiltonian PDE, specifically those of (nonlinear) Schrödinger type. (Other classes of Hamiltonian PDE, such as nonlinear wave or Airy type equations, also exhibit similar features, but we will restrict attention to Schrödinger for sake of concreteness.) These equations formally resemble Hamiltonian ODE, which can be viewed as finite-dimensional measure-preserving dynamical systems with a continuous time parameter $t \in \mathbf{R}$. However, this resemblance is not rigorous because Hamiltonian PDE have infinitely many degrees of freedom rather than finitely many; at a technical level, this means that the dynamics takes place on a highly non-compact space (e.g. the energy surface), whereas much of the theory of finite-dimensional dynamics implicitly relies on at least local compactness of the domain. Nevertheless, in many dispersive settings (e.g. when the spatial domain is Euclidean) it seems that almost all of the infinitely many degrees of freedom are so "pseudorandom" or "radiative" as to have an essentially trivial (or more precisely, linear and free) impact on the dynamics, leaving only a mysterious "core" of essentially finite-dimensional (or more precisely, compact) dynamics which is still very poorly understood at present.

To illustrate these rather vague assertions, let us first begin with the free linear Schrödinger equation $iu_t + \Delta u = 0$, where $u : \mathbf{R} \times \mathbf{R}^n \to \mathbf{C}$ has some specified initial data $u(0) = u_0 : \mathbf{R}^n \to \mathbf{C}$, which for simplicity we shall place in the Schwartz class. It is not hard to show, using Fourier analysis, that a unique smooth solution, well-behaved at spatial infinity, to this equation exists, and furthermore obeys the $L^2(\mathbf{R}^n)$ conservation law

$$\int_{\mathbf{R}^n} |u(t,x)|^2 dx = \int_{\mathbf{R}^n} |u_0(x)|^2 \, dx, \tag{3.4}$$

which can be interpreted physically as the law of conservation of probability. By using the fundamental solution for this equation, one can also obtain the pointwise decay estimate

$$\lim_{t \to \infty} |u(t,x)| = 0 \text{ for all } x \in \mathbf{R}^n$$

and in a similar spirit, we have the local decay estimate

$$\lim_{t\to\infty} \int_K |u(t,x)|^2 dx = 0 \tag{3.5}$$

for all compact $K \subset \mathbf{R}^n$.

The two properties (3.4), (3.5) may appear contradictory at first, but what they imply is that the solution is *dispersing* or *radiating* its (fixed amount of) L^2 mass into larger and larger regions of space, so that the amount of mass that any given compact set captures will go to zero as time goes to infinity. This type of dispersion - asymptotic orthogonality to any fixed object - should be compared with the notion of strong mixing discussed in the previous lecture. The analogous notion of weak mixing, by the way, is the slightly weaker statement

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T \int_K |u(t,x)|^2 dx dt = 0 \tag{3.6}$$

for all compact $K \subset \mathbf{R}^n$.

[There is also a very useful and interesting quantitative version of this analysis, known as *profile decomposition*, in which a solution (or sequence of solutions) to the free linear Schrödinger equation can be split into a small number of "structured" components which are localised in spacetime and in frequency, plus a "pseudorandom" term which is dispersed in spacetime, and is small in various useful norms. These decompositions have recently begun to play a major role in this subject, but it would be too much of a digression to discuss them here. See however my lecture notes [Ta2006f] for more discussion.]

To summarise so far, for the free linear Schrödinger equation all solutions are radiative or "pseudorandom". Now let us generalise a little bit by throwing in a (time-independent) potential function $V : \mathbf{R}^n \to \mathbf{R}$, which for simplicity we shall also place in the Schwartz class, leading to the familiar linear Schrödinger equation $i\partial_t u + \Delta u = Vu$. This equation still has unique smooth global solutions, decaying at spatial infinity, for Schwartz data u_0, and still obeys the L^2 conservation law (3.4). What about the dispersion properties (3.5) or (3.6)? Here there is a potential obstruction to dispersion (or pseudorandomness), namely that of *bound states*. Indeed, if we can find a solution $\psi \in L^2(\mathbf{R}^n)$ to the time-independent linear Schrödinger equation $-\Delta\psi + V\psi = -E\psi$ for some E ¿ 0, then one easily verifies that the function $u(t,x) := e^{-iEt}\psi(x)$ is a solution to the time-varying linear Schrödinger equation which refuses to disperse in the sense of (3.5) or (3.6); indeed, we have the opposite property that the L^2 density $|u(t,x)|^2$ is static in time. One can then use the principle of superposition to create some more non-dispersing solutions by adding several bound states together, or perhaps adding some bound states to some radiating states. The famous RAGE theorem [Ru1969, Am1973, En1978] asserts, roughly speaking, that there are no further types of states, and that every state decomposes uniquely into a bound state and a radiating state. For instance, if a solution fails to obey the weak dispersion property (3.6), then it must necessarily have a non-zero correlation (inner product) with a bound state. (If instead it fails to obey the strong dispersion property (3.5), the situation is trickier, as there is unfortunately a third type of state, a "singular continuous spectrum" state,

which one might correlate with.) More generally, an arbitrary solution will decompose uniquely into the sum of a radiating state obeying (3.6), and a unconditionally convergent linear combination of bound states. The proof of these facts largely rests on the spectral theorem for the underlying Hamiltonian $-\Delta+V$; the bound states correspond to pure point spectrum, the weak dispersion property (3.6) corresponds to continuous spectrum, and the strong dispersion property (3.5) corresponds to absolutely continuous spectrum. Thus the RAGE theorem gives a nice connection between dynamics and spectral theory. Let us now turn to nonlinear Schrödinger equations. There are a large number of such equations one could study, but let us restrict attention to a particularly intensively studied case, the *cubic nonlinear Schrödinger equation* (NLS)

$$iu_t + \Delta u = \mu |u|^2 u$$

where μ is either equal to $+1$ (the *defocusing* case) or -1 (the *focusing* case). (This particular equation arises often in physics as the leading approximation to a Taylor expansion to more complicated dispersive models, such as those for plasmas, mesons, or Bose-Einstein condensates.) We specify initial data $u(0,x) = u_0(x)$ as per usual, and to avoid technicalities we place this initial data in the Schwartz class. Unlike the linear case, it is no longer automatic that smooth solutions exist globally in time, although it is not too hard to at least establish local existence of smooth solutions. There are thus several basic questions:

1. (Global existence) Under what conditions do smooth solutions u to NLS exist globally in time?

2. (Asymptotic behaviour, global existence case) If there is global existence, what is the limiting behaviour of $u(t)$ in the limit as t goes to infinity?

3. (Asymptotic behaviour, blowup case) If global existence fails, what is the limiting behaviour of $u(t)$ in the limit as t approaches the maximal time of existence?

For reasons of time and space I will focus only on Questions 1 and 2, although Question 3 is very interesting (and very difficult). The answer to these questions is rather complicated (and still unsolved in several cases), depending on the sign μ of the nonlinearity, the ambient dimension n, and the size of the initial data. Here are some sample results regarding Question 1 (most of which can be found for instance in [Caz2003] or [Ta2006d]):

- If $n = 1$, then one has global smooth solutions for arbitrarily large data and any choice of sign.

- For $n = 2,3,4$, one has global smooth solutions for arbitrarily large data in the defocusing case (this is particularly tricky[RyVi2007] in the energy-critical case $n = 4$), and for small data in the focusing case. For large data in the focusing case, finite time blowup is possible.

- For $n \geq 5$, one has global smooth solutions for small data with either sign. For large data in the focusing case, finite time blowup is possible. For large data in the defocusing case, the existence of global smooth solutions are unknown

even for spherically symmetric data, indeed this problem, being supercritical, is of comparable difficulty to the Navier-Stokes global regularity problem (Section 1.4).

Incidentally, the relevance of the sign μ can be seen by considering the conserved Hamiltonian

$$H(u_0) = H(u(t)) := \int_{\mathbf{R}^n} \frac{1}{2}|\nabla u(t,x)|^2 + \mu\frac{1}{4}|u(t,x)|^4\,dx.$$

In the defocusing case the Hamiltonian is positive definite and thus coercive; in the focusing case it is indefinite, though in low dimensions and in conjunction with the L^2 conservation law one can sometimes recover coercivity.

Let us now assume henceforth that the solution exists globally (and, to make a technical assumption, also assume that the solution stays bounded in the energy space $H^1(\mathbf{R}^n)$) and consider Question 2. As in the linear case, we can see two obvious possible asymptotic behaviours for the solution u(t). Firstly there is the dispersive or radiating scenario in which (3.5) or (3.6) occurs. (For technical reasons relating to Galilean invariance, we have to allow the compact set K to be translated in time by an arbitrary time-dependent displacement $x(t)$, unless we make the assumption of spherical symmetry; but let us ignore this technicality.) This scenario is known to take place when the initial data is sufficiently small. (Indeed, it is conjectured to take place whenever the data is "strictly smaller" in some sense than that of the smallest non-trivial bound state, aka the *ground state*; there has been some recent progress on this conjecture [KeMe2006], [HoRo2008] in the spherically symmetric case.) In dimensions $n = 1,3,4$, this scenario is also known to be true for large data in the defocusing case (the case $n = 1$ by inverse scattering considerations[No1980], the case $n = 3$ by Morawetz inequalities[GiVe1985], and the case $n = 4$ by the recent work in [RyVi2007]; the $n = 2$ case is a major open problem).

The opposite scenario is that of a *nonlinear bound state* $u(t,x) = e^{-iEt}\psi(x)$, where $E > 0$ and ψ solves the time-independent NLS $-\Delta\psi + \mu|\psi|^2\psi = -E\psi$. From the Pohozaev identity or the Morawetz inequality one can show that non-trivial bound states only exist in the focusing case $\mu = -1$, and in this case one can construct such states, for instance by using the work of Berestycki and Lions[BeLi1980]. Solutions constructed using these nonlinear bound states are known as *stationary solitons* (or stationary solitary waves). By applying the Galilean invariance of the NLS equation one can also create *travelling solitons*. With some non-trivial effort one can also combine these solitons with radiation (as was done recently in three dimensions [Be2007]), and one should also be able to combine distant solitons with each other to form *multisoliton solutions* (this has been achieved in one dimension by inverse scattering methods[No1980], as well as for the gKdV equation[MaMeTs2002] which is similar in many ways to NLS.) Presumably one can also form solutions which are a superposition of multisolitons and radiation.

The *soliton resolution conjecture* asserts that for "generic" choices of (arbitrarily large) initial data to an NLS with a global solution, the long-time behaviour of the solution should eventually resolve into a finite number of receding solitons (i.e. a multisoliton solution), plus a radiation component which decays in senses such as of (3.5)

or (3.6). (For short times, all kinds of things could happen, such as soliton collisons, solitons fragmenting into radiation or into smaller solitons, etc., and indeed this sort of thing is observed numerically.) This conjecture (which is for instance discussed in [So2006], [Ta2004], [Ta2007d]) is still far out of reach of current technology, except in the special one-dimensional case $n = 1$ when the equation miraculously becomes completely integrable, and the solutions can be computed rather explicitly via inverse scattering methods, as was for instance carried out by Novoksenov[No1980]. In that case the soliton resolution conjecture was indeed verified for generic data (in which the associated Lax operator had no repeated eigenvalues or resonances), however for exceptional data one could have a number of exotic solutions, such as a pair of solitons receding at a logarithmic rate from each other, or of periodic or quasiperiodic "breather solutions" which are not of soliton form.

Based on this one-dimensional model case, we expect the soliton resolution conjecture to hold in higher dimensions also, assuming sufficient uniform bounds on the global solution to prevent blowup or "weak turbulence" from causing difficulties. However, the fact that a good resolution into solitons is only expected for "generic" data rather than all data makes the conjecture extremely problematic, as almost all of our tools are based on a worst-case analysis and thus cannot obtain results that are only supposed to be true generically. (This is also a difficulty which seems to obstruct the global solvability of Navier-Stokes, as discussed in Section 1.4.) Even in the spherically symmetric case, which should be much simpler (in particular, the solitons must now be stationary and centred at the origin), the problem is wide open.

Nevertheless, there is some recent work which gives a small amount of progress towards the soliton resolution conjecture. For spherically symmetric energy-bounded global solutions (of arbitrary size) to the focusing cubic NLS in three dimensions, it is a result of myself [Ta2004] that the solution ultimately decouples into a radiating term obeying (3.5), plus a "weakly bound state" which is asymptotically orthogonal to all radiating states, is uniformly smooth, and exhibits a weak decay at spatial infinity. If one is willing to move to five and higher dimensions and to weaken the strength of the nonlinearity (e.g. to consider quadratic NLS in five dimensions) then a stronger result[Ta2007d] is available under similar hypotheses, namely that the weakly bound state is now *almost periodic*, ranging inside of a fixed compact subset of energy space, thus providing a "dispersive compact attractor" for this equation. In principle, this brings us back to the realm of dynamical systems, but we have almost no control on what this attractor is (though it contains all the soliton states and respects the symmetries of the equation), and so it is unclear what the next step should be. (There is a similar result in the non-radial case which is more complicated to state: see [Ta2007d] for more details.)

3.1.4 Notes

These articles were originally posted on Apr 5-8, 2007 at

terrytao.wordpress.com/2007/04/05

terrytao.wordpress.com/2007/04/07

terrytao.wordpress.com/2007/04/08

3.2 Ostrowski Lecture: The uniform uncertainty principle and compressed sensing

As mentioned in Section 2.2, *compressed sensing* is a relatively new measurement paradigm which seeks to capture the "essential" aspects of a high-dimensional object using as few measurements as possible. There are many contexts in which compressed sensing is potentially useful, but for this talk (which is focussed on theory rather than applications) I will just consider a single toy model arising from Fourier analysis. Specifically, the object we seek to measure will be a complex function $f : \mathbf{Z}/N\mathbf{Z} \to \mathbf{C}$ on the cyclic group of N elements (or if you wish, a column vector of N complex numbers). In practice one should think of N as being of moderately large size, e.g. $N \sim 10^6$. Given such a function, one can form its (discrete) Fourier transform $\hat{f} : \mathbf{Z}/N\mathbf{Z} \to \mathbf{C}$; for this talk it will be convenient to normalise this Fourier transform as

$$\hat{f}(\xi) = \frac{1}{\sqrt{N}} \sum_{x \in \mathbf{Z}/N\mathbf{Z}} f(x) e^{-2\pi i x\xi/N}.$$

We suppose that we can measure some (but perhaps not all) of the Fourier coefficients of f, and ask whether we can reconstruct f from this information; the objective is to use as few Fourier coefficients as possible. More specifically, we fix a set $\Lambda \subset \mathbf{Z}/N\mathbf{Z}$ of "observable" frequencies, and pose the following two questions:

1. Let N be a known integer, let $f : \mathbf{Z}/N\mathbf{Z} \to \mathbf{C}$ be an unknown function, let $\Lambda \subset \mathbf{Z}/N\mathbf{Z}$ a known set of frequencies, and let $c_\xi = \hat{f}(\xi)$ be a sequence of known Fourier coefficients of f for all $\xi \in \Lambda$. Is it possible to reconstruct f uniquely from this information?

2. If so, what is a practical algorithm for finding f?

For instance, if Λ is the whole set of frequencies, i.e. $\Lambda = \mathbf{Z}/N\mathbf{Z}$, then the answer to Q1 is "yes" (because the Fourier transform is injective), and an answer to Q2 is provided by the Fourier inversion formula

$$f(x) = \frac{1}{\sqrt{N}} \sum_{\xi \in \mathbf{Z}/N\mathbf{Z}} c_\xi e^{2\pi i x/\xi}$$

which can be computed quite quickly, for instance by using the fast Fourier transform.

Now we ask what happens when Λ is a proper subset of $\mathbf{Z}/N\mathbf{Z}$. Then the answer to Q1, as stated above, is "no" (and so Q2 is moot). One can see this abstractly by a degrees-of-freedom argument: the space of all functions f on N points has N degrees of freedom, but we are only making $|\Lambda|$ measurements, thus leaving $N - |\Lambda|$ remaining degrees of freedom in the unknown function f. If $|\Lambda|$ is strictly less than N, then there are not enough measurements to pin down f precisely. More concretely, we can easily use the Fourier inversion formula to concoct a function f which is not identically zero, but whose Fourier transform vanishes on Λ (e.g. consider a plane wave whose frequency lies outside of Λ). Such a function is indistinguishable from the zero function as far as the known measurements are concerned.

However, we can hope to recover unique solvability for this problem by making an additional hypothesis on the function f. There are many such hypotheses one could make, but for this toy problem we shall simply assume that f is *sparse*. Specifically, we fix an integer S between 1 and N, and say that a function f is *S-sparse* if f is non-zero in at most S places, or equivalently if the support $supp(f) := \{x \in \mathbf{Z}/N\mathbf{Z} : f(x) \neq 0\}$ has cardinality less than or equal to S. We now ask the following modified versions of the above two questions:

1. Let S and N be known integers, let $f : \mathbf{Z}/N\mathbf{Z} \to \mathbf{C}$ be an unknown S-sparse function, let $\Lambda \subset \mathbf{Z}/N\mathbf{Z}$ a known set of frequencies, and let $c_\xi = \hat{f}(\xi)$ be a sequence of known Fourier coefficients of f for all $\xi \in \Lambda$. Is it possible to reconstruct f uniquely from this information?

2. If so, what is a practical algorithm for finding f?

Note that while we know how sparse f is, we are not given to know exactly *where* f is sparse - there are S positions out of the N total positions where f might be non-zero, but we do not know which S positions these are. The fact that the support is not known a priori is one of the key difficulties with this problem. Nevertheless, setting that problem aside for the moment, we see that f now has only S degrees of freedom instead of N, and so by repeating the previous analysis one might now hope that the answer to Q1 becomes yes as soon as $|\Lambda| \geq S$, i.e. one takes at least as many measurements as the sparsity of f.

Actually, one needs at least $|\Lambda| \geq 2S$ (if $2S$ is less than or equal to N), for the following reason. Suppose that $|\Lambda|$ was strictly less than $2S$. Then the set of functions supported on $\{1,\ldots,2S\}$ has more degrees of freedom than are measured by the Fourier coefficients at Λ. By elementary linear algebra, this therefore means that there exists a function f supported on $\{1,\ldots,2S\}$ whose Fourier coefficients vanish on Λ, but is not identically zero. If we split $f = f_1 - f_2$ where f_1 is supported on $\{1,\ldots,S\}$ and f_2 is supported on $\{S+1,\ldots,2S\}$, then we see that f_1 and f_2 are two distinct S-sparse functions whose Fourier transforms agreed on Λ, thus contradicting unique recoverability of f.

One might hope that this necessary condition is close to being sharp, so that the answer to the modified Q1 is yes as soon as $|\Lambda|$ is larger than $2S$. By modifying the arguments of the previous paragraph we see that Q1 fails if and only if there exists a non-zero $2S$-sparse function whose Fourier transform vanished on all of Λ, but one can hope that this is not the case, because of the following heuristic:

Principle 3.3 (Uncertainty principle)**.** *(informal) If a function is sparse and not identically zero, then its Fourier transform should be non-sparse.*

This type of principle is motivated by the *Heisenberg uncertainty principle* in physics; the size of the support of f is a proxy for the spatial uncertainty of f, whereas the size of the support of $\hat{f}$ is a proxy for the momentum uncertainty of f. There are a number of ways to make this principle precise. One standard one is

Proposition 3.4 (Discrete uncertainty principle)**.** *If f is not identically zero, then* $|\mathrm{supp}(f)| \times |\mathrm{supp}(\hat{f})| \geq N$.

Proof. Combine the Plancherel identity

$$\sum_{x \in \mathbf{Z}/N\mathbf{Z}} |f(x)|^2 = \sum_{\xi \in \mathbf{Z}/N\mathbf{Z}} |\hat{f}(\xi)|^2$$

with the elementary inequalities

$$\sup_{x \in \mathbf{Z}/N\mathbf{Z}} |f(x)| \leq \frac{1}{\sqrt{N}} \sum_{\xi \in \mathbf{Z}/N\mathbf{Z}} |\hat{f}(\xi)|$$

$$\sum_{x \in \mathbf{Z}/N\mathbf{Z}} |f(x)|^2 \leq |supp(f)|(\sup_{x \in \mathbf{Z}/N\mathbf{Z}} |f(x)|)^2$$

and

$$\sum_{\xi \in \mathbf{Z}/N\mathbf{Z}} |\hat{f}(\xi)| \leq |supp(\hat{f})|^{1/2}(\sum_{\xi \in \mathbf{Z}/N\mathbf{Z}} |\hat{f}(\xi)|^2)^{1/2}.$$

□

By applying this principle we see that we obtain unique recoverability as soon as $|\Lambda| > N - N/2S$, but this is a rather lousy criterion - for most values of S, it means that one has to measure almost all of the Fourier coefficients to recover f! At that point one may as well measure all of the coefficients so that one can recover both sparse and non-sparse f easily via the Fourier inversion formula.

There are cases in which the condition $|\Lambda| > N - N/2S$ cannot be improved. A good example is provided by the *Dirac comb*, in which N is a square number, and f is the indicator function of the multiples $\{0, \sqrt{N}, 2\sqrt{N}, \ldots, N - \sqrt{N}\}$ of $\sqrt{N}$. As is well known, this function is its own Fourier transform. If we then set Λ to be the complement of the multiples of $\sqrt{N}$, then we see that even though we are measuring almost all the frequencies - $N - \sqrt{N}$ of them to be precise - we cannot distinguish the $\sqrt{N}$-sparse Dirac comb f from the zero function; we have unluckily chosen only those frequencies Λ which totally fail to intersect the support of the Fourier transform of f.

One can concoct several further counterexamples of this type, but they require various subgroups of $\mathbf{Z}/N\mathbf{Z}$ (such as the multiples of $\sqrt{N}$, when N is square). One can ask what happens when the ambient group has no non-trivial subgroups, i.e. when N is prime. Then things become better, thanks to a classical result of Chebotarev:

Lemma 3.5 (Chebotarev lemma)**.** *Let N be prime. Then every square minor of the Fourier matrix $(e^{2\pi ijk/N})_{1 \leq j,k \leq N}$ is invertible.*

This lemma has been rediscovered or reproved at least seven times in the literature (I myself rediscovered it once [Ta2005]); there is a nice survey paper on Chebotarev's work in [StLe1996] which summarises some of this. (A nice connection to the setting of this talk, pointed out to me by Hendrik Lenstra here at Leiden: Chebotarev's lemma was originally conjectured to be true by Ostrowski.) As a quick corollary of this lemma, we obtain the following improvement to the uncertainty principle in the prime order case:

Corollary 3.6 (Uncertainty principle for cyclic groups of prime order)**.** *If N is prime and f is not identically zero, then $|\operatorname{supp}(f)| + |\operatorname{supp}(\hat{f})| \geq N + 1$.*

From this one can quickly show that one does indeed obtain unique recoverability for S-sparse functions in cyclic groups of prime order whenever one has $|\Lambda| \geq 2S$, and that this condition is absolutely sharp. (There is also a generalisation of the above uncertainty principle to composite Ndue to Meshulam[Me2006].)

This settles the (modified) Q1 posed above, at least for groups of prime order. But it does not settle Q2 - the question of exactly *how* one recovers f from the given data $N, S, \Lambda, (c_\xi)_{\xi\in\Lambda}$. One can consider a number of simple-minded strategies to recover f:

1. **Brute force**. If we knew precisely what the support $\mathrm{supp}(f)$ of f was, we can use linear algebra methods to solve for f in terms of the coefficients c_ξ, since we have $|\Lambda|$ equations in S unknowns (and Lemma 3.5 guarantees that this system has maximal rank). So we can simply exhaust all the possible combinations for $supp(f)$ (there are roughly $\binom{N}{S}$ of these) and apply linear algebra to each combination. This works, but is horribly computationally expensive, and is completely impractical once N and S are of any reasonable size, e.g. larger than 1000.

2. l^0 **minimisation**. Out of all the possible functions f which match the given data (i.e. $\hat{f}(\xi) = c_\xi$ for all $\xi \in \Lambda$), find the sparsest such solution, i.e. the solution which minimises the "l^0 norm" $\|f\|_{l^0} := \sum_{x\in\mathbf{Z}/N\mathbf{Z}} |f(x)|^0 = |\mathrm{supp}(f)|$. This works too, but is still impractical: the general problem of finding the sparsest solution to a linear system of equations contains the infamous *subset sum* decision problem as a special case (we'll leave this as an exercise to the reader) and so this problem is NP-hard in general. (Note that this does *not* imply that the original problem Q1 is similarly NP-hard, because that problem involves a specific linear system, which turns out to be rather different from the specific linear system used to encode subset-sum.)

3. l^2 **minimisation** (i.e. the method of least squares). Out of all the possible functions f which match the given data, find the one of least energy, i.e. which minimises the l^2 norm $\|f\|_{l^2} := (\sum_{x\in\mathbf{Z}/N\mathbf{Z}} |f(x)|^2)^{1/2}$. This method has the advantage (unlike 1. and 2.) of being extremely easy to carry out; indeed, the minimiser is given explicitly by the formula $f(x) = \frac{1}{\sqrt{N}} \sum_{\xi\in\Lambda} c_\xi e^{2\pi i x\xi/N}$. Unfortunately, this minimiser not guaranteed at all to be S-sparse, and indeed the uncertainty principle suggests in fact that the l^2 minimiser will be highly non-sparse.

So we have two approaches to Q2 which work but are computationally infeasible, and one approach which is computationally feasible but doesn't work. It turns out however that one can take a "best of both worlds" approach halfway between method 2. and method 3., namely:

4. l^1 **minimisation** (or *basis pursuit*): Out of all the possible functions f which match the given data, find the one of least mass, i.e. which minimises the l^1 norm $\|f\|_{l^1} := \sum_{x\in\mathbf{Z}/N\mathbf{Z}} |f(x)|$.

The key difference between this minimisation problem and the l^0 problem is that the l^1 norm is *convex*, and so this minimisation problem is no longer NP-hard but can be solved in reasonable (though not utterly trivial) time by convex programming techniques such as the simplex method. So the method is computationally feasible; the

only question is whether the method actually works to recover the original S-sparse function f.

Before we reveal the answer, we can at least give an informal geometric argument as to why l^1 minimisation is more likely to recover a sparse solution than l^2 minimisation. The set of all f whose Fourier coefficients match the observed data c_ξ forms an affine subspace of the space of all functions. The l^2 minimiser can then be viewed geometrically by taking l^2 balls (i.e. Euclidean balls) centred at the origin, and gradually increasing the radius of the ball until the first point of contact with the affine subspace. In general, there is no reason to expect this point of contact to be sparse (i.e. to lie on a high-codimension coordinate subspace). If however we replace l^2 with l^1, then the Euclidean balls are replaced by octahedra, which are much "pointier" (especially in high dimensions) and whose corners lie on coordinate subspaces. So the point of first contact is now much more likely to be sparse. The idea of using l^1 as a "convex relaxation" of l^0 is a powerful one in applied mathematics; see for instance [Tr2006].

It turns out that if Λ and f are structured in a perverse way, then basis pursuit does not work (and more generally, any algorithm to solve the problem is necessarily very unstable). We already saw the Dirac comb example, which relied on the composite nature of N. But even when N is prime, we can construct pseudo-Dirac comb examples which exhibit the problem: if f is for instance a discretised bump function adapted to an arithmetic progression such as $\{-\lfloor\sqrt{N}\rfloor,\ldots,\lfloor\sqrt{N}\rfloor\}$, then elementary Fourier analysis reveals that the Fourier transform of f will be highly concentrated (though not completely supported) on a dual progression (which in the above example will also be basically $\{-\lfloor\sqrt{N}\rfloor,\ldots,\lfloor\sqrt{N}\rfloor\}$, and have a rapidly decreasing tail away from this progression. (This is related to the well-known fact that the Fourier transform of a Schwartz function is again a Schwartz function.) If we pick Λ to be far away from this progression - e.g. $\Lambda = \{\lfloor N/3\rfloor,\ldots,\lfloor 2N/3\rfloor\}$, then the Fourier transform will be very small on Λ. As a consequence, while we know abstractly that exact reconstruction of f is possible if N is a large prime assuming infinite precision in the measurements, any presence of error (e.g. roundoff error) will mean that f is effectively indistinguishable from the zero function. In particular it is not hard to show that basis pursuit fails in general in this case.

The above counterexamples used very structured examples of sets of observed frequencies Λ, such as arithmetic progressions. On the other hand, it turns out, remarkably enough, that if instead one selects *random* sets of frequencies Λ of some fixed size $|\Lambda|$ (thus choosing uniformly at arndom among all the $\binom{N}{|\Lambda|}$ possibilities), then things become much better. Intuitively, this is because all the counterexamples that obstruct solvability tend to have their Fourier transform supported in very structured sets, and the dichotomy between structure and randomness means that a random subset of $\mathbf{Z}/N\mathbf{Z}$ is likely to contain a proportional fraction of all structured sets. One specific manifestation of this is

Proposition 3.7 (Uniform Uncertainty Principle (UUP))**.** *If Λ is a random set with $|\Lambda| \gg S\log^4 N$, then with overwhelming probability (at least $1 - O(N^{-A})$ for any fixed*

A), we have the approximate local Plancherel identity

$$\sum_{\xi\in\Lambda} |\hat{f}(\xi)|^2 \approx \frac{|\Lambda|}{N} \sum_{x\in\mathbf{Z}/N\mathbf{Z}} |f(x)|^2$$

for all 4S-sparse functions f, where by $X \approx Y$ we mean that X and Y differ by at most 10% (say). (N is not required to be prime.)

The above formulation is a little imprecise; see [CaTa2006], [RuVe2006] for more rigorous versions. This principle asserts that if the random set Λ is just a little bit bigger than S (by a couple logs), then it is not possible for the Fourier transform of an S-sparse function to avoid Λ, and moreover the set Λ must receive its "fair share" of the l^2 energy, as predicted by Plancherel's theorem. The "uniform" nature of this principle refers to the fact that it applies for *all* S-sparse functions f, with no exceptions. For a single function f, this type of localisation of the Plancherel identity is quite easy to prove using Chernoff's inequality. To extend this to all sparse f, the main strategy (first used in this type of context in [Bo1989]) is to exploit the fact that the set of sparse f has low metric entropy and so can be described efficiently by a relatively small number of functions. (The principle cannot possibly extend to *all* functions f, since it is certainly possible to create non-zero functions whose Fourier transform vanishes everywhere on Λ.)

By using this principle (and variants of this principle), one can indeed show that basis pursuit works:

Theorem 3.8. *[CaRoTa2006], [CaTa2006] Suppose Λ is a random set with $|\Lambda| \gg S\log N$. Then any given S-sparse function f will be recovered exactly by l^1 minimisation with overwhelming probability. If one makes the stronger hypothesis $|\Lambda| \gg S\log^4 N$, then with overwhelming probability* all *S-sparse functions will be recovered exactly by l^1 minimisation. (Again, N is not required to be prime.)*

Roughly speaking, the idea in the latter result is to use the UUP to show that the Fourier coefficients of any sparse (or l^1-bounded) competitor with disjoint support to the true S-sparse solution is going to be rather orthogonal to the true solution, and thus unlikely to be present in an l^1 minimiser. The former result is more delicate and combinatorial, and requires computing high moments of random Fourier minors.

The method is rather robust; there is some followup work[CaRoTa2006b] which demonstrates stability of the basis pursuit method with respect to several types of noise; see also the survey [Ca2006]. It can also be abstracted from this toy Fourier problem to a more general problem of recovering sparse or compressible data from few measurements. As long as the measurement matrix obeys an appropriate generalisation of the UUP, the basis pursuit methods are quite effective (both in theory, in numerical experiments, and more recently in laboratory prototypes).

3.2.1 Notes

This article was originally posted on Apr 15, 2007 at
terrytao.wordpress.com/2007/04/15

Utpal Sarkar raised the interesting question of whether some analogue of Corollary 3.6 for arbitrary abelian groups (beyond those in [Me2006]) could be established under the additional assumption that f and $\hat{f}$ are not supported in subgroups (or cosets of subgroups).

3.3 Milliman Lecture Series: Recent developments in arithmetic combinatorics

This week I am visiting the University of Washington in Seattle, giving the Milliman Lecture Series for 2007-2008. My chosen theme here is "Recent developments in arithmetic combinatorics". In my first lecture, I will speak (once again) on how methods in additive combinatorics have allowed us to detect additive patterns in the prime numbers, in particular discussing my joint work with Ben Green. In the second lecture I will discuss how additive combinatorics has made it possible to study the invertibility and spectral behaviour of random discrete matrices, in particular discussing my joint work with Van Vu; and in the third lecture I will discuss how sum-product estimates have recently led to progress in the theory of expanders relating to Lie groups, as well as to sieving over orbits of such groups, in particular presenting work of Jean Bourgain and his coauthors.

3.3.1 Additive combinatorics and the primes

Additive combinatorics is focused, among other things, on the task of studying additive patterns in general sets of integers (or more generally, sets in an additive group). It is descended in part from the more classical subject of *additive number theory*: the study of additive patterns, structures, and operations on explicit sets of integers, such as the primes $\mathscr{P} := \{2,3,5,7,11,\ldots\}$ and the squares $\mathscr{S} := \{0,1,4,9,16,\ldots\}$. Here are some typical results and conjectures in the subject for both the primes and the squares, side by side for comparison:

- Lagrange's four square theorem: For every positive integer N, there exists a pattern in $\mathscr{S}$ of the form $a,b,c,N-a-b-c$.
- Vinogradov's theorem: For every sufficiently large integer N, there exists a pattern in $\mathscr{P}$ of the form $a,b,c,N-a-b-c$.
- Fermat's two square theorem: For every prime number $N = 1 \bmod 4$, there exists a pattern in $\mathscr{S}$ of the form $a,N-a$.
- Even Goldbach conjecture: For every even number $N \geq 4$, there exists a pattern in $\mathscr{P}$ of the form $a,N-a$.
- Fermat's four square theorem: There does not exist any pattern in $\mathscr{S}$ of the form $a,a+b,a+2b,a+3b$ with $b \neq 0$.
- Green-Tao theorem: For any $k \geq 1$, there exist infinitely many patterns in $\mathscr{P}$ of the form $a,a+b,\ldots,a+(k-1)b$ with $b \neq 0$.
- Pell's equation: There are infinitely many patterns in $\mathscr{S}$ of the form $a,2a+1$.
- Sophie Germain conjecture: There are infinitely many patterns in $\mathscr{P}$ of the form $a,2a+1$.

I have deliberately phrased the above results in a unified format, namely that of counting additive patterns with one or more free parameters $a, b, \ldots$ in either the squares or the primes. However, this apparent unification is actually an illusion: the results involving square numbers are much older (the Pell equation solutions, for instance, was essentially known to Diophantus, as well as the ancient Indians) and are proven using completely different methods than for the prime numbers. For the square numbers, there are some key algebraic identities and connections, ranging from the high-school factorisations $a^2 - b^2 = (a-b)(a+b)$, $a^2 - 2b^2 = (a-\sqrt{2}b)(a+\sqrt{2}b)$ and $a^2 + b^2 = (a-ib)(a+ib)$ to deeper connections between quadratic forms and elliptic curves, which allow one to prove the results on the left-hand column via the methods of algebraic number theory. For the primes, on the other hand, there are very few usable algebraic properties available: one has *local* (mod q) information, such as the fact that all primes are odd (with one exception), or adjacent to a multiple of 6 (with two exceptions), but there are essentially no *global* algebraic identities or structures to exploit amongst the primes (except, perhaps, for the identities such as the Euler product formula $\zeta(s) = \prod_p (1-p^{-s})^{-1}$ connecting the prime numbers to the Riemann zeta function and its relatives, although this only directly helps one count *multiplicative* patterns in the primes rather than additive ones). So, whereas the square numbers can be profitably studied by cleverly exploiting their special algebraic structure, when dealing with the prime numbers it is in fact better to rely instead on more general tools which require very little structural control on the set being studied. In particular, in recent years we have learned that the methods of *additive combinatorics*, which offers tools to count additive patterns in *arbitrary* sets of integers (or more generally, subsets of an additive group), can be remarkably effective in additive prime number theory. Thus - rather counter-intuitively - some of our strongest results about additive patterns in primes have been obtained by using very little information about the primes at all!

To give a very simple example of how additive combinatorics can be applied to the primes, let us consider the problem of finding *parallelograms* inside the primes - patterns of the form $a, a+b, a+c, a+b+c$ with b, c positive integers; for instance, 3, 7, 43, 47 is a parallelogram of primes. It is very hard to produce any parallelograms of primes by algebraic means (such as an explicit formula); however, there is a simple combinatorial argument that shows that such parallelograms exist in abundance. The only actual information needed about the primes for this argument is the *prime number theorem*, which says that the number of primes less than a large number N is equal to $(1+o(1))N/\log N$ in the limit $N \to \infty$. (Actually, we won't even need the full strength of the prime number theorem; the weaker statement that there are $\gg N/\log N$ primes less than N which was known to Chebyshev and can be established by elementary means based on the prime factorisation of $\binom{2N}{N}$, will suffice.)

Let N be a large number, then there are $(1+o(1))N/\log N$ primes less than N. This allows us to form roughly $(\frac{1}{2}+o(1))N^2/\log^2 N$ differences $p-q$ of primes $1 < q < p \leq N$. But each of these differences takes values between 1 and N. For N large enough, we can thus use the pigeonhole principle to conclude that there are two differences $p-q$ and $p'-q'$ of primes which have the same value, which implies that the quadruplet p, q, q', p' forms a parallelogram. In fact, a slight refinement this argument (using the Cauchy-Schwarz inequality, which can provide a more quantitative version of the pigeonhole principle) shows that there are $\gg N^3/\log^4 N$ parallelograms of primes less

than N, and in particular that there are infinitely many parallelograms of primes.

The above example shows how one can detect additive patterns in the primes using very little information about the primes themselves; in the above case, the only information we actually needed about the primes was about their cardinality. (Indeed, the argument is not really about primes at all, and is best viewed as a general statement about dense sets of integers, known as the *Szemerédi cube lemma*.) More generally, the strategy of the additive combinatorial approach is to minimise the number of facts one actually needs to establish about the primes, and rely primarily on tools which are valid for rather general classes of sets of integers.

A good example of this type of tool is *Szemerédi's theorem*[Sz1975], which asserts any set of integers A of positive density contains arbitrarily long arithmetic progressions; as with the case of parallelograms, the only information needed about the set is that it is large. This theorem does not directly apply to the prime numbers $\mathscr{P}$, as they have density zero, but it turns out that there is a trick (which Ben Green and I call the *transference principle*) which (very roughly speaking) lets one locate a dense set of integers A which "models" the primes, in the sense that there is a relationship between additive patterns in A and additive patterns in $\mathscr{P}$. (The relationship here is somewhat analogous to Monte Carlo integration, which uses the average value of a function f on a sparse pseudorandom set to approximate the average value of f on a much larger domain.) As a consequence of this principle, Ben and I were able to use Szemerédi's theorem to establish that the primes contained arbitrarily long arithmetic progressions. There have since been a number of similar results in which Szemerédi-type results for dense sets of integers have been transferred to yield similar statements about the primes and related sets (e.g. constellations in Gaussian primes[Ta2006g], or polynomial patterns in the ordinary primes[TaZi2008]).

In this talk, though, I am not going to discuss the above results further, but instead focus on the task of using additive combinatorics to detect more general classes of additive patterns in sets of integers such as the primes, with the philosophy of always trying to use as little structural information about these sets as possible.

To illustrate some of the ideas, let us consider the odd Goldbach conjecture, which asserts that any odd integer larger than 5 can be expressed as the sum of three primes. Let's first tackle a model problem in the same spirit: we'll work in a cyclic group $\mathbf{Z}/N\mathbf{Z}$ instead of the integers, we will pick three sets A, B, C in this group as well as an element x, and we will ask whether x can be expressed as the sum of an element a from A, an element b from B, and an element c from C.

Of course, to make any headway on this problem we have to make some assumptions on A, B, C. Let us first assume that A, B, C are fairly dense subsets of $\mathbf{Z}/N\mathbf{Z}$, say $|A|, |B|, |C| \geq \frac{1}{10}N$. Even with such large sets, there is no guarantee that every element x can be expressed as a sum of elements from A, B, and C respectively. For instance, if $A = B = C = \{1, \ldots, \lfloor N/10 \rfloor + 1\}$, we see that only about 30% of the numbers in $\mathbf{Z}/N\mathbf{Z}$ can be expressed in this way. Or, if N is a multiple of 10 and $A = B = C = \{10, 20, 30, \ldots\}$ consist of those elements in $\mathbf{Z}/N\mathbf{Z}$ which are multiples of 10, then only 10% of the elements in $\mathbf{Z}/N\mathbf{Z}$ can be expressed in this fashion. Thus there are some non-trivial obstructions to this Goldbach-type problem.

However, it turns out that if one of the sets, say A, is sufficiently "uniform" or "pseudorandom", then one can always solve this Goldbach-type problem, regardless of

what the other two sets are. This type of fact is often established by Fourier-analytic means (or by closely related techniques, such as spectral theory), but let me give a heuristic combinatorial argument to indicate why one would expect this type of phenomenon to occur. We will work in the contrapositive: we assume that we can find an x which cannot be expressed as the sum of elements from A, B, and C, and somehow eliminate the role of x, B, and C to to deduce some "non-uniformity" or "structure" for A.

So, suppose that $x \neq a+b+c$ for all a in A, b in B, c in C. This implies that $x-a-b$ always avoids C. Thus $x-a-b$ does not range freely throughout $\mathbf{Z}/N\mathbf{Z}$, but is instead concentrated in a set of 90% the size or smaller. Because of this more confined space, one would expect more "collisions" than usual, in the sense that there should be more solutions to the equation $x-a-b = x-a'-b'$ with a,a' in A and b,b' in B than one would normally expect. Rearranging, we conclude that there are more solutions to the equation $a-a' = b-b'$ than one might first expect. This means that the differences $a-a'$ and the differences $b-b'$ have to cluster in the same region of $\mathbf{Z}/N\mathbf{Z}$, which then suggests that we should have more collisions $a-a' = a''-a'''$ with a,a',a'',a'' in A than one might first expect. To put it another way, A should contain a higher than expected number of parallelograms $a, a+r, a+s, a+r+s$ (also known as *additive quadruples*).

The above argument can be made rigorous by two quick applications of the Cauchy-Schwarz inequality. If we had $|A|,|B|,|C| \geq \delta N$ for some $\delta > 0$, say, then it is not hard to use Cauchy-Schwarz to show that A will contain at least $\delta^4 N^3$ parallelograms (where we allow degenerate parallelograms, in order to simplify the formulae a little); but if there existed an x which was not the sum of an element from A, an element from B, and an element of C, one can use this to conclude that A must have a few more parallelograms, in fact it must have at least $(1+c\delta)\delta^4 N^3$ for some absolute constant $c > 0$.

Taking contrapositives, we conclude that if A has a near-minimal number of parallelograms (between $\delta^4 N^3$ and $(1+c\delta)\delta^4 N^3$), then we can solve this Goldbach-type problem for any x and any choice of sets B, C of density δ.

So, by using elementary additive combinatorics, we can reduce Goldbach-type problems to the problem of counting parallelograms in a given set A. But how can one achieve the latter task? It turns out that for this specific problem, there is an elegant formula from Fourier analysis: the number of parallelograms in a set $A \subset \mathbf{Z}/N\mathbf{Z}$ is equal to

$$N^3 \sum_{\xi \in \mathbf{Z}/N\mathbf{Z}} |\hat{1}_A(\xi)|^4 \tag{3.7}$$

where $\hat{1}_A(\xi)$ is the Fourier coefficient of the indicator function 1_A at frequency ξ:

$$\hat{1}_A(\xi) := \frac{1}{N} \sum_{x \in A} e^{-2\pi i x\xi/N}.$$

The connection between Fourier analysis and parallelograms $a, a+r, a+s, a+r+s$ might not be immediately explicable, but a link between the two can be discerned from the algebraic identity

$$e^{-2\pi i a\xi/N} e^{2\pi i (a+r)\xi/N} e^{2\pi i (a+s)\xi/N} e^{-2\pi i (a+r+s)\xi/N} = 1$$

which is a fancy way of saying that the linear function $x \mapsto x\xi/N \bmod 1$ has a vanishing second derivative.

Anyway, returning to the formula (3.7), in the case when A has density exactly δ, thus $|A| = \delta N$, we see that the number of parallelograms is equal to

$$(\delta^4 + \sum_{\xi \neq 0} |\hat{1}_A(\xi)|^4) N^3.$$

Thus we see (informally, at least) that a set A is going to have near-minimal number of parallelograms precisely when it is *Fourier-pseudorandom* in the sense that its Fourier coefficients at non-zero frequencies are all small, or in other words that the set A exhibits no correlation or bias with respect to any non-trivial linear phase function $x \mapsto e^{2\pi i x\xi/N}$. (It is instructive to consider our two counterexamples to the toy Goldbach problem, namely $A = \{1, \ldots, \lfloor N/10 \rfloor + 1\}$ and $A = \{10, 20, 30, \ldots\}$. The first set is biased with respect to the phase $x \mapsto e^{2\pi i x/N}$; the second set is biased with respect to $x \mapsto e^{2\pi i x/10}$.)

This gives us a strategy to solve Goldbach-type problems: if we can show somehow that a set A does not correlate strongly with any non-trivial linear phase function, then it should be sufficiently Fourier pseudorandom that there is no further obstruction to the Goldbach problem. If instead A does closely resemble something related to a non-trivial linear phase function, then that is quite a bit of structural information on A and that we should be able to solve the Goldbach type problem by explicit algebraic counting of solutions (as is for instance the case in the two model examples $A = \{1, 2, \ldots, \lfloor N/10 \rfloor + 1\}$ and $A = \{10, 20, 30, \ldots\}$ discussed earlier).

In the case of sets of integers such as the primes, this type of strategy is known as the *Hardy-Littlewood circle method*. It was successfully used by Vinogradov to establish his theorem that every sufficiently large odd number is the sum of three primes (and thus every sufficiently large number is the sum of four primes); the problem boils down to getting sufficiently strong estimates for exponential sums over primes such as $\sum_{p<N} e^{2\pi i \alpha p}$. In the "major arc" case where α is rational (or very close to rational) with small denominator then methods from multiplicative number theory, based on zeta functions and L-functions, become useful; in contrast, in the complementary "minor arc" case where α behaves irrationally, one can use more analytic methods (based, ultimately, on the equidistribution of multiples of α modulo 1 on the unit circle, and on the obvious fact that the product of two primes is a non-prime) to obtain good estimates. (I hope to discuss this in more detail in a later post.) A similar argument was used by van der Corput to establish that the prime numbers contained arbitrarily many arithmetic progressions of length three. These arguments are actually quite quantitative and precise; for instance, Vinogradov's theorem not only gives the existence of a representation $N = p_1 + p_2 + p_3$ of any sufficiently large odd number as the sum of three primes, but in fact gives an asymptotically precise formula as $N \to \infty$ as to how many such representations exist. Similarly, van der Corput's argument gives an asymptotically precise formula as to how many arithmetic progressions of length three consisting of primes less than N there are, as $N \to \infty$.

This strategy unfortunately fails miserably for the *even* Goldbach problem, which asks whether every even number greater than 2 is the sum of two primes; it turns out that there is no useful analogue of the parallelogram in this problem, basically due to

the fact that there is only one free parameter in the pattern one is looking for. However, it is possible to adapt the strategy to more complicated patterns with two or more free parameters, such as arithmetic progressions of length greater than three. For instance, if one wants to find arithmetic progressions of length 4 in a set A, it turns out that this problem is controlled by the number of *parallelopipeds*

$$a, a+r, a+s, a+t, a+r+s, a+r+t, a+s+t, a+r+s+t$$

that A contains, in much the same way that the odd Goldbach problem was controlled by parallelograms. So, if one knows how to count how many parallelopipeds there are in a set, one can (in principle) count how many progressions of length 4 there are as well (and one can also count a large variety of other patterns too). One would then hope for an elegant formula analogous to (3.7) to count these objects, but unfortunately it seems that no such formula exists. Part of the problem is that while parallelograms are closely tied to the linear (or Fourier) phases $x \mapsto x\xi/N$, because such phases have vanishing second derivative, parallelopipeds are more naturally tied to the larger class of phases which have vanishing third derivative, such as the quadratic phases $x \mapsto x^2\xi/N$. (Actually, there are also many more "pseudoquadratic" phases, such as $x \mapsto \lfloor \alpha x \rfloor \beta x$ for various real numbers α, β, whose third derivative exhibits some cancellation but does not vanish entirely, and are connected to flows on nilmanifolds, but I will not discuss them in detail here.) With this much larger class of potentially relevant phases, it appears that there is no useful analogue of the formula (3.7) (basically because there are so many such phases out there, most of which having no significant correlation with the set A, that the noise from these irrelevant phases drowns out the signal from those few phases that are actually important). Nevertheless, there are a set of tools, developed initially by Timothy Gowers, in what might loosely be called *quadratic Fourier analysis*, which can make precise the connection between parallelopipeds and correlations with quadratic (or pseudoquadratic) phases; there is also the beginnings of a more general theory connecting higher dimensional parallelopipeds and higher degree phases. This is still work in progress, but we have already been able to use the theory to understand several types of linear patterns already; for instance, Ben and I showed that the number of arithmetic progressions of length four consisting of primes less than a given large number N is equal to

$$(\frac{3}{4}\prod_{p\geq 5}(1-\frac{3p-1}{(p-1)^3}+o(1)))\frac{N^2}{\log^4 N} \approx 0.4764\frac{N^2}{\log^4 N}.$$

Very briefly, the role of additive combinatorics (and generalised Fourier analysis) is to replace problems of counting patterns involving multiple prime parameters, with that of counting correlations that involve a single prime parameter (e.g. computing a sum $\sum_{p<N} e^{2\pi i \alpha p^2}$ for various real numbers α), which is significantly easier (though not entirely trivial) and amenable to a current technology from analytic number theory. So we don't dispense with the number theory entirely, but thanks to combinatorics we can reduce the amount of difficult number theoretical work that we have to do to a feasible level.

3.3.2 Additive combinatorics and random matrices

In many areas of physics, chemistry, and computer science, one often has to study large matrices A, which for sake of discussion we shall take to be square, thus $A = (a_{ij})_{1 \leq i,j \leq n}$ is an $n \times n$ matrix for some large integer n, and the a_{ij} are real or complex numbers. In some cases A will be very structured (e.g. self-adjoint, upper triangular, sparse, circulant, low rank, etc.) but in other cases one expects A to be so complicated that there is no usable structure to exploit in the matrix. In such cases, one often makes the non-rigorous (but surprisingly accurate, in practice) assumption that A behaves like a random matrix, whose entries are drawn independently and identically from a single probability distribution (which could be continuous, discrete, or a combination of both). Each choice of distribution determines a different random matrix ensemble. Two particularly fundamental examples of continuous ensembles are:

1. The *real Gaussian random matrix ensemble*, in which each a_{ij} is distributed independently according to the standard normal distribution $N(0,1) \equiv \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,dx$. This ensemble has the remarkable feature of being invariant under the orthogonal group $O(n)$; if R is a rotation or reflection matrix in $O(n)$, and A is distributed as a real Gaussian random matrix, then RA and AR are also distributed as a real Gaussian random matrix. (This is ultimately because the product of n Gaussian measures $\frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,dx$ on $\mathbf{R}$ is a Gaussian measure $\frac{1}{(2\pi)^{n/2}} e^{-|x|^2/2}\,dx$ on $\mathbf{R}^n$, which is manifestly rotation-invariant.)

2. The *complex Gaussian random matrix ensemble*, in which the a_{ij} are distributed according to a complex normal distribution $\frac{1}{\pi} e^{-|z|^2}\,dz$. This ensemble has the feature of being invariant under the unitary group U(n).

Two particularly fundamental examples of discrete ensembles are

1. The *Bernoulli ensemble*, in which each a_{ij} is distributed independently and uniformly in the set $\{-1,+1\}$, thus A is a random matrix of signs.

2. The *lazy (or sparse) Bernoulli ensemble*, in which each a_{ij} is independently equal to -1 or $+1$ with probability $p/2$, and equal to 0 with probability $1-p$, for some fixed $0 \leq p \leq 1$, thus A is a sparse matrix of signs of expected density p.

The Bernoulli and sparse Bernoulli ensembles arise naturally in computer science and numerical analysis, as they form a simple model for simulating the effect of numerical roundoff error (or other types of digital error) on a large matrix. Continuous ensembles such as the Gaussian ensembles, in contrast, are natural models for matrices in the analog world, and in particular in physics and chemistry. [For reasons that are still somewhat mysterious, these ensembles, or more precisely their self-adjoint counterparts, also seem to be good models for various important statistics in number theory, such as the statistics of zeroes of the Riemann zeta function, but this is not the topic of my discussion here.] There are of course many other possible ensembles of interest that one could consider, but I will stick to the Gaussian and Bernoulli ensembles here for simplicity.

If A is drawn randomly from one of the above matrix ensembles, then we have a very explicit understanding of how each of the coefficients of the matrix A behaves. But in practice, we want to study more "global" properties of the matrix A which involve rather complicated interactions of all the coefficients together. For instance, we could be interested in the following (closely related) questions:

- *Dynamics*. Given a typical vector $x \in \mathbf{R}^n$, what happens to the iterates $Ax, A^2x, \ldots, A^mx$ in the limit $m \to \infty$?
- *Expansion and contraction*. Given a non-zero vector x, how does the norm $\|Ax\|$ of Ax compare with the norm $\|x\|$ of x? What is the largest ratio $\|Ax\|/\|x\|$? The smallest ratio? The average ratio?
- *Invertibility*. Is the equation $Ax = b$ solvable for every vector b? Do small fluctuations in b always cause small fluctuations in x, or can they cause large fluctuations? (In other words, is the invertibility problem stable?)

As any student of linear algebra knows, these questions can be answered satisfactorily if one knows the eigenvalues $\lambda_1, \ldots, \lambda_n \in \mathbf{C}$ (counting multiplicity) and singular values $\sigma_1 \geq \ldots \geq \sigma_n \geq 0$ of the matrix A. (As the matrix A is not self-adjoint, the eigenvalues can be complex-valued even if the coefficients of A are real-valued; however, the singular values are always non-negative reals, because AA^* is self-adjoint and positive semi-definite.) For instance:

- The largest eigenvalue magnitude $\sup_{1 \leq k \leq n} |\lambda_k|$ determines the maximal rate of exponential growth or decay of A^mx (ignoring polynomial growth corrections coming from repeated eigenvalues), while the smallest eigenvalue magnitude $\inf_{1 \leq k \leq n} |\lambda_k|$ determines the minimal rate of exponential growth.
- The ratio $\|Ax\|/\|x\|$ has a maximal value of σ_1, a minimal value of σ_n, and a root mean square value of $(\frac{1}{n}\sum_{k=1}^n \sigma_k^2)^{1/2}$ if the orientation of x is selected uniformly at random.
- The matrix A is invertible if and only if all eigenvalues are non-zero, or equivalently if σ_n is positive.
- The stability of the invertibility problem is controlled by the condition number σ_1/σ_n.

So, one of the fundamental problems in the theory of random matrices is to understand how the eigenvalues and singular values of a random matrix A are distributed. (More generally, it is of interest to study the eigenvalues and singular values of $A+B$, where A is drawn from a standard random matrix ensemble, and B is a fixed deterministic matrix, but for simplicity we will not discuss this case here.) But how does one get a handle on these numbers?

The direct approach of working with the characteristic equation $\det(A - \lambda I) = 0$ (or $\det(AA^* - \sigma^2 I) = 0$) looks very unpromising; one is asking to find the roots of a large degree polynomial, most of whose coefficients depend in a hopelessly complicated way on the coefficients on A.

In the special cases of the real and complex Gaussian ensembles, there is a massive amount of algebraic structure coming from the action of $O(n)$ and $U(n)$ that allows one to explicitly compute various multidimensional integrals, and this approach actually works! One gets a very explicit and useful explicit formula for the joint eigenvalue distribution (first worked out by Ginibre, I believe) this way. But for more general ensembles, such as the Bernoulli ensemble, such algebraic structure is not present, and so it is unlikely that any useful explicit formula for the joint eigenvalue distribution exists. However, one can still obtain a lot of useful information if, instead of trying to locate each eigenvalue or singular value directly, one instead tries to compute various special averages (e.g. moments) of these eigenvalues or singular values. For instance, from undergraduate linear algebra we have the fundamental formulae

$$\mathrm{tr}(A) = \sum_{k=1}^{n} \lambda_k$$

and

$$\det(A) = \prod_{k=1}^{n} \lambda_k$$

connecting the trace and determinant of a matrix A to its eigenvalues, and more generally

$$\mathrm{tr}(A^m) = \sum_{k=1}^{n} \lambda_k^m \tag{3.8}$$

$$\det(A - zI) = \prod_{k=1}^{n} (\lambda_k - z)$$

and similarly for the singular values, we have

$$\mathrm{tr}((AA^*)^m) = \sum_{k=1}^{n} \sigma_k^{2m}$$

$$\det(AA^* - zI) = \prod_{k=1}^{n} (\sigma_k^2 - z).$$

So, if one can easily compute traces and determinants of the matrix A (and various other matrices related to A), then one can in principle get quite a bit of control on the eigenvalues and singular values. It is also worth noting that the eigenvalues and singular values are related to each other in several ways; for instance, we have the identity

$$\prod_{k=1}^{n} |\lambda_k| = \prod_{k=1}^{n} \sigma_k \tag{3.9}$$

(which comes from comparing the determinants of A and AA^*), the inequality

$$\sigma_n \leq \sup_{1 \leq k \leq n} |\lambda_k| \leq \sigma_1 \tag{3.10}$$

(which comes from looking at the ratio $\|Ax\|/\|x\|$ when x is an eigenvector), and the inequality

$$\sum_{1 \leq k \leq n} |\lambda_k|^2 \leq \sum_{1 \leq k \leq n} \sigma_k^2 \tag{3.11}$$

(which is easiest to see by using QR (or KAN) decomposition of the eigenvector matrix to rotate the matrix A to be upper triangular, and then computing the trace of AA^*).

Let's give some simple examples of this approach. If we take A to be the Gaussian or Bernoulli ensemble, then the trace of A has expectation zero, and so we know that the sum $\sum_{k=1}^n \lambda_k$ has expectation zero also. (Actually, this is easy to see for symmetry reasons: A has the same distribution as $-A$, and so the distribution of eigenvalues also has a symmetry around the origin.) The trace of AA^*, by contrast, is the sum of the squares of all the matrix coefficients, and will be close to n^2 (for the Bernoulli ensemble, it is exactly n^2); thus we see that $\sum_{k=1}^n \sigma_k^2 \sim n^2$, and so by (3.11) we have $\sum_{k=1}^n |\lambda_k|^2 = O(n^2)$. So we see that the eigenvalues and singular values should be about $O(\sqrt{n})$ on the average. By working a little harder (e.g. by playing with very high moments of AA^*) one can show that the largest singular value is also going to be $O(\sqrt{n})$ with high probability, which then implies by (3.10) that all eigenvalues and singular values will be $O(\sqrt{n})$. Unfortunately, this approach does not seem to yield much information on the *least* singular value, which plays a major role in the invertibility and stability of A.

It is now natural to normalise the eigenvalues and singular values of A by $\frac{1}{\sqrt{n}}$, and consider the distribution of the set of normalised eigenvalues $\{\frac{1}{\sqrt{n}}\lambda_k : 1 \leq k \leq n\}$. If one plots these normalised eigenvalues numerically in the complex plane for moderately large n (e.g. $n = 100$), one sees a remarkable distribution; the eigenvalues appear to be uniformly distributed in the unit circle $D := \{z \in \mathbf{C} : |z| \leq 1\}$. (For small n, there is a little bit of a clustering on the real line, just because polynomials with real coefficients tend to have a couple of real zeroes, but this clustering goes away in the limit as n goes to infinity.) This phenomenon is known as the *circular law*; more precisely, if we let n tend to infinity, then for every sufficiently nice set R in the plane (e.g. one could take R to be a rectangle), one has

$$\lim_{n\to\infty} \frac{1}{n}\{1 \leq k \leq n : \frac{1}{\sqrt{n}}\lambda_k \in R\} \to \frac{1}{\pi}|R \cap D|.$$

(Technically, this formulation is known as the *strong circular law*; there is also a *weak circular law*, which asserts that one has convergence in probability rather than almost sure convergence. But for this lecture I will ignore these distinctions.)

The circular law was first proven in the case of the complex Gaussian ensemble by Mehta[Me1967], using an explicit formula for the joint distribution of the eigenvalues. But for more general ensembles, in which explicit formulae were not available, progress was more difficult. The method of moments (in which one uses (3.8) to compute the sums of powers of the eigenvalues) is not very useful because of the cancellations caused by the complex nature of the eigenvalues; indeed, one can show that $\mathrm{tr}((\frac{1}{\sqrt{n}}A)^m)$ is roughly zero for every m, which is consistent with the circular law but also does not preclude, for instance, all the eigenvalues clustering at the origin. [For random *self-adjoint* matrices, the moment method works quite well, leading for instance to Wigner's semi-circular law.]

The first breakthrough was by Girko[Gi1984], who observed that the eigenvalue distribution could be recovered from the quantities

$$\frac{1}{n}\log|\det(\frac{1}{\sqrt{n}}A - zI)| = \frac{1}{n}\sum_{k=1}^{n}\log|\frac{1}{\sqrt{n}}\lambda_k - z| \tag{3.12}$$

for complex z (this expression is closely related to the *Stieltjes transform* $\frac{1}{n}\sum_{k=1}^{n}\frac{1}{z-\lambda_k}$ of the normalised eigenvalue distribution of A, being an antiderivative of the real part of this transform). To compute this quantity, Girko then used the formula (3.9) to relate the determinant of $\frac{1}{\sqrt{n}}A - zI$ with the singular values of this matrix. The singular value distribution could then be computed by the moment method (note that singular values, unlike eigenvalues, are real and non-negative, and so we do not have cancellation problems). Putting this all together and doing a large number of algebraic computations, one eventually obtains (formally, at least) a proof of the circular law.

There was however a technical difficulty with the above analysis, which was that the formula (3.9) becomes very unstable when the least singular value is close to zero (basically because of a division by zero problem). This is not merely a technical issue but is fundamental to the general problem of controlling eigenvalues of non-self-adjoint matrices $\frac{1}{\sqrt{n}}A$: these eigenvalues can become very unstable near a region of *pseudospectrum*, which can be defined as a complex number z such that the least singular value of $\frac{1}{\sqrt{n}}A - zI$ is small. The classic demonstration of this comes from the perturbed shift matrices

$$A_\varepsilon := \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & & & \ddots & \\ 0 & 0 & 0 & \dots & 1 \\ \varepsilon & 0 & 0 & \dots & 0 \end{pmatrix}.$$

For sake of discussion let us take n to be even. When $\varepsilon = 0$, this matrix is singular, with least singular value $\sigma_1 = 0$ and with all n generalised eigenvalues equal to 0. But when ε becomes positive, the least singular value creeps up to ε, but the n eigenvalues move rapidly away from the origin, becoming $\varepsilon^{1/n}e^{2\pi ij/n}$ for $j = 1,\dots,n$. This is ultimately because the zero set of the characteristic polynomial $z^n - \varepsilon$ is very sensitive to the value of ε when that parameter is close to zero.

So, in order to make the circular law argument complete, one needs to get good lower bounds on the least singular value of the random matrix A (as well as variants of this matrix, such as $\frac{1}{\sqrt{n}}A - zI$). In the case of continuous (non-Gaussian) ensembles, this was first done by Bai [Ba1997]. To illustrate the basic idea, let us look at a toy problem, to show that the least singular value of a real Gaussian matrix A is usually non-zero (i.e. A is invertible with high probability). For this, we use some linear algebra. Let $X_1,\dots,X_n$ denote the rows of A, which we can view as vectors in $\mathbf{R}^n$. Then the least singular value of A vanishes precisely when $X_1,\dots,X_n$ lie on a hyperplane. This implies that one of the vectors here is a linear combination of the other $n-1$; by symmetry, we conclude that the probability that the least singular value vanishes is

bounded by n times the probability that X_n (say) is a linear combination of $X_1,\ldots,X_{n-1}$. But $X_1,\ldots,X_n$ span a hyperplane at best; and X_n has a continuous distribution and so has only a zero probability of lying in the hyperplane. Thus the least singular value vanishes with probability zero. It turns out that this argument is robust enough to also show that the least singular value not only avoids zero, but in fact keeps a certain distance away from it; for instance, it is not hard to show with this method that the least singular value is at least $1/n^2$ (say) with probability $1-o(1)$, basically because each row vector is not only not a linear combination of the other rows, but in fact keeps a certain distance away from the space spanned by the other rows, with high probability.

For discrete random matrices, one runs into a new difficulty: a row vector such as X_n is no longer continuously distributed, and so can in fact concentrate on a hyperplane with positive probability. For instance, in the Bernoulli case, a row vector X is just a random corner of the discrete cube $\{-1,+1\}^n$. There are certain hyperplanes which X can visit quite frequently; for instance, X will have a probability $1/2$ of lying in the hyperplane $\{(x_1,\ldots,x_n) : x_1 = x_2\}$. In particular, there is a probability $1/2^n$ that all rows $X_1,\ldots,X_n$ lie in this hyperplane, which would cause the Bernoulli matrix A to be non-invertible. (In terms of A itself, what is going on is that there is a $1/2^n$ chance that the first column and second column coincide, which will of course destroy invertibility.) In particular, A now has a non-zero chance of being singular. [It is in fact conjectured that the singularity probability is close to this value, or more precisely equal to $(1/2+o(1))^n$. The best known upper bound currently for this probability is $(1/\sqrt{2}+o(1))^n$, due to Bourgain, Vu, and Wood.]

One can hope to sum the singularity probability over all hyperplanes, but there are of course infinitely many hyperplanes in $\mathbf{R}^n$. Fortunately, only a few of these will have a particularly high chance of capturing X. The problem now hinges on getting a sufficiently strong control on the number of hyperplanes which are "rich" in the sense that they have a high probability of capturing X (or equivalently, that they have a large intersection with the discrete cube $\{-1,+1\}^n$).

This is where (finally!) the additive combinatorics comes in. Let $v = (v_1,\ldots,v_n) \in \mathbf{R}^n$ be a normal vector to a given hyperplane V. (For instance, if $V = \{(x_1,\ldots,x_n) : x_1 = x_2\}$, one could take $v = (1,-1,0,\ldots,0)$.) Then V is rich if and only if the random variable

$$\varepsilon_1 v_1 + \ldots + \varepsilon_n v_n \tag{3.13}$$

vanishes a large proportion of the time, where $\varepsilon_1,\ldots,\varepsilon_n \in \{-1,+1\}$ are independent signs. (One can view (3.13) as the result of an n-step random walk, in which the j^{th} step of the walk has magnitude $|v_i|$.) To put it another way, if $(v_1,\ldots,v_n)$ is associated to a rich hyperplane, then there are many additive relations amongst the coefficients $v_1,\ldots,v_n$. What kinds of sets of numbers have such a strong amount of structure? (This problem is known as the *inverse Littlewood-Offord problem*; the forward Littlewood-Offord problem concerns how often (3.13) vanished for a given set of numbers $v_1,\ldots,v_n$.)

Well, one way that many of the sums (3.13) could vanish is if many of the v_i are themselves zero; for instance, we have already seen that if $(v_1,\ldots,v_n) = (1,-1,0,\ldots,0)$, then half of the sums (3.13) vanish. But this is a rather degenerate case, and it is intuitively obvious that the more non-zero terms one has in the random walk, the less likely

it is that the sum is going to vanish. Indeed, there is an observation of Erdős[Er1945] (a quick application of Sperner's theorem) that if k of the coefficients $v_1,\ldots,v_n$ are non-zero, then (3.13) can only vanish at most $O(1/\sqrt{k})$ of the time. This bound is sharp; if, for instance, $v_1 = \ldots = v_k = 1$ and $v_{k+1} = \ldots = v_n = 0$, then the theory of random walks tells us that (3.13) is distributed in a roughly Gaussian and discrete fashion around the origin with standard deviation $\sqrt{k}$, and so (3.13) should vanish about $O(1/\sqrt{k})$ of the time (and one can check this easily enough, at least when k is even, using Stirling's formula). [In fact, this is the exact optimum, as follows from Erdős' argument.]

So, now suppose that all the v_i are non-zero. Then Erdős' result tells us that (3.13) vanishes at most $O(1/\sqrt{n})$ of the time. But in most cases one can do a lot better; if, for instance, the v_i are linearly independent over the rationals, then (3.13) in fact never vanishes at all. It turns out that by intensively using the tools from additive combinatorics (including Fourier analysis and the geometry of numbers) one can obtain a satisfactory classification of the vectors $v = (v_1,\ldots,v_n)$ for which (3.13) has a high chance of vanishing; the precise description is technical, but basically in order for (3.13) to equal zero often, most of the coordinates of v must lie inside an generalised arithmetic progression of reasonably small size and dimension. Using such facts, it is possible to get good bounds on the singularity probability and on the least singular value of random discrete matrices such as Bernoulli matrices, leading in particular to the circular law for such discrete matrices (various formulations of this law have been recently obtained in [GoTi2008], [GoTi2008b], [PaZh2008], [TaVu2008]). [To get the best bounds on the singularity probability, one uses a slightly different argument, using Fourier analysis and additive combinatorics to compare the vanishing probability of a random walk with that of a lazy random walk, thus creating a relationship between the singularity of Bernoulli matrices and sparse Bernoulli matrices; see [TaVu2007] for details.]

This theory for understanding singularity behaviour of discrete random matrices promises to have applications to some other areas of mathematics as well. For instance, the subject of *smoothed analysis*, in which the introduction of random noise to various numerical algorithms (such as the simplex method) increases the stability of the algorithm, can use this theory to extend the theoretical results in that subject from continuous noise models to discrete noise models (such as those created by roundoff error).

3.3.3 Sum-product estimates, expanders, and exponential sums

This is my final Milliman lecture, in which I talk about the sum-product phenomenon in arithmetic combinatorics, and some selected recent applications of this phenomenon to uniform distribution of exponentials, expander graphs, randomness extractors, and detecting (or *sieving*) almost primes in group orbits, particularly as developed by Bourgain and his co-authors.

In the previous two lectures we had concentrated on *additive combinatorics* - the study of additive operations and patterns on sets; this can be viewed as a combinatorial analogue of abelian group theory. Now we will look at *arithmetic combinatorics* - the simultaneous study of additive *and* multiplicative operations on sets; this is sort of a combinatorial analogue of commutative algebra.

There are many questions to study here, but the most basic is the *sum-product problem*, which we can state as follows. Let A be a finite non-empty set of elements of a ring R (e.g. finite sets of integers, or elements of a cyclic group $\mathbf{Z}/q\mathbf{Z}$, or sets of matrices over some ring). Then we can form the *sum set*

$$A+A := \{a+b : a,b \in A\}$$

and the *product set*

$$A \cdot A := \{a \cdot b : a,b \in A\}$$

To avoid degeneracies, let us assume that none (or very few) of the elements in A are zero divisors (as this may cause $A \cdot A$ to become very small). Then it is easy to see that $A+A$ and $A \cdot A$ will be at least as large as A itself.

Typically, both of these sets will be much larger than A itself; indeed, if we select A at random, we generically expect $A+A$ and $A \cdot A$ to have cardinality comparable to $|A|^2$. But when A enjoys additive or multiplicative structure, the sets $A+A$ or $A \cdot A$ can be of size comparable to A. For instance, if A is an arithmetic progression $\{a, a+r, a+2r, \ldots, a+(k-1)r\}$ or an additive subgroup in the ring R (modulo zero divisors, such as 0), then $|A+A| \sim |A|$. Similarly, if A is a geometric progression $\{a, ar, ar^2, \ldots, ar^{k-1}\}$ or a multiplicative subgroup in the ring R, then $|A \cdot A| \sim |A|$. And of course, if A is both an additive and a multiplicative subgroup of R (modulo zero divisors), i.e. if A is a subring of R, then $|A+A|$ and $|A \cdot A|$ are both comparable in size to $|A|$. These examples are robust with respect to small perturbations; for instance, if A is a dense subset of an arithmetic progression or additive subgroup, then it is still the case that $A+A$ is comparable in size to A. There are also slightly more complicated examples of interest, such as *generalised arithmetic progressions*, but we will not discuss these here.

Now let us work in the ring of integers $\mathbf{Z}$. This ring has no non-trivial finite additive subgroups or multiplicative subgroups (and it certainly has no non-trivial finite subrings), but it of course has plenty of arithmetic progressions and geometric progressions. But observe that it is rather difficult for a finite set A of integers to resemble both an arithmetic progression and a geometric progression simultaneously (unless A is very small). So one expects at least one of $A+A$ and $A \cdot A$ to be significantly larger than A itself. This claim was made precise Erdős and Szemerédi[ErSz1983], who showed that

$$\max(|A+A|, |A \cdot A|) \gg |A|^{1+\varepsilon} \tag{3.14}$$

for some absolute constant $\varepsilon > 0$. The value of this constant as improved steadily over the years; the best result currently is due to Solymosi[So2005], who showed that one can take ε arbitrarily close to $3/11$. Erdős and Szemerédi in fact conjectured that one can take ε arbitrarily close to 1 (i.e. for any finite set of integers A, either the sum set or product set has to be very close to its maximal size of $|A|^2$), but this conjecture seems out of reach at present. Nevertheless, even just the epsilon improvement over the trivial bound of $|A|$ is already quite useful. It is the first example of what is now called the *sum-product phenomenon*: if a finite set A is not close to an actual subring, then either the sum set $A+A$ or the product set $A \cdot A$ must be significantly larger than A itself. One

can view (3.7) as a "robust" version of the assertion that the integers contain no non-trivial finite subrings; (3.7) is asserting that in fact the integers contain no non-trivial finite sets which even come close to behaving like a subring.

In 1999, Tom Wolff (personal communication) posed the question of whether the sum-product phenomenon held true in finite fields $\mathbf{F}_p$ of prime order (note that such fields have no non-trivial subrings), and in particular whether (3.14) was true when $A \subset \mathbf{F}_p$, and A was not close to being all of $\mathbf{F}_p$, in the sense that $|A| \leq p^{1-\delta}$ for some $\delta > 0$; of course one would need ε to depend on δ. (Actually, Tom only posed the question for $|A| \sim p^{1/2}$, being motivated by finite field analogues of the Kakeya problem[Wo1999], but the question was clearly of interest for other ranges of A as well.) This question was solved in the affirmative by Bourgain, Katz, and myself[BoKaTa2004] (in the range $p^\delta \leq |A| \leq p^{1-\delta}$) and then by Bourgain, Glibichuk, and Konyagin[BoGlKo2006] (in the full range $1 \leq |A| \leq p^{1-\delta}$); the result is now known as the *sum-product theorem* for $\mathbf{F}_p$ (and there have since been several further proofs and refinements of this theorem). The fact that the field has prime order is key; if for instance we were working in a field of order p^2, then by taking A to be the subfield of order p we see that both $A+A$ and $A \cdot A$ have exactly the same size as A. So any proof of the sum-product theorem must use at some point the fact that the field has prime order.

As in the integers, one can view the sum-product theorem as a robust assertion of the obvious statement that the field $\mathbf{F}_p$ contains no non-trivial subrings. So the main difficulty in the proof is to find a proof of this latter fact which is robust enough to generalise to this combinatorial setting. The standard way to classify subrings is to use Lagrange's theorem that the order of a subgroup divides the order of the whole group, which is proven by partitioning the whole group into cosets of the subgroup, but this argument is very unstable and does not extend to the combinatorial setting. But there are other ways to proceed. The argument of Bourgain, Katz, and myself (which is based on an earlier argument of Edgar and Miller[EdMi2003]), roughly speaking, proceeds by investigating the "dimension" of $\mathbf{F}_p$ relative to A, or in other words the least number of elements $v_1, \ldots, v_d$ in $\mathbf{F}_p$ such that every element of $\mathbf{F}$ can be expressed in the form $a_1 v_1 + \ldots + a_d v_d$. Note that the number of such representations is equal to $|A|^d$. The key observation is that as $|\mathbf{F}_p|$ is prime, it cannot equal $|A|^d$ if $d > 1$, and so by the pigeonhole principle some element must have more than one representation. One can use this "linear dependence" to reduce the dimension by 1 (assuming that A behaves a lot like a subring), and so can eventually reduce to the $d = 1$ case, which is prohibited by our assumption $A < p^{1-\delta}$. (The hypothesis $|A| > p^\delta$ is needed to ensure that the initial dimension d is bounded, so that the iteration only requires a bounded number of steps.) The argument of Bourgain, Glibichuk, and Konyagin uses a more algebraic method (a variant of the polynomial method of Stepanov[St1969]), using the basic observation that the number of zeroes of a polynomial (counting multiplicity) is bounded by the degree of that polynomial to obtain upper bounds for various sets (such as the number of parallelograms in A). More recently, a short argument of Garaev[Ga2008] proceeds using the simple observation that if A is any non-trivial subset of $\mathbf{F}_p$, then there must exist $a \in A$ such that $a+1 \notin A$; applying this to the "fraction field" $Q[A] := \{(a-b)/(c-d) : a,b,c,d \in A, c \neq d\}$ of A one can conclude that $Q[A]$ does not in fact behave like a field, and hence A does not behave like a ring.

The sum-product phenomenon implies that if a set $A \subset \mathbf{F}_p$ of medium size $p^\delta \leq$

$|A| \leq p^{1-\delta}$ is multiplicatively structured (e.g. it is a geometric progression or a multiplicative subgroup) then it cannot be additively structured: $A+A$ is significantly larger than A. It turns out that with a little bit of extra work, this observation can be iterated: $A+A+A+A$ is even larger than A, and so on and so forth, and in fact one can show that $kA = \mathbf{F}_p$ for some bounded k depending only on δ, where $kA := A + \ldots + A$ is the k-fold sumset of A. (The key to this iteration essentially lies in the inclusion $(kA) \cdot (kA) \subset k^2(A^2)$, which is a consequence of the distributive law. The use of this law unfortunately breaks the symmetry between multiplication and addition that one sees in the sum-product estimates.) Thus any multiplicatively structured subset A of $\mathbf{F}_p$ of medium size must eventually additively generate the whole field. As a consequence of this, one can show that A is an "additive expander", which roughly speaking means that $A+B$ is spread out on a significantly larger set than B for any medium-sized set B. (In more probabilistic language, if one considered the random walk whose steps were drawn randomly from A, then this walk would converge extremely rapidly to the uniform distribution.) From that observation (and some more combinatorial effort), one can in fact conclude that multiplicatively structured sets must be distributed uniformly in an additive sense; if they concentrated too much in, say, a subinterval of $\mathbf{F}_p$, then this could be used to contradict the additive expansion property.

Let me note one cute application of this technology, due to Bourgain[Bo2005], to the Diffie-Hellman key exchange protocol[DiHe1976] and its relatives in cryptography. Suppose we have two people, Alice and Bob, who want to communicate privately and securely, but have never met each other before and can only contact each other via an unsecured network (e.g. the internet, or physical mail), in which anyone can eavesdrop. How can Alice and Bob achieve this?

If one was sending a physical object (e.g. a physical letter) by physical mail (which could be opened by third parties), one could proceed as follows.

1. Alice places the object in a box, and locks the box with her own padlock, keeping the key. She then mails the locked box to Bob. Anyone who intercepts the box cannot open it, since they don't have Alice's key.

2. Of course, Bob can't open the box either. But what he can do instead is put his own padlock on the box (keeping the key), and sends the doubly locked box back to Alice.

3. Alice can't unlock Bob's padlock... but she can unlock her own. So she removes her lock, and sends the singly locked box back to Bob.

4. Bob can unlock his own padlock, and so retreives the object safely. At no point was the object available to any interceptor.

A similar procedure (a slight variant of the Diffie-Hellman protocol, essentially the Massey-Omura cryptosystem) can be used to transmit a digital message g (which one should think of as just being a number) from Alice to Bob over an unsecured network, as follows:

1. Alice and Bob agree (over the unsecured network) on some large prime p (larger than the maximum size of the message g).

2. Alice "locks" the message g by raising it to a power a mod p, where Alice generates the "key" a randomly and keeps it secret. She then sends the locked message g^a mod p to Bob.

3. Bob can't decode this message (he doesn't know a), but he doubly locks the message by raising the message to his own power b, and returns the doubly locked message g^{ab} mod p back to Alice.

4. Alice then "unlocks" her part of the message by taking the a^{th} root (which can be done by exploiting Cauchy's theorem) and sends g^b mod p back to Bob.

5. Bob then takes the b^{th} root of the message and recovers g.

An eavesdropper (let's call her Eve) could intercept p, as well as the three "locked" values g^a, g^b, g^{ab} mod p, but does not directly recover g. Now, it is possible that one could use this information to reconstruct g (indeed, if one could quickly take discrete logarithms, then this would be a fairly easy task) but no feasible algorithm for this is known (if p is large, e.g. 500+ digits); the problem is generally believed to be roughly comparable in difficulty to that of factoring large numbers. But no-one knows how to rigorously prove that the Diffie-Hellman reconstruction problem is hard (e.g. non-polynomial time); indeed, this would imply $P \neq NP$, since this reconstruction problem is easily seen to be in NP (though it is not believed to be NP-complete).

Using the sum-product technology, Bourgain was at least able to show that the Diffie-Hellman protocol was secure (for sufficiently large p) if Eve was only able to see the high bits of g^a, g^b, g^{ab} mod p, thus pinning down g^a, g^b, g^{ab} to intervals. The reason for this is that the set $\{(g^a, g^b, g^{ab}) \in \mathbf{F}_p^3 : a, b \in \mathbf{Z}\}$ has a lot of multiplicative structure (indeed, it is a multiplicative subgroup of the ring $\mathbf{F}_p^3$) and so should be uniformly distributed in an additive sense (by adapting the above sum-product technology to $\mathbf{F}_p^3$).

Another application of sum-product technology was to build efficient randomness extractors - deterministic algorithms that can create high-quality (very uniform) random bits from several independent low-quality (non-uniform) random sources; such extractors are of importance in computer science and cryptography. Basically, the sum-product estimate implies that if $A, B, C \subset \mathbf{F}_p$ are sets of medium size, then the set $A + B \cdot C$ is significantly larger than A, B, or C. As a consequence, if X, Y, Z are independent random variables in $\mathbf{F}_p$ which are not too narrowly distributed(in particular, they are not deterministic, and thus distributed only on a single value), one can show (with the assistance of some additive combinatorics) that the random variable $X + YZ$ is significantly more uniformly distributed than X, Y, or Z. Iterating this leads to some surprisingly good randomness extractors, as was first observed by Barak, Impagliazzo, and Wigderson[BaImWi2006].

Another application of the above sum-product technology was to get a product estimate in matrix groups, such as $SL_2(\mathbf{F}_p)$. Indeed, Helfgott[He2008] was able to show that if A was a subset of $SL_2(\mathbf{F}_p)$ of medium or small size, and it was not trapped inside a proper subgroup of $SL_2(\mathbf{F}_p)$, then $A \cdot A \cdot A$ was significantly larger than A itself. (One needs to work with triple products here instead of double products for a rather trivial reason: if A was the union of a subgroup and some external element, then $A \cdot A$ is still comparable in size to A, but $A \cdot A \cdot A$ will be much larger. This result may not

immediately look like a sum-product estimate, because there is no obvious addition, but it is concealed within the matrix multiplication law for $SL_2(\mathbf{F}_p)$. The key observation in Helfgott's argument, which relies crucially on the sum-product estimate, is that if V is a collection of diagonal matrices in $SL_2(\mathbf{F}_p)$ of medium size, and g is a non-diagonal matrix element, then the set $\mathrm{tr}(VgVg^{-1})$ is significantly larger than V itself. If one works out explicitly what this trace is, one sees a sum-product type of result emerging. Conversely, if the trace $\mathrm{tr}(A)$ of a group-like set A is large, then the conjugacy classes in A are fairly small (since trace is conjugation-invariant), which forces many pairs in A to commute, which creates large sets V of simultaneously commuting (and hence simultaneously diagonalisable) elements, due to the fact that if two elements in SL_2 commute with a third, then they are quite likely to commute with each other. The tension between these two implications is what underlies Helfgott's results.

The estimate of Helfgott shows that multiplication by medium-size sets in $SL_2(\mathbf{F}_p)$ expands rapidly across the group (unless it is trapped in a subgroup). As a consequence of Helfgott's estimate, Bourgain and Gamburd[BoGa2006] were able to show that if S was any finite symmetric set of matrices in $SL_2(\mathbf{Z})$ which generated a sufficiently large (or more precisely, Zariski dense) subgroup of $SL_2(\mathbf{Z})$, and S_p was the projection of S to $SL_2(\mathbf{Z}_p)$, then the random walk using S_p on $SL_2(\mathbf{Z}_p)$ was very rapidly mixing, so that after about $O(\log p)$ steps, the walk was very close to uniform. (The precise statement was that the Cayley graph associated to S_p for each p formed an expander family.) Quite recently, Bourgain, Gamburd, and Sarnak[BoGaSa2006] have applied these results (and generalisations thereof) to the problem of detecting (or sieving) almost primes in thin algebraically generated sets. To motivate the problem, we observe that many classical questions in prime number theory can be rephrased as one of detecting prime points $(p_1,\ldots,p_d) \in \mathscr{P}^d$ in algebraic subsets $\mathscr{O}$ of a lattice $\mathbf{Z}^d$. For instance, the twin prime problem asks whether the line $\mathscr{O} = \{(n,n+2) \in \mathbf{Z}^2\}$ contains infinitely many prime points. In general, these problems are very difficult, especially once one considers sets described by polynomials rather than linear functions; even the one-dimensional problem of determining whether the set $\mathscr{O} = \{n^2+1 : n \in \mathbf{Z}\}$ contains infinitely many primes has been open for quite a long time (though it is worth mentioning the celebrated result of Friedlander and Iwaniec[FrIw1998] that the somewhat larger set $\mathscr{O} = \{n^2+m^4 : n,m \in \mathbf{Z}\}$ is known to have infinitely many primes).

So prime points are hard to detect. However, by using methods from sieve theory, one can often detect *almost prime* points in various sets $\mathscr{O}$ - points whose coordinates are the products of only a few primes. For instance, a famous theorem of Chen[Ch1973] shows that the line $\mathscr{O} = \{(n,n+2) \in \mathbf{Z}^2\}$ contains infinitely many points which are almost prime in the sense that the first coordinate is prime, and the second coordinate is the product of at most two primes. The basic idea of sieve theory is to sift out primes and almost primes by removing all points whose coordinates are divisible by small factors (and then, due to various generalisations of the inclusion-exclusion principle, one has to add back in points which are divisible by multiple small factors, and so forth). See Section 1.10 for further discussion. In order for sieve theory to work well, one needs to be able to accurately count the size of the original set $\mathscr{O}$ (or more precisely, the size of this set restricted to a ball or a similar object), and also need to count how many points in that set have a certain residue class modulo q, for various values of q. (For instance, to sieve out twin primes or twin almost primes in the interval $\{1,\ldots,N\}$,

one needs to count how many elements n in that interval are such that n and $n+2$ are both invertible modulo q (i.e. coprime to q) for various values of q.)

For arbitrary algebraic sets $\mathcal{O}$, these tasks are very difficult. For instance, even the basic task of determining whether a set $\mathcal{O}$ described by several polynomials is non-empty is essentially Hilbert's tenth problem, which is undecidable in general. But if the set $\mathcal{O}$ is generated by a group Λ acting on $\mathbf{Z}^d$ (in some polynomial fashion), thus $\mathcal{O} = \Lambda b$ for some point $b \in \mathbf{Z}^d$, then the problems become much more tractable. If the group Λ is generated by some finite set S, and we restrict attention to group elements with some given word length, the problem of understanding how $\mathcal{O}$ is distributed modulo q is equivalent to asking how random walks on S of a given length distribute themselves on $(\mathbf{Z}/q\mathbf{Z})^d$. This latter problem is very close to the problem solved by the mixing results of Bourgain and Gamburd mentioned earlier, which is where the link to sum-product estimates arises from. Indeed, Bourgain, Gamburd, and Sarnak have now shown that rather general classes of algebraic sets generated by subgroups of $SL_2(\mathbf{Z})$ will contain infinitely many almost primes, as long as there are no obvious algebraic obstructions; the methods should hopefully extend to more general groups, such as subgroups of $SL_n(\mathbf{Z})$.

3.3.4 Notes

These articles were originally posted on Dec 4-6, 2007 at

terrytao.wordpress.com/2007/12/04

terrytao.wordpress.com/2007/12/05

terrytao.wordpress.com/2007/12/06

Thanks to intoverflow, Harald Helfgott, MK, Mark Meckes, ninguem, and Tom Smith for corrections and references.

Harald Helfgott remarked that perhaps the right framework for sum-product estimates was that of abelian groups G acting on other abelian groups A (thus A is a $\mathbf{Z}[G]$-module); given any subsets G', A' of G and A respectively that obey various non-degeneracy conditions, one should be able to take a bounded number of combinations of G' and A' to generate either about $|G'||A'|$ elements of A, or else to generate the entire submodule $\langle\langle G'\rangle\langle A'\rangle\rangle$.

Helfgott also remarked that the fact that two elements in SL_2 that commute with a third are likely to commute with another also holds in SL_n, and more generally in any semisimple group of Lie type, since a generic element of a semisimple Lie group is regular semisimple. Emmanuel Kowalski pointed out that this latter result is explicitly stated in [St1965].

Bibliography

[AgWo1998] S. Agnihotri, C. Woodward, *Eigenvalues of products of unitary matrices and quantum Schubert calculus*, Math. Res. Lett. **5** (1998), no. 6, 817–836.

[AjChNeSz1982] M. Ajtai, V. Chvátal, M. Newborn, and E. Szemerédi, *Crossing-free subgraphs*, Annals of Discrete Mathematics **12** (1982), 9–12.

[AlFiKrSz2000] N. Alon, E. Fischer, M. Krivelevich, M. Szegedy, *Efficient testing of large graphs*, Combinatorica **20** (2000), no. 4, 451–476.

[AlSh2005] N. Alon, A. Shapira, *Every monotone graph property is testable*, STOC'05: Proceedings of the 37th Annual ACM Symposium on Theory of Computing, 128–137, ACM, New York, 2005.

[Am1973] W. Amrein, V. Georgescu, *On the characterization of bound states and scattering states in quantum mechanics*, Helv. Phys. Acta **46** (1973/74), 635–658.

[Ar2006] E. Arrondo, *Another elementary proof of the Nullstellensatz*, American Mathematical Monthly, **113** (2) (2006), 169–171.

[Ar1929] E. Artin, *Ein Mechanisches System mit quasiergodischen Bahnen*, Abh. Math. Sem. Univ. Hamburg **3** (1924), 170–175.

[At1982] M. Atiyah, *Convexity and commuting Hamiltonians*, Bull. London Math. Soc. **14** (1982), no. 1, 1–15.

[AvGeTo2008] J. Avigad, P. Gerhardy, H. Towsner, *Local stability of ergodic averages*, preprint.

[Ba1973] G. Bachelis, *On the upper and lower majorant properties in $L^p(G)$*, Quart. J. Math. Oxford Ser. **24** (1973), 119–128.

[Ba1997] Z. D. Bai, Circular law, *Ann. Probab.* **25** (1997), 494–529.

[BaImWi2006] B. Barak, R. Impagliazzo, A. Wigderson, *Extracting randomness using few independent sources*, SIAM J. Comput. **36** (2006), no. 4, 1095–1118.

[BaDadVWa2008] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, *A Simple Proof of the Restricted Isometry Property for Random Matrices (aka "The Johnson-Lindenstrauss Lemma Meets Compressed Sensing")*, preprint.

[BaKe2008] R. Baraniuk, K. Kelly, et al., *Compressive Imaging: A New Single Pixel Camera*, at `www.dsp.ece.rice.edu/cs/cscamera`

[BaKe2008b] R. Baraniuk, K. Kelly, et al., *Compressive sensing resources*, at `www.dsp.ece.rice.edu/cs`

[BeKaMa1984] J.T. Beale, T. Kato, A. Majda, *Remarks on the breakdown of smooth solutions for the* 3*-D Euler equations*, Comm. Math. Phys. **94** (1984), no. 1, 61–66.

[Be2007] M. Beceanu, *A Centre-Stable Manifold for the Focussing Cubic NLS in* R^{1+3}, preprint.

[Be1975] W. Beckner, *Inequalities in Fourier analysis*, Ann. of Math. **102** (1975), no. 1, 159–182.

[Be1946] F. A. Behrend, *On sets of integers which contain no three terms in arithmetic progression*, Proc. Nat. Acad. Sci., **32** (1946), 331–332.

[Be2001] P. Belkale, *Local systems on* $\mathbf{P}^1 - S$ *for S a finite set*, Compositio Math. **129** (2001), no. 1, 67–86.

[Be2006] P. Belkale, *Geometric proofs of Horn and saturation conjectures*, J. Algebraic Geom. **15** (2006), no. 1, 133–173.

[Be2008] P. Belkale, *Quantum generalization of the Horn conjecture*, J. Amer. Math. Soc. **21** (2008), no. 2, 365–408.

[BeLi1980] H. Berestycki, P.-L. Lions, *Existence of a ground state in nonlinear equations of the Klein-Gordon type*, Variational inequalities and complementarity problems (Proc. Internat. School, Erice, 1978), pp. 35–51, Wiley, Chichester, 1980.

[BeZe1992] A. Berenstein, A. Zelevinsky, *Triple multiplicities for* $\mathrm{sl}(r+1)$ *and the spectrum of the exterior algebra of the adjoint representation*, J. Algebraic Combin. **1** (1992), no. 1, 7–22.

[Be2003] V. Bergelson, *Minimal idempotents and ergodic Ramsey theory*, Topics in Dynamics and Ergodic Theory 8-39, London Math. Soc. Lecture Note Series 310, Cambridge Univ. Press, Cambridge, 2003.

[BlPo2002] V. Blondel, N. Portier, *The presence of a zero in an integer linear recurrent sequence is NP-hard to decide*, Lin. Alg. Appl. **351-352** (2002), 91–98.

[BoTh1995] B. Bollobás, A. Thomason, *Projections of bodies and hereditary properties of hypergraphs*, Bull. London Math. Soc. **27** (1995), no. 5, 417–424.

[Bo1977] E. Bombieri, *The asymptotic sieve*, Rend. Accad. Naz. XL (5) 1/2 (1975/76), 243–269 (1977).

[Bo1986] J. Bourgain, *A Szemerédi type theorem for sets of positive density in* R^k, Israel J. Math. **54** (1986), no. 3, 307–316.

[Bo1989] J. Bourgain, *Bounded orthogonal systems and the $\Lambda(p)$-set problem*, Acta Math. **162** (1989), no. 3-4, 227–245.

[Bo1990] J. Bourgain, *Double recurrence and almost sure convergence*, J. Reine Angew. Math. **404** (1990), 140–161.

[Bo1999] J. Bourgain, *On triples in arithmetic progression*, Geom. Func. Anal. **9** (1999), 968–984.

[Bo1999b] J. Bourgain, *Global well-posedness of defocusing 3D critical NLS in the radial case,* J. Amer. Math. Soc. **12** (1999), 145–171.

[Bo2003] J. Bourgain, *On the Erdős-Volkmann and Katz-Tao ring conjectures*, Geom. Funct. Anal. **13** (2003), no. 2, 334–365.

[Bo2005] J. Bourgain, *Estimates on exponential sums related to the Diffie-Hellman distributions*, Geom. Funct. Anal. **15** (2005), no. 1, 1–34.

[Bo2008] J. Bourgain, *Roth's theorem on arithmetic progressions revisited*, preprint.

[BoGa2006] J. Bourgain, A. Gamburd, *New results on expanders*, C. R. Math. Acad. Sci. Paris **342** (2006), no. 10, 717–721

[BoGaSa2006] J. Bourgain, A. Gamburd, P. Sarnak, *Sieving and expanders*, C. R. Math. Acad. Sci. Paris 343 (2006), no. 3, 155–159.

[BoGlKo2006] J. Bourgain, A. Glibichuk, S. Konyagin, *Estimates for the number of sums and products and for exponential sums in fields of prime order*, J. London Math. Soc. (2) 73 (2006), no. 2, 380–398.

[BoKaTa2004] J. Bourgain, N. Katz, T. Tao, *A sum-product estimate in finite fields, and applications*, Geom. Funct. Anal. **14** (2004), no. 1, 27–57.

[BoMi1987] J. Bourgain, V. D. Milman, *New volume ratio properties for convex symmetric bodies in R^n*, Invent. Math. **88** (1987), no. 2, 319–340.

[Bu1974] L. Bunimovič, *The ergodic properties of certain billiards*, (Russian) Funkcional. Anal. i Priložen. **8** (1974), no. 3, 73–74.

[BuZw2004] N. Burq, M. Zworski, *Control theory and high energy eigenfunctions*, Forges Les Eaux proceedings, 2004.

[BuZw2005] N. Burq, M. Zworski, *Bouncing ball modes and quantum chaos*, SIAM Rev. **47** (2005), no. 1, 43–49.

[Ca2006] E. Candés, *Compressive sampling*, Proceedings of the International Congress of Mathematicians, Madrid, Spain, 2006.

[CaRoTa2006] E. Candés, J. Romberg, T. Tao, *Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information*, IEEE Inf. Theory **52** (2006), 489–509.

[CaRoTa2006b] E. Candés, J. Romberg, T. Tao, *Stable Signal Recovery from Incomplete and Inaccurate Measurements*, Comm. Pure Appl. Math. 59 (2006), 1207–1223.

[CaRuTaVe2005] E. Candés, M. Rudelson, T. Tao, R. Vershynin, *Error Correction via Linear Programming*, Proc. 46th Annual IEEE Symposium on Foundations of Computer Science (FOCS05), IEEE, 2005. pp. 295–308.

[CaTa2005] E. Candés, T. Tao, *Decoding by Linear Programming*, IEEE Inf. Theory **51** (2005), 4203–4215.

[CaTa2006] E. Candés, T. Tao, *Near Optimal Signal Recovery From Random Projections: Universal Encoding Strategies?*, IEEE Inf. Theory **52** (2006), 5406–5425.

[CaTa2007] E. Candés, T. Tao, *The Dantzig selector: statistical estimation when p is much larger than n*, Annals of Statistics **35** (2007), 2313–2351.

[Ca1966] L. Carleson, *On convergence and growth of partial sums of Fourier series*, Acta Math., vol. **116** (1966), 135–157.

[Caz2003] T. Cazenave, *Semilinear Schrödinger equations*, Courant Lecture Notes in Mathematics, **10**. New York University, Courant Institute of Mathematical Sciences, AMS, 2003.

[Ch2008] M-C. Chang, *Product theorems in SL_2 and SL_3*, preprint.

[ChMa1995] L. Chayes, J. Machta, *On the behavior of the surface tension for spin systems in a correlated porous medium*, J. Statist. Phys. **79** (1995), no. 1-2, 117–164.

[ChMcKWi1998] L. Chayes, D. McKellar, B. Winn, *Percolation and Gibbs states multiplicity for ferromagnetic Ashkin-Teller models on* $\mathbf{Z}^2$, J. Phys. A **31** (1998), no. 45, 9055–9063.

[Ch1973] J. R. Chen, *On the representation of a larger even integer as the sum of a prime and the product of at most two primes*, Sci. Sinica **16** (1973), 157–176.

[ChLuTi2006] X. Chen, P. Lu, G. Tian, *A note on uniformization of Riemann surfaces by Ricci flow*, Proc. Amer. Math. Soc. **134** (2006), no. 11, 3391–3393.

[Ch2001] M. Christ, *On certain elementary trilinear operators*, Math. Res. Lett. **8** (2001), no. 1-2, 43–56.

[Ch1988] M. Christ, *Weak type* $(1,1)$ *bounds for rough operators*, Ann. of Math. (2) **128** (1988), no. 1, 19–42

[ChKi2001] M. Christ, A. Kiselev, *WKB asymptotic behavior of almost all generalized eigenfunctions for one-dimensional Schrödinger operators with slowly decaying potentials*, J. Funct. Anal. **179** (2001), no. 2, 426–447.

[CdV1985] Y. Colin de Verdiére, *Ergodicité et fonctions propres du laplacien*, Bony-Sjöstrand-Meyer seminar, 1984–1985, Exp. No. 13, 8 pp., École Polytech., Palaiseau, 1985.

[CoKeStTaTa2008] J. Colliander, M. Keel, G. Staffilani, H. Takaoka, and T. Tao, *Global well-posedness and scattering for the energy-critical nonlinear Schrödinger equation in* $\mathbf{R}^3$, to appear in Annals of Math.

[CoLe1988] J. Conze, E. Lesigne, *Sur un théoréme ergodique pour des mesures diagonales*, C. R. Acad. Sci. Paris Sér. I Math. **306** (1988), no. 12, 491–493

[Co2000] J. Conway, *Universal quadratic forms and the fifteen theorem*, Quadratic forms and their applications (Dublin, 1999), 23–26, Contemp. Math., 272, Amer. Math. Soc., Providence, RI, 2000.

[Cr2008] E. Croot, *The Minimal Number of Three-Term Arithmetic Progressions Modulo a Prime Converges to a Limit*, preprint.

[Da1986] S. Dani, *On the orbits of unipotent flows on homogeneous spaces. II*, Ergodic Thy. Dynam. Systems **6** (1986), 167–182.

[DeKi1999] P. Deift, R. Killip, *On the absolutely continuous spectrum of one-dimensional Schrödinger operators with square summable potentials*, Comm. Math. Phys. **203** (1999), no. 2, 341–347.

[De1974] P. Deligne, *La conjecture de Weil I.*, Inst. Hautes Études Sci. Publ. Math., **48** (1974), pp. 273–308.

[De2008] C. Demeter, *Divergence of combinatorial averages and the unboundedness of the trilinear Hilbert transform*, preprint.

[DeTaTh2007] C. Demeter, T. Tao, C. Thiele, *A trilinear maximal inequality via arithmetic combinatorics*, available at `www.math.ucla.edu/\~{}tao/preprints/Expository/maximal.dvi`

[De2007] H. Derksen, *A Skolem-Mahler-Lech theorem in positive characteristic and finite automata*, Invent. Math. **168** (2007), no. 1, 175–224.

[DeWe2000] H. Derksen, J. Weyman, *Semi-invariants of quivers and saturation for Littlewood-Richardson coefficients*, J. Amer. Math. Soc. **13** (2000), no. 3, 467–479.

[dV2008] R. de Vore, *Deterministic constructions of compressed sensing matrices*, preprint.

[Di1904] L. E. Dickson, *A new extension of Dirichlet's theorem on prime numbers*, Messenger of Math. **33** (1904), 155–161.

[DiHe1976] W. Diffie, M. Hellman, *New Directions in Cryptography*, IEEE Transactions on Information Theory, vol. IT-22, Nov. 1976, pp: 644-654.

[Do2006] D. Donoho, *For most large underdetermined systems of equations, the minimal l_1-norm near-solution approximates the sparsest near-solution*, Comm. Pure Appl. Math. **59** (2006), no. 7, 907–934.

[Ed2004] Y. Edel, *Extensions of generalized product caps*, Designs, Codes, and Cryptography, **31** (2004), 5–14.

[EdMi2003] G. Edgar, C. Miller, *Borel subrings of the reals*, Proc. Amer. Math. Soc. **131** (2003), no. 4, 1121–1129

[EeSa1964] J. Eells, J. Sampson, *Harmonic mappings of Riemannian manifolds*, Amer. J. Math. **86** (1964), 109–160.

[EiMaVe2008] M. Einsiedler, G. Margulis, A. Venkatesh, *Effective equidistribution for closed orbits of semisimple groups on homogeneous spaces*, preprint.

[Ei1905] A. Einstein, *Ist die Trägheit eines Körpers von dessen Energieinhalt abhängig?*, Annalen der Physik **18** (1905), 639-643.

[El1997] G. Elekes, *On the number of sums and products*, Acta Arith. **81** (1997), 365–367.

[ElHa1969] P. Elliot, H. Halberstam, *A conjecture in prime number theory*, Symp. Math. **4** (1968-1969), 59–72.

[ElKi2001] G. Elekes, Z. Király, *On the combinatorics of projective mappings*, J. Algebraic Combin. **14** (2001), no. 3, 183–197.

[ElSz2008] G. Elek, B. Szegedy, *Limits of Hypergraphs, Removal and Regularity Lemmas. A Non-standard approach*, preprint.

[En1978] V. Enss, *Asymptotic completeness for quantum mechanical potential scattering. I. Short range potentials*, Comm. Math. Phys. **61** (1978), no. 3, 285–291.

[Er1945] P. Erdős, *On a lemma of Littlewood and Offord*, Bull. Amer. Math. Soc. **51** (1945), 898–902.

[Er1947] P. Erdős, *Some remarks on the theory of graphs*, Bull. Am. Math. Soc. **53** (1947), 292–294.

[Er1949] P. Erdős, *On a new method in elementary number theory*, Proc. Nat. Acad. Sci. U.S.A. **35** (1949), 374–384.

[ErSz1983] P. Erdős, E. Szeremédi, *On sums and products of integers*, Studies in Pure Mathematics; To the memory of Paul Turán. P. Erdos, L. Alpar, and G. Halasz, editors. Akademiai Kiado - Birkhauser Verlag, Budapest - Basel-Boston, Mass. 1983, 213-218.

[EsSeSv2003] L. Eskauriaza, G. Serëgin, G., V. Sverák, *$L^{3,\infty}$-solutions of Navier-Stokes equations and backward uniqueness*, (Russian) Uspekhi Mat. Nauk **58** (2003), no. 2(350), 3–44; translation in Russian Math. Surveys **58** (2003), no. 2, 211–250.

[EvScSc2002] J. Evertse, H. Schlickewei, W. Schmidt, *Linear equations in variables which lie in a multiplicative group*, Ann. of Math. (2) 155 (2002), no. 3, 807–836.

[FNdB2003] F. Faure, S. Nonnenmacher, S. De Biévre, *Scarred eigenstates for quantum cat maps of minimal periods*, Comm. Math. Phys. **239** (2003), no. 3, 449–492.

[Fe1973] C. Fefferman, *Pointwise convergence of Fourier series*, Ann. of Math. **98** (1973), 551–571.

[Fe2006] C. Fefferman, *Existence and smoothness of the Navier-Stokes equation*, Millennium Prize Problems, Clay Math. Inst., Cambridge, MA, 2006, 57–67.

[Fi1989] D. Fisher, *Lower bounds on the number of triangles in a graph*, J. Graph Theory **13** (1989), no. 4, 505–512.

[FoKa1972] C. Fortuin, P. Kasteleyn, *On the random-cluster model. I. Introduction and relation to other models*, Physica **57** (1972), 536–564.

[Fr1998] M. Freedman, *K-sat on groups and undecidability*, Annual ACM Symposium on Theory of Computing archive, Proceedings of the thirtieth annual ACM symposium on Theory of computing, 1998, 572–576.

[FrIw1998] J. Friedlander, H. Iwaniec, *The polynomial $X^2 + Y^4$ captures its primes*, Ann. of Math. (2) **148** (1998), no. 3, 945–1040.

[FrPa2008] S. Friedlander, N. Pavlovíc, *Dyadic models for the equations of fluid motion*, preprint.

[Fr1973] G. Freiman, *Foundations of a structural theory of set addition*. Translated from the Russian. Translations of Mathematical Monographs, Vol 37. American Mathematical Society, Providence, R. I., 1973. vii+108 pp.

[FrKa1999] A. Frieze, R. Kannan, *Quick approximation to matrices and applications*, Combinatorica **19** (1999), no. 2, 175–220.

[Fu2000] B. Fulton, *Eigenvalues, invariant factors, highest weights, and Schubert calculus*, Bull. Amer. Math. Soc. (N.S.) **37** (2000), no. 3, 209–249.

[Fu1977] H. Furstenberg, *Ergodic behavior of diagonal measures and a theorem of Szemerédi on arithmetic progressions*, J. Analyse Math. **31** (1977), 204–256.

[Fu1981] H. Furstenberg, *Recurrence in Ergodic theory and Combinatorial Number Theory*, Princeton University Press, Princeton NJ 1981.

[FuKaOr1982] H. Furstenberg, Y. Katznelson, D. Ornstein, *The ergodic theoretical proof of Szemerédi's theorem*, Bull. Amer. Math. Soc. **7** (1982), 527–552.

[FuWe1996] H. Furstenberg, B. Weiss, *A mean ergodic theorem for $(1/N)\sum_{n=1}^{N} f(T^n x)g(T^{n^2}x)$*, Convergence in ergodic theory and probability (Columbus, OH, 1993), 193–227, Ohio State Univ. Math. Res. Inst. Publ., 5, de Gruyter, Berlin, 1996.

[Ga1993] G. Gamow, R. Penrose (Foreword), Mr. Tompkins in Paperback (Omnibus of Mr. Tompkins in Wonderland and Mr Tompkins Explores the Atom), Cambridge University Press, 1993.

[GaKeRu1998] A. Gandolfi, M. Keane, L. Russo, *On the uniqueness of the infinite occupied cluster in dependent two-dimensional site percolation*, Ann. Probab. **16** (1988), no. 3, 1147–1157.

[Ga2008] M. Garaev, *An explicit sum-product estimate in* $\mathbb{F}_p$, preprint.

[GeLi1993] P. Gérard, E. Leichtnam, *Ergodic properties of eigenfunctions for the Dirichlet problem*, Duke Math. J. **71** (1993), no. 2, 559–607.

[GiTr2008] A. Gilbert, J. Tropp, *Signal recovery from partial information via Orthogonal Matching Pursuit*, preprint.

[GiStTrVe2007] A. C. Gilbert, M. J. Strauss, J. A. Tropp, R. Vershynin, *One Sketch for All: Fast Algorithms for Compressed Sensing*, in Proceedings of the ACM Symposium on the Theory of Computing (STOC 2007), 2007

[GiVe1985] J. Ginibre, G. Velo, *Scattering theory in the energy space for a class of nonlinear Schrdinger equations*, J. Math. Pures Appl. (9) **64** (1985), no. 4, 363–401.

[Gi1984] V. L. Girko, *Circular law*, Theory Probab. Appl. (1984), 694–706.

[GlKo2008] A. Glibichuk, S. Konyagin, *Additive properties of product sets in fields of prime order*, preprint.

[GoYiPi2008] D. Goldston, J. Pintz, C. Yıldırım, *Primes in Tuples II*, preprint.

[Go1959] A. Goodman, *On sets of acquaintances and strangers at any party*, Amer. Math. Monthly **66** (1959), 778–783.

[GoTi2008] F. Götze, A.N. Tikhomirov, On the circular law, *preprint*

[GoTi2008b] F. Götze, A.N. Tikhomirov, The Circular Law for Random Matrices, *preprint*

[Go1997] T. Gowers, *Lower bounds of tower type for Szemerédi's uniformity lemma*, Geom. Func. Anal. **7** (1997), 322–337.

[Go2000] T. Gowers, *The two cultures of mathematics*, in: Mathematics: Frontiers and Perspectives, International Mathematical Union. V. Arnold, M. Atiyah, P. Lax, B. Mazur, Editors. American Mathematical Society, 2000.

[Go2001] W. T. Gowers, *A new proof of Szemerédi's Theorem,* Geom. Funct. Anal. **11** (2001), no. 3, 465–588.

[Gr2001] G. Greaves, Sieves in number theory, Springer, Berlin 2001.

[Gr2005] B. Green, *A Szemerédi-type regularity lemma in abelian groups, with applications*, Geom. Funct. Anal. **15** (2005), no. 2, 340–376.

[Gr2005b] B. Green, *Finite field models in additive combinatorics*, Surveys in combinatorics 2005, 1–27, London Math. Soc. Lecture Note Ser., 327, Cambridge Univ. Press, Cambridge, 2005

[GrKo2006] B. Green, S. Konyagin, *On the Littlewood problem modulo a prime*, preprint.

[GrRu2004] B. Green, I. Ruzsa, *On the Hardy-Littlewood majorant problem*, Math. Proc. Cambridge Philos. Soc. **137** (2004), 511–517.

[GrRu2007] B. Green, I. Ruzsa, *Freiman's theorem in an arbitrary abelian group*, J. Lond. Math. Soc. (2) **75** (2007), no. 1, 163–175.

[GrSi2008] B. Green, O. Sisask, *On the maximal number of three-term arithmetic progressions in subsets of* $\mathbf{Z}/p\mathbf{Z}$, preprint.

[GrTa2008] B. Green and T. Tao, *The primes contain arbitrarily long arithmetic progressions*, to appear in Annals of Math.

[GrTa2008b] B. Green, T. Tao, *Linear equations in primes*, to appear, Annals of Math.

[GrTa2008c] B. Green, T. Tao, *An inverse theorem for the Gowers* $U^3(G)$ *norm*, preprint.

[GrTa2008d] B. Green, T. Tao, *Quadratic uniformity of the Möbius function*, preprint.

[GrTa2008e] B. Green, T. Tao, *Compressions, convex geometry, and the Freiman-Bilu theorem*, Quarterly J. Math. **57** (2006), 495–504.

[GrTa2008f] B. Green, T. Tao, *The quantitative behaviour of polynomial orbits on nilmanifolds*, preprint.

[Gr1981] M. Gromov, *Groups of polynomial growth and expanding maps*, Inst. Hautes Études Sci. Publ. Math. No. 53 (1981), 53–73.

[GuSt1982] V. Guillemin, S. Sternberg, *Convexity properties of the moment mapping*, Invent. Math. **67** (1982), no. 3, 491–513.

[HaHaHiKa2005] V. Halava, T. Harju, M. Hirvensalo, J. Karhumaki, *Skolem's problem: on the border between decidability and undecidability*, Tech report 683, Turku Unviersity Computer Science, 2005.

[HaRi1974] H. Halberstam and H. E. Richert, Sieve methods, Academic Press (1974)

[HaLlSe1998] Y. Hamiduone, A. Lladó, O. Serra, *On subsets with small product in torsion-free groups*, Combinatorica **18** (1998), 529–540.

[Ha1982] R. Hamilton, *Three-manifolds with positive Ricci curvature*, J. Differential Geom. **17** (1982), no. 2, 255–306.

[Ha1988] R. Hamilton, *The Ricci flow on surfaces*, Mathematics and general relativity (Santa Cruz, CA, 1986), 237–262, Contemp. Math., 71, Amer. Math. Soc., Providence, RI, 1988.

[Ha1985] G. Hansel, *A simple proof of the Skolem-Mahler-Lech theorem*, Automata, languages and programming (Nafplion, 1985), 244–249, Lecture Notes in Comput. Sci., 194, Springer, Berlin, 1985.

[Ha1976] B. Hansson, *The existence of group preference functions*, Public Choice **28** (1976), 89-98.

[HaLi1923] G.H. Hardy and J.E. Littlewood *Some problems of "partitio numerorum"; III: On the expression of a number as a sum of primes,* Acta Math. **44** (1923), 1–70.

[HB1983] R. Heath-Brown, *Prime twins and Siegel zeros*, Proc. London Math. Soc. (3) **47** (1983), no. 2, 193–224.

[HB2001] R. Heath-Brown, *Primes represented by* $x^3 + 2y^3$, Acta Math. **186** (2001), no. 1, 1–84.

[HBMo2002] R. Heath-Brown, B. Moroz, *Primes represented by binary cubic forms*, Proc. London Math. Soc. (3) **84** (2002), no. 2, 257–288.

[He2006] H. Helfgott, *The parity problem for reducible cubic forms*, J. London Math. Soc. (2) **73** (2006), no. 2, 415–435

[He2008] H. Helfgott, *Growth and generation in* $SL_2(\mathbf{Z}/p\mathbf{Z})$, Ann. of Math. **167** (2008), 601-623.

[He1991] E. Heller, *Wavepacket dynamics and quantum chaology*, Chaos et physique quantique (Les Houches, 1989), 547–664, North-Holland, Amsterdam, 1991.

[HeKa2006] A. Henriques, J. Kamnitzer, *The octahedron recurrence and* $\mathfrak{gl}_n$ *crystals*, Adv. Math. **206** (2006), 211–249.

[He1930] J. Herbrand, *Recherches sur la théorie de la démonstration*, PhD thesis, University of Paris, 1930.

[Hi1974] N. Hindman, *Finite sums from sequences within cells of a partition of N*, J. Comb. Th. A **17** (1974), 1–11.

[HoRo2008] J. Holmer, S. Roudenko, *A sharp condition for scattering of the radial 3d cubic nonlinear Schroedinger equation*, preprint.

[HoPy1974] D. Holton, W. Pye, Creating Calculus. Holt, Rinehart and Winston, 1974.

[Ho1962] A. Horn, *Eigenvalues of sums of Hermitian matrices*, Pacific J. Math. **12** (1962) 225–241.

[HoKr2005] B. Host, B. Kra, *Nonconventional ergodic averages and nilmanifolds*, Ann. of Math. (2) **161** (2005), no. 1, 397–488.

[Is2008] Y. Ishigami, *A Simple Regularization of Hypergraphs*, preprint.

[Iw1974] H. Iwaniec, *Primes represented by quadratic polynomials in two variables*, Collection of articles dedicated to Carl Ludwig Siegel on the occasion of his seventy-fifth birthday, V. Acta Arith. **24** (1973/74), 435–459.

[Jo1948] F. John, *Extremum problems with inequalities as subsidiary conditions*, Studies and Essays presented to R. Courant on his 60th birthday, Interscience Publishers Inc., New York, NY 1948, 187–204.

[Jo1979] F. John, *Blow-up of solutions of non-linear wave equations in three dimensions*, Manuscript. Math. **28** (1979), 235–268.

[JoLi1984] W. Johnson, J. Lindenstrauss, *Extensions of Lipschitz maps into a Hilbert space*, Contemporary Mathematics, **26** (1984), 189–206.

[Le1983] T. Leighton, *Complexity Issues in VLSI*, Foundations of Computing Series, MIT Press, Cambridge, MA, 1983.

[LoWh1949] L. Loomis, H. Whitney, *An inequality related to the isoperimetric inequality*, Bull. Amer. Math. Soc **55**, (1949) 961–962.

[Ka1966] M. Kac, *Can one hear the shape of a drum?*, American Mathematical Monthly **73** (1966), part II, 1–23.

[KaLeMi2008] M. Kapovich, B. Leeb, J. Milson, *The generalized triangle inequalities in symmetric spaces and buildings with applications to algebra*, preprint.

[Ka1990] M. Kashiwara, *Crystalizing the q-analogue of universal enveloping algebras*, Comm. Math. Phys. **133** (1990), no. 2, 249–260.

[Ka1993] T. Kato, Abstract evolution equations, linear and quasilinear, revisited. Functional analysis and related topics, 1991 (Kyoto), 103–125, Lecture Notes in Math., 1540, Springer, Berlin, 1993.

[KaUg2007] S. Katok, I. Ugarcovici, *Symbolic dynamics for the modular surface and beyond*, Bull. Amer. Math. Soc. **44** (2007) 87–132.

[Ka1968] G. Katona, *A theorem of finite sets*, Theory of Graphs, P. Erdős and G. Katona (eds.), Akadémiai Kiadó and Academic Press, 1968.

[KaPa2005] N. Katz, N. Pavlović, *Finite time blow-up for a dyadic model of the Euler equations*, Trans. Amer. Math. Soc. **357** (2005), no. 2, 695–708.

[KaSh2008] N. Katz, C-Y. Shen, *A Slight Improvement to Garaev's Sum Product Estimate*, preprint.

[KaTa1999] N. Katz, T. Tao, *Bounds on arithmetic projections, and applications to the Kakeya conjecture*, Math. Res. Lett. **6** (1999), no. 5-6, 625–630.

[KaTa2001] N. Katz, T. Tao, *Some connections between Falconer's distance set conjecture and sets of Furstenburg type*, New York J. Math. 7 (2001), 149–187.

[KeMe2006] C. Kenig and F. Merle, *Global well-posedness, scattering, and blowup for the energy-critical, focusing, non-linear Schrödinger equation in the radial case,* Invent. Math. **166** (2006), 645–675.

[KiTaVi2008] R. Killip, T. Tao, M. Visan, *The cubic nonlinear Schrodinger equation in two dimensions with radial data*, preprint.

[KiSo1972] A.P. Kirman, D. Sondermann, *Arrows theorem, many agents, and invisible dictators*, Journal of Economic Theory **5** (1972), pp. 267-277.

[Ki1984] F. Kirwan, *Convexity properties of the moment mapping. III*, Invent. Math. **77** (1984), no. 3, 547–552.

[Kl2008] B. Kleiner, *A new proof of Gromov's theorem on groups of polynomial growth*, preprint.

[Kl2000] S. Klainerman, *PDE as a unified subject*, GAFA 2000 (Tel Aviv, 1999), Geom. Funct. Anal. 2000, Special Volume, Part I, 279–315.

[Kl1998] A. Klyachko, *Stable bundles, representation theory and Hermitian operators* Selecta Math. (N.S.) **4** (1998), no. 3, 419–445.

[KoTa2001] H. Koch, D. Tataru, *Well-posedness for the Navier-Stokes equations*, Adv. Math. **157** (2001), no. 1, 22–35.

[Ko2008] U. Kohlenbach, Applied Proof Theory: Proof Interpretations and Their Use in Mathematics. Springer Verlag, Berlin, 1–536, 2008.

[KnTa1999] A. Knutson, T. Tao, *The honeycomb model of* $\mathrm{GL}_n(C)$ *tensor products. I. Proof of the saturation conjecture*, J. Amer. Math. Soc. **12** (1999), no. 4, 1055–1090.

[KnTa2001] A. Knutson, T. Tao, *Honeycombs and sums of Hermitian matrices*, Notices Amer. Math. Soc. **48** (2001), no. 2, 175–186.

[KnTa2001b] A. Knutson, T. Tao, *Honeycomb applet*, available at `www.math.ucla.edu/\~{}tao/java/Honeycomb.html`

[KnTa2003] A. Knutson, T. Tao, *Puzzles and (equivariant) cohomology of Grassmannians*, Duke Math. J. **119** (2003), no. 2, 221–260.

[KnTaWo2004] A. Knutson, T. Tao, C. Woodward, *A positive proof of the Littlewood-Richardson rule using the octahedron recurrence*, Electron. J. Combin. **11** (2004), no. 1, Research Paper 61, 18 pp.

[Kr1963] J.B. Kruskal, *The number of simplices in a complex*, Mathematical Optimization Techniques, R. Bellman (ed.), University of California Press, 1963.

[KuNe1974] L. Kuipers and H. Neiderreiter, Uniform Distribution of Sequences, Wiley, New York, 1974.

[Ku1930] K. Kuratowski, *Sur le probléme des courbes gauches en topologie*, Fund. Math. 15 (1930), 271-283.

[Ku1992] G. Kuperberg, *A low-technology estimate in convex geometry*, Internat. Math. Res. Notices **9** (1992), 181–183.

[Ku2008] G. Kuperberg, *From the Mahler conjecture to Gauss linking integrals*, preprint.

[La2000] M. Lacey, *The bilinear maximal functions map into L^p for $2/3 < p \leq 1$*, Ann. of Math. (2) **151** (2000), no. 1, 35–57.

[LaTh1997] M. Lacey, C. Thiele, *L^p estimates on the bilinear Hilbert transform for $2 < p < \infty$*, Ann. of Math. (2) **146** (1997), no. 3, 693–724.

[LaTh1999] M. Lacey, C. Thiele, *On Calderón's conjecture*, Ann. of Math. (2) **149** (1999), no. 2, 475–496.

[LaTh2000] M. Lacey, C. Thiele, *A proof of boundedness of the Carleson operator*, Math. Res. Lett. **7** (2000), no. 4, 361–370.

[Le1953] C. Lech, *A note on recurring series*, Arkiv fur Matematik, Band 2 nr 22 (1953), 417–421.

[Le1933] J. Leray, *Étude de diverses équations intégrales nonlinéaires et de quelques problémes que pose lhy-drodynamique*, J. Math. Pure Appl. **12** (1933), 1-82.

[Li2001] E. Lindenstrauss, *Pointwise theorems for amenable groups*, Invent. Math. **146** (2001), 259–295.

[Li2003] Y. Li, *Chaos in PDEs and Lax pairs of Euler equations*, Acta Appl. Math. **77** (2003), no. 2, 181–214.

[LiWa2002] M.-C. Liu, T. Wang, *On the Vinogradov bound in the three primes Goldbach conjecture*, Acta Arith. **105** (2002), no. 2, 133–175.

[LoSz2006] L. Lovász, B. Szegedy, *Limits of dense graph sequences*, J. Combin. Theory Ser. B **96** (2006), no. 6, 933–957.

[LoSz2008] L. Lovász, B. Szegedy, *Graph limits and testing hereditary properties*, preprint.

[LuZh1997] E. Lutwak, G. Zhang, *Blaschke-Santaló inequalities*, J. Diff. Geom. **45** (1997), 1–16.

[McK2007] D. McKellar, Math Doesn't Suck: How to Survive Middle-School Math Without Losing Your Mind or Breaking a Nail, Hudson Street Press, 2007.

[Ma1935] K. Mahler, *Eine arithmetische Eigenshaft der Taylor coefficienten rationaler Funktionen*, Proc. Acad. Sci. Amst. **38** (1935), 51–60.

[Ma1939] K. Mahler, *Ein Minimalproblem fur konvexe Polygone*, Mathematica (Zutphen) B, 118-127 (1939);

[Ma1956] K. Mahler, *On the Taylor coefficients of rational functions*, Proc. Cambridge Philos. Soc. **52** (1956), 39–48.

[MaZy1939] J. Marcinkiewicz, A. Zygmund, *Quelques inégalités pour les opérations linéaires*, Fund Math. **32** (1939), 115-121.

[Ma1989] G. Margulis, *Discrete subgroups and ergodic theory*, Number theory, trace formulas and discrete groups (Oslo, 1987), 377–398, Academic Press, Boston, MA, 1989.

[MaMeTs2002] Y. Martel, F. Merle, T.-P. Tsai, *Stability and asymptotic stability in the energy space of the sum of N solitons for subcritical gKdV equations*, Comm. Math. Phys. **231** (2002), no. 2, 347–373.

[Me1967] M.L. Mehta, Random Matrices and the Statistical Theory of Energy Levels, Academic Press, New York, NY, 1967.

[Me1995] R. Meshulam, *On subsets of finite abelian groups with no 3-term arithmetic progressions*, J. Combin. Theory Ser. A., **71** (1995), 168–172.

[Me2006] R. Meshulam, An uncertainty principle for finite abelian groups, Eur. J. Comb. **27** (2006), 63–67.

[MiVa2003] J. Miller, B. Van Loon, Darwin for Beginners. Pantheon, 2003.

[MoSh2002] G. Mockenhaupt, W. Schlag, *On the Hardy-Littlewood majorant problem for random sets*, preprint, 2002.

[MoVa1973] H. Montgomery, R. Vaughan, *The large sieve*, Mathematika **20** (1973), 119–134.

[Mo2005] D. Witte Morris, Ratner's theorems on unipotent flows. Chicago Lectures in Mathematics. University of Chicago Press, Chicago, IL, 2005.

[Mo1921] M. Morse, *Recurrent geodesics on a surface of negative curvature*, Trans. Amer. Math. Soc. **22** (1921), 84–100.

[Mu2007] K. Mulmuley, *Geometric complexity theory VI: the flip via saturated and positive integer programming in representation theory and algebraic geometry*, preprint.

[MuTaTh2003] C. Muscalu, T. Tao, C. Thiele, *A counterexample to a multilinear endpoint question of Christ and Kiselev*, Math. Res. Letters **10** (2003), 237–246.

[MuTaTh2003b] C. Muscalu, T. Tao, C. Thiele, *A Carleson-type theorem for a Cantor group model of the Scattering Transform*, Nonlinearity **19** (2003), 219–246

[Ni2008] V. Nikiforov, *An extension of Maclaurin's inequalities*, preprint.

[NoSt1963] E. Nordhaus, B. Stewart, *Triangles in an ordinary graph*, Canad. J. Math. **15** (1963), 33–41.

[No1980] V. Novoksenov, *Asymptotic behavior as $t \to \infty$ of the solution of the Cauchy problem for a nonlinear Schrödinger equation*, Dokl. Akad. Nauk SSSR **251** (1980), no. 4, 799–802.

[Op1929] A. Oppenheim, *The minima of indefinite quaternary quadratic forms*, Proc. Natl. Acad. Sci. USA **15** (1929), 724-727.

[PaRaTaTo2006] J. Pach, R. Radoičić, G. Tardos, G. Tóth, *Improving the crossing lemma by finding more crossings in sparse graphs*, Discrete Comput. Geom. **36** (2006), no. 4, 527–552.

[PaVa2005] I. Pak, E. Vallejo, *Combinatorics and geometry of Littlewood-Richardson cones*, European J. Combin. **26** (2005), no. 6, 995–1008.

[PaZh2008] G. Pan and W. Zhou, *Circular law, Extreme singular values and potential theory*, preprint.

[PaHa1977] J. Paris, L. Harrington, *A Mathematical Incompleteness in Peano Arithmetic*, In Handbook for Mathematical Logic (Ed. J. Barwise). Amsterdam, Netherlands: North-Holland, 1977.

[Pe2002] G. Perelman, *The entropy formula for the Ricci flow and its geometric applications*, preprint.

[Pe2003] G. Perelman, *Ricci flow with surgery on three-manifolds*, preprint.

[Pe2003b] G. Perelman, *Finite extinction time for the solutions to the Ricci flow on certain three-manifolds*, preprint.

[Pe2001] C. Pereyra, *Lecture notes on dyadic harmonic analysis*, In "Second Summer school in analysis and mathematical physics. Topics in analysis: harmonic, complex, nonlinear and quantization," Cuernavaca Morelos, Mexico, June 12-22, 2000. S. Pérez-Esteva, C.Villegas eds. Contemporary Mathematics 289 AMS, Ch. I, p. 1-61 (2001).

[Pl1969] H. Plünnecke, *Eigenschaften un Abschätzungen von Wirkingsfunktionen*, BMwF-GMD-22 Gesellschaft für Mathematik und Datenverarbeitung, Bonn 1969.

[Ra1930] F.P.Ramsey, *On a problem of formal logic*, Proc. London Math. Soc. **30** (1930), 264–285.

[Ra1991] M. Ratner, *Raghunatan's topological conjecture and the distribution of unipotent flows*, Duke Math. J. **61** (1991) no. 1, 235–280.

[Ra2008] A. Razborov, *A Product Theorem in Free Groups*, preprint.

[Ra2008b] A. Razborov, *On the Minimal Density of Triangles in Graphs*, preprint.

[Ra2008c] A. Razborov, *Flag algebras*, preprint.

[Re1986] S. Reisner, *Zonoids with minimal volume-product*, Math. Z. **192** (1986), no. 3, 339–346.

[RoSc2007] V. Rödl, M. Schacht, *Regular partitions of hypergraphs*, Regularity Lemmas, Combinatorics, Probability and Computing, **16** (2007), 833–885.

[RoSh1957] C.A. Rogers, G.C. Shephard, *The difference body of a convex body*, Arch. Math. **8** (1957), 220–233.

[Ro2002] M. Rosen, Number theory in function fields. Graduate Texts in Mathematics, 210. Springer-Verlag, New York, 2002.

[Ro1953] K.F. Roth, *On certain sets of integers*, J. London Math. Soc. **28** (1953), 245–252.

[Ro1955] K. F. Roth, *Rational approximations to algebraic numbers*, Mathematika, **2** (1955), 1–20.

[RuSa1994] Z. Rudnick, P. Sarnak, *The behaviour of eigenstates of arithmetic hyperbolic manifolds*, Comm. Math. Phys. **161** (1994), no. 1, 195–213.

[RuVe2006] M. Rudelson, R.Vershynin, *Sparse reconstruction by convex relaxation: Fourier and Gaussian measurements*, CISS 2006 (40th Annual Conference on Information Sciences and Systems)

[Ru1969] D. Ruelle, *A remark on bound states in potential-scattering theory*, Nuovo Cimento A **61** (1969) 655–662.

[Ru1996] I. Ruzsa, *Sums of finite sets*, Number Theory: New York Seminar; Springer-Verlag (1996), D.V. Chudnovsky, G.V. Chudnovsky and M.B. Nathanson editors.

[RuSz1978] I. Ruzsa, E. Szemerédi, *Triple systems with no six points carrying three triangles*, Colloq. Math. Soc. J. Bolyai **18** (1978), 939–945.

[RyVi2007] E. Ryckman and M. Visan, *Global well-posedness and scattering for the defocusing energy-critical nonlinear Schrödinger equation in* $\mathbf{R}^{1+4}$, Amer. J. Math. **129** (2007), 1–60.

[SaUh1981] J. Sacks, K. Uhlenbeck, *The existence of minimal immersions of* 2-*spheres*, Ann. of Math. (2) **113** (1981), no. 1, 1–24.

[SR1981] J. Saint-Raymond, *Sur le volume des corps convexes symétriques*, Initiation Seminar on Analysis: G. Choquet-M. Rogalski-J. Saint-Raymond, 20th Year: 1980/1981, Exp. No. 11, 25 pp., Publ. Math. Univ. Pierre et Marie Curie, 46, Univ. Paris VI, Paris, 1981.

[Sa1949] L. Santaló, *Un invariante afin para los cuerpos convexos del espacio de n dimensiones*, Portugalie Math. **8** (1949), 155–161.

[Sa1998] Y. Saouter, *Checking the odd Goldbach conjecture up to* 10^{20}, Math. Comp. **67** (1998), no. 222, 863–866.

[Sc1985] J. Schaeffer, *The equation* $u_{tt} - \Delta u = |u|^p$ *for the critical value of* p, Proc. Roy. Soc. Edinburgh Sect. A **101** (1985), no. 1-2, 31–44.

[Sc1976] V. Scheffer, *Partial regularity of solutions to the Navier-Stokes equations*, Pacific J. Math. **66** (1976), no. 2, 535–552.

[Sc1993] V. Scheffer, *An inviscid flow with compact support in space-time*, J. Geom. Anal. **3** (1993), no. 4, 343–401.

[Sc1989] E. Schrödinger, Statistical thermodynamics, Dover, 1989.

[Sc1916] I. Schur, *Über die Kongruenz* $x^m + y^m = z^m (mod p)$, Jber. Deutsch. Math.-Verein. **25** (1916), 114–116.

[Sc2006] P. Scowcroft, *Nonnegative solvability of linear equations in certain ordered rings*, Trans. Amer. Math. Soc. **358** (2006), 3535-3570.

[Se1949] A. Selberg, *An elementary proof of the prime number theorem*, Ann. Math. **50** (1949), 305–313

[ShWrMa2008] Y. Sharon, J. Wright, Y. Ma, *Computation and relaxation of conditions for equivalence between ell-1 and ell-0 minimization*, preprint.

[Si1994] A. Sidorenko, *An analytic approach to extremal problems for graphs and hypergraphs*, Extremal problems for finite sets (Visegrád, 1991), 423–455, Bolyai Soc. Math. Stud., 3, János Bolyai Math. Soc., Budapest, 1994.

[Sk1933] T. Skolem, *Einige Sätze über gewisse Reihenentwichlungen und exponentiale Bezeihungen mit Anwendung auf diophantische Gleichungen*, Oslo Vid. Akad. Skriften I. **6** (1933).

[Sn1974] A. Šnirelman, *Ergodic properties of eigenfunctions* (Russian) Uspehi Mat. Nauk **29** (1974), no. 6(180), 181–182.

[So2006] A. Soffer, *Soliton dynamics and scattering*, International Congress of Mathematicians. Vol. III, 459–471, Eur. Math. Soc., Zrich, 2006.

[So2005] J. Solymosi, *On the number of sums and products*, Bull. London Math. Soc. **37** (2005), no. 4, 491–494.

[So2008] J. Solymosi, *On sumsets and product sets of complex numbers*, Journal de Theorie des Nombres de Bordeaux **17** (2005), 921–924.

[Sp2005] D. Speyer, *Horn's Problem, Vinnikov Curves and Hives*, Duke Journal of Mathematics **127** (2005), 395–428.

[St1961] E. M. Stein, *On limits of sequences of operators*, Ann. of Math. (2) **74** 1961 140–170.

[St1970] E. M. Stein, *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, 1970.

[St1965] R. Steinberg, *Regular elements of semi-simple algebraic groups*, Publ. IHES **25** (1965), 49–80.

[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.

[StLe1996] P. Stevenhagen, H.W. Lenstra Jr., *Chebotarëv and his density theorem*, Math. Intelligencer **18** (1996), no. 2, 26–37.

[Sz1997] L. Székely, *Crossing numbers and hard Erdős problems in discrete geometry*, Combin. Probab. Comput. **6** (1997), 353–358.

[Sz1975] E. Szemerédi, *On sets of integers containing no k elements in arithmetic progression*, Acta Arith. **27** (1975), 299–345.

[Sz1978] E. Szemerédi, *Regular partitions of graphs*, in "Problémes Combinatoires et Théorie des Graphes, Proc. Colloque Inter. CNRS," (Bermond, Fournier, Las Vergnas, Sotteau, eds.), CNRS Paris, 1978, 399–401.

[Ta1950] J.T. Tate, Fourier analysis in number fields and Hecke's zeta-functions, 1950 Princeton Ph.D. thesis.

[Ta2001] T. Tao, 254A Lecture notes 5, 2001, available at `www.math.ucla.edu/\%7Etao/254a.1.01w/notes5.ps`

[Ta2004] T. Tao, *On the asymptotic behavior of large radial data for a focusing nonlinear Schrödinger equation*, Dynamics of PDE **1** (2004), 1–48.

[Ta2005] T. Tao, *An uncertainty principle for cyclic groups of prime order*, Math. Res. Lett. **12** (2005), no. 1, 121–127.

[Ta2006] T. Tao, *The dichotomy between structure and randomness, arithmetic progressions, and the primes*, 2006 ICM proceedings, Vol. I., 581–608.

[Ta2006a] T. Tao, *The dichotomy between structure and randomness*, unpublished slides, available at `www.math.ucla.edu/\~tao/preprints/Slides/icmslides2.pdf`

[Ta2006b] T. Tao, *Santaló's inequality*, available at `www.math.ucla.edu/\~tao/preprints/Expository/santalo.dvi`

[Ta2006c] T. Tao, *Perelmans proof of the Poincare conjecture a nonlinear PDE perspective*, unpublished. Available at `arxiv.org/abs/math.DG/0610903`

[Ta2006d] T. Tao, *Nonlinear dispersive equations: local and global analysis*, CBMS regional series in mathematics, 2006

[Ta2006e] T. Tao, *Arithmetic progressions and the primes*, Collectanea Mathematica (2006), Vol. Extra., 37-88.

[Ta2006f] T. Tao, *Global behaviour of nonlinear dispersive and wave equations*, Current Developments in Mathematics 2006, International Press. 255-340.

[Ta2006g] T. Tao, *The Gaussian primes contain arbitrarily shaped constellations*, J. dAnalyse Mathematique **99** (2006), 109–176.

[Ta2006h] T. Tao, *Szemerédi's regularity lemma revisited*, Contrib. Discrete Math. **1** (2006), no. 1, 8–28.

[Ta2007] T. Tao, *Global regularity for a logarithmically supercritical defocusing nonlinear wave equation for spherically symmetric data*, J. Hyperbolic Diff. Eq. **4** (2007), 259–266.

[Ta2007b] T. Tao, *The ergodic and combinatorial approaches to Szemerédi's theorem*, Centre de Recerches Mathématiques, CRM Proceedings and Lecture Notes Vol. **43** (2007), 145–193.

[Ta2007c] T. Tao, *What is good mathematics?*, Mathematical Perspectives, Bull. Amer. Math. Soc. **44** (2007), 623–634.

[Ta2007d] T. Tao, *A (concentration-)compact attractor for high-dimensional nonlinear Schrödinger equations*, Dynamics of PDE **4** (2007), 1–53.

[Ta2007e] T. Tao, *A correspondence principle between (hyper)graph theory and probability theory, and the (hyper)graph removal lemma*, J. dAnalyse Mathematique **103** (2007), 1–45.

[Ta2007f] 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, *Product set estimates in noncommutative groups*, preprint.

[Ta2008b] T. Tao, *A quantitative formulation of the global regularity problem for the periodic Navier-Stokes equation*, preprint.

[Ta2008c] T. Tao, *A quantitative version of the Besicovitch projection theorem via multiscale analysis*, preprint.

[TaViZh2008] T. Tao, M. Visan, and X. Zhang, *Global well-posedness and scattering for the mass-critical nonlinear Schrödinger equation for radial data in high dimensions,* to appear in Duke Math. J.

[TaViZh2008b] T. Tao, M. Visan, X. Zhang, *Minimal-mass blowup solutions of the mass-critical NLS,* to appear in Forum Math.

[TaVu2006] T. Tao, V. Vu, *Additive Combinatorics,* Cambridge University Press, 2006.

[TaVu2007] T. Tao, V. Vu, *On the singularity probability of random Bernoulli matrices*, J. Amer. Math. Soc. **20** (2007), 603–628.

[TaVu2008] T. Tao, V. Vu, *Random Matrices: The circular Law*, Communications in Contemporary Mathematics, **10** (2008), 261–307.

[TaZi2008] T. Tao, T. Ziegler, *The primes contain arbitrarily long polynomial progressions*, to appear, Acta Math.

[ThYo2008] H. Thomas, A. Yong, *An S_3-symmetric Littlewood-Richardson rule*, preprint.

[Ti1930] E. Titchmarsh, *A divisor problem*, Rend. Circ. Math. Palermo **54**(1930), 414–429.

[Tr2006] J. Tropp, *Just relax: Convex programming methods for identifying sparse signals*, IEEE Trans. Info. Theory **51** (2006), 1030–1051.

[Tu1941] P. Turán, *Egy gráfelméleti szélsőértékfeladatról*, Mat Fiz Lapok **48** (1941), 436–452.

[vdC1939] J.G. van der Corput, *Über Summen von Primzahlen und Primzahlquadraten,* Math. Ann. **116** (1939), 1–50.

[Vi1937] I. M. Vinogradov, *Some theorems concerning the primes,* Mat. Sbornik. N.S. **2** (1937), 179–195.

[Vi2007] M. Visan, *The defocusing energy-critical nonlinear Schrödinger equation in higher dimensions,* Duke Math. J. **138** (2007), 281-374.

[Wa1936] A. Walfisz, *Zur additiven Zahlentheorie. II*, Math. Z. **40** (1936), no. 1, 592–607.

[Wi2008] A. Wigderson, *The power and weakness of randomness (when you are short on time)*, slides, available at `www.math.ias.edu/\~{}avi/TALKS/LATIN.ppt`

[Wo1999] T. Wolff, *Recent work connected with the Kakeya problem*. Prospects in mathematics (Princeton, NJ, 1996), 129–162, Amer. Math. Soc., Providence, RI, 1999.

[Ze1990] S. Zelditch, *Quantum transition amplitudes for ergodic and for completely integrable systems*, J. Funct. Anal. **94** (1990), no. 2, 415–436.

[Ze2004] S. Zelditch, *Note on quantum unique ergodicity*, Proc. Amer. Math. Soc. **132** (2004), 1869–1872.

[Zi2007] T. Ziegler, *Universal characteristic factors and Furstenberg averages*, J. Amer. Math. Soc. **20** (2007), no. 1, 53–97.