# Compactness and contradiction

Terence Tao

Department of Mathematics, UCLA, Los Angeles, CA 90095
*E-mail address*: tao@math.ucla.edu

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

# Contents

# Preface

In February of 2007, I converted my "What's new" web page of research updates into a blog at `terrytao.wordpress.com`. This blog has since grown and evolved to cover a wide variety of mathematical topics, ranging from my own research updates, to lectures and guest posts by other mathematicians, to open problems, to class lecture notes, to expository articles at both basic and advanced levels. In 2010, I also started writing shorter mathematical articles on my Google Buzz feed at

`profiles.google.com/114134834346472219368/buzz` .

This book collects some selected articles from both my blog and my Buzz feed from 2010, continuing a series of previous books [**Ta2008**], [**Ta2009**], [**Ta2009b**], [**Ta2010**], [**Ta2010b**], [**Ta2011**], [**Ta2011b**], [**Ta2011c**] based on the blog.

The articles here are only loosely connected to each other, although many of them share common themes (such as the titular use of *compactness and contradiction* to connect finitary and infinitary mathematics to each other). I have grouped them loosely by the general area of mathematics they pertain to, although the dividing lines between these areas is somewhat blurry, and some articles arguably span more than one category. Each chapter is roughly organised in increasing order of length and complexity (in particular, the first half of each chapter is mostly devoted to the shorter articles from my Buzz feed, with the second half comprising the longer articles from my blog.

### A remark on notation

For reasons of space, we will not be able to define every single mathematical term that we use in this book. If a term is italicised for reasons other than

emphasis or for definition, then it denotes a standard mathematical object, result, or concept, which can be easily looked up in any number of references. (In the blog version of the book, many of these terms were linked to their Wikipedia pages, or other on-line reference pages.)

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

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

Asymptotic notation is discussed further in Section 3.5.

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 $X$.

If $E$ is a subset of a domain $X$, we use $1_E : X \to \mathbf{R}$ to denote the *indicator function* of $X$, thus $1_E(x)$ equals 1 when $x \in E$ and 0 otherwise.

## Acknowledgments

I am greatly indebted to many readers of my blog and buzz feed, including Dan Christensen, David Corfield, Quinn Culver, Tim Gowers, Greg Graviton, Zaher Hani, Bryan Jacobs, Bo Jacoby, Sune Kristian Jakobsen, Allen Knutson, Ulrich Kohlenbach, Mark Meckes, David Milovich, Timothy Nguyen, Michael Nielsen, Anthony Quas, Pedro Lauridsen Ribeiro, Jason Rute, Américo Tavares, Willie Wong, Qiaochu Yuan, Pavel Zorin, and several anonymous commenters, for corrections and other comments, which can be viewed online at

`terrytao.wordpress.com`

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

*Chapter 1*

# Logic and foundations

## 1.1. Material implication

The *material implication* "If $A$, then $B$" (or "$A$ implies $B$") can be thought of as the assertion "$B$ is *at least as true as* $A$" (or equivalently, "$A$ is *at most as true as* $B$"). This perspective sheds light on several facts about the material implication:

(1) *A falsehood implies anything* (the *principle of explosion*). Indeed, any statement $B$ is at least as true as a falsehood. By the same token, if the hypothesis of an implication fails, this reveals nothing about the conclusion.

(2) *Anything implies a truth.* In particular, if the conclusion of an implication is true, this reveals nothing about the hypothesis.

(3) *Proofs by contradiction.* If $A$ is at most as true as a falsehood, then it is false.

(4) *Taking contrapositives.* If $B$ is at least as true as $A$, then $A$ is at least as false as $B$.

(5) *"If and only if" is the same as logical equivalence.* "$A$ if and only if $B$" means that $A$ and $B$ are *equally true*.

(6) *Disjunction elimination.* Given "If $A$, then $C$" and "If $B$, then $C$", we can deduce "If ($A$ or $B$), then $C$", since if $C$ is at least as true as $A$, and at least as true as $B$, then it is at least as true as either $A$ or $B$.

(7) *The principle of mathematical induction.* If $P(0)$ is true, and each $P(n+1)$ is at least as true as $P(n)$, then all of the $P(n)$ are true. (Note, though, that if one is only 99% certain of each implication

"$P(n)$ implies $P(n+1)$", then the chain of deductions can break down fairly quickly. It is thus dangerous to apply mathematical induction outside of rigorous mathematical settings. See also Section 6.9 for further discussion.)

(8) *Material implication is not causal.* The material implication "Is $A$, then $B$" is a statement purely about the truth values of $A$ and $B$, and can hold even if there is no causal link between $A$ and $B$. (e.g. "If $1+1=2$, then Fermat's last theorem is true.".)

## 1.2. Errors in mathematical proofs

Formally, a mathematical proof consists of a sequence of mathematical statements and deductions (e.g. "If $A$, then $B$"), strung together in a logical fashion to create a conclusion. A simple example of this is a linear chain of deductions, such as $A \implies B \implies C \implies D \implies E$, to create the conclusion $A \implies E$. In practice, though, proofs tend to be more complicated than a linear chain, often acquiring a tree-like structure (or more generally, the structure of a directed acyclic graph), due to the need to branch into cases, or to reuse a hypothesis multiple times. Proof methods such as proof by contradiction, or proof by induction, can lead to even more intricate loops and reversals in a mathematical argument.

Unfortunately, not all proposed proofs of a statement in mathematics are actually correct, and so some effort needs to be put into verification of such a proposed proof. Broadly speaking, there are two ways that one can show that a proof can fail. Firstly, one can find a "local", "low-level" or "direct" objection to the proof, by showing that one of the steps (or perhaps a cluster of steps, see below) in the proof is invalid. For instance, if the implication $C \implies D$ is false, then the above proposed proof $A \implies B \implies C \implies D \implies E$ of $A \implies E$ is invalid (though it is of course still conceivable that $A \implies E$ could be proven by some other route).

Sometimes, a low-level error cannot be localised to a single step, but rather to a cluster of steps. For instance, if one has a circular argument, in which a statement $A$ is claimed using $B$ as justification, and $B$ is then claimed using $A$ as justification, then it is possible for both implications $A->B$ and $B->A$ to be true, while the deduction that $A$ and $B$ are then both true remains invalid[1].

Another example of a low-level error that is not localisable to a single step arises from ambiguity. Suppose that one is claiming that $A \implies B$ and $B \implies C$, and thus that $A \implies C$. If all terms are unambiguously

[1]Note though that there are important and valid examples of *near*-circular arguments, such as proofs by induction, but this is not the topic of my discussion today.

well-defined, this is a valid deduction. But suppose that the expression $B$ is ambiguous, and actually has at least two distinct interpretations, say $B_1$ and $B_2$. Suppose further that the $A \implies B$ implication presumes the former interpretation $B = B_1$, while the $B \implies C$ implication presumes the latter interpretation $B = B_2$, thus we actually have $A \implies B_1$ and $B_2 \implies C$. In such a case we can no longer validly deduce that $A \implies C$ (unless of course we can show in addition that $B_1 \implies B_2$). In such a case, one cannot localise the error to either $A \implies B$ or $B \implies C$ until $B$ is defined more unambiguously. This simple example illustrates the importance of getting key terms defined precisely in a mathematical argument.

The other way to find an error in a proof is to obtain a "high-level" or "global" objection, showing that the proof, if valid, would necessarily imply a further consequence that is either known or strongly suspected to be false. The most well-known (and strongest) example of this is the *counterexample*. If one possesses a counterexample to the claim $A \implies E$, then one instantly knows that the chain of deduction $A \implies B \implies C \implies D \implies E$ must be invalid, even if one cannot immediately pinpoint where the precise error is at the local level. Thus we see that global errors can be viewed as "non-constructive" guarantees that a local error must exist somewhere.

A bit more subtly, one can argue using the structure of the proof itself. If a claim such as $A \implies E$ could be proven by a chain $A \implies B \implies C \implies D \implies E$, then this might mean that a parallel claim $A' \implies E'$ could then also be proven by a parallel chain $A' \implies B' \implies C' \implies D' \implies E'$ of logical reasoning. But if one also possesses a counterexample to $A' \implies E'$, then this implies that there is a flaw somewhere in this parallel chain, and hence (presumably) also in the original chain. Other examples of this type include proofs of some conclusion that mysteriously never use in any essential way a crucial hypothesis (e.g. proofs of the non-existence of non-trivial integer solutions to $a^n + b^n = c^n$ that mysteriously never use the hypothesis that $n$ is strictly greater than 2, or which could be trivially adapted to cover the $n = 2$ case).

While global errors are less constructive than local errors, and thus less satisfying as a "smoking gun", they tend to be significantly more robust. A local error can often be patched or worked around, especially if the proof is designed in a fault-tolerant fashion (e.g. if the proof proceeds by factoring a difficult problem into several strictly easier pieces, which are in turn factored into even simpler pieces, and so forth). But a global error tends to invalidate not only the proposed proof as it stands, but also all reasonable perturbations of that proof. For instance, a counterexample to

$A \implies E$ will automatically defeat any attempts to patch the invalid argument $A \implies B \implies C \implies D \implies E$, whereas the more local objection that $C$ does not imply $D$ could conceivably be worked around.

It is also a lot quicker to find a global error than a local error, at least if the paper adheres to established standards of mathematical writing. To find a local error in an $N$-page paper, one basically has to read a significant fraction of that paper line-by-line, whereas to find a global error it is often sufficient to skim the paper to extract the large-scale structure. This can sometimes lead to an awkward stage in the verification process when a global error has been found, but the local error predicted by the global error has not yet been located. Nevertheless, global errors are often the most serious errors of all.

It is generally good practice to try to structure a proof to be fault tolerant with respect to local errors, so that if, say, a key step in the proof of Lemma 17 fails, then the paper does not collapse completely, but contains at least some salvageable results of independent interest, or shows a reduction of the main problem to a simpler one. Global errors, by contrast, cannot really be defended against by a good choice of proof structure; instead, they require a good choice of proof strategy that anticipates global pitfalls and confronts them directly.

One last closing remark: as error-testing is the complementary exercise to proof-building, it is not surprising that the standards of rigour for the two activities are dual to each other. When one is building a proof, one is expected to adhere to the highest standards of rigour that are practical, since a single error could well collapse the entire effort. But when one is testing an argument for errors or other objections, then it is perfectly acceptable to use heuristics, hand-waving, intuition, or other non-rigorous means to locate and describe errors. This may mean that some objections to proofs are not watertight, but instead indicate that either the proof is invalid, or some accepted piece of mathematical intuition is in fact inaccurate. In some cases, it is the latter possibility that is the truth, in which case the result is deemed "paradoxical", yet true. Such objections, even if they do not invalidate the paper, are often very important for improving one's intuition about the subject.

## 1.3. Mathematical strength

The early twentieth century philosopher Ludwig Wittingstein famously argued that every mathematical theorem was a tautology, and thus all such theorems contained a trivial amount of content. There is a grain of truth to this: when a difficult mathematical problem is finally solved, it is often the case that the solution does make the original problem look significantly

easier than one had previously thought. Indeed, one could take the somewhat counter-intuitive point of view that progress in mathematics can be measured by how much of mathematics has been made trivial (or at least easier to understand than previously).

On the other hand, there is a definite sense that some mathematical theorems are "stronger" than others, even if from a strictly logical point of view they are equivalent. A theorem can be strong because its conclusions are strong, because its hypotheses (or underlying axiom system used in the proof) are weak, or for some combination of the two reasons.

What makes a theorem strong? This is not a precise, well-defined concept. But one way to measure the strength of a theorem is to test it against a class of questions and problems that the theorem is intended to assist with solving. For instance, one might gauge the strength of a theorem in analytic number theory by the size of the error terms it can give on various number-theoretic quantities; one might gauge the strength of a theorem in PDE by how large a class of initial data the theorem is applicable to, and how much control one gets on the solution as a consequence; and so forth.

All other things being equal, universal statements ("$P(x)$ is true for all $x$") are stronger than existential statements ("$P(x)$ is true for some $x$"), assuming of course that one is quantifying over a non-empty space. There are also statements of intermediate strength (e.g. "$P(x)$ is true for "many" $x$", or "$P(x)$ is true for "almost all" $x$", for suitably precise quantifications of "many" or "almost all"). In a similar vein, statements about special types of objects (e.g. about special functions) are usually not as strong as analogous statements about general types of objects (e.g. arbitrary functions in some function space), again assuming all other things are equal[2].

Asymptotic statements (e.g. statements that only have content when some parameter $N$ is "sufficiently large", or in the limit as $N$ goes to infinity) are usually not as strong as non-asymptotic statements (which have meaningful content even for fixed $N$). Again, this is assuming that all other things are equal. In a similar vein, approximate statements are not as strong as exact ones.

Statements about "easy" or well-understood objects are usually not as strong as statements about "difficult" or poorly understood objects. For instance, statements about solutions to equations over the reals tend to be much weaker than their counterparts concerning equations over the integers; results about linear operators tend to be much weaker than corresponding results about nonlinear operators; statements concerning arithmetic functions that are sensitive to prime factorisation (e.g. the Mobius function or

[2]In practice, there is often a tradeoff; to obtain more general statements, one has to weaken the conclusion.

von Mangoldt function) are usually significantly stronger than analogous statements about non-arithmetical functions (e.g. the logarithm function); and so forth.

When trying to read and understand a long and complicated proof, one useful thing to do is to take a look at the strength of various key statements inside the argument, and focus on those portions of the argument where the strength of the statements increases significantly (e.g. if statements that were only known for a few values of a variable $x$, somehow became amplified into statements that were true for many instances of $x$). Such amplifications often contain an essential trick or idea which powers the entire argument, and understanding those crucial steps often brings one much closer to understanding the argument as a whole.

By the same token, if the proof ends up being flawed, it is quite likely that at least one of the flaws will be associated with a step where the statements being made became unexpectedly stronger by a suspiciously large amount, and so one can use the strength of such statements as a way to quickly locate flaws in a dubious argument.

The notion of the strength of a statement need not be absolute, but may depend on the context. For instance, suppose one is trying to read a convoluted argument that is claiming a statement which is true in all dimensions $d$. If the argument proceeds by induction on the dimension $d$, then it is useful to adopt the perspective that any statement in dimension $d+1$ should be considered "stronger" than a statement in dimension $d$, even if this latter statement would ordinarily be viewed as a stronger statement than the former if the dimensions were equal. With this perspective, one is then motivated to look for the passages in the argument in which statements in dimension $d$ are somehow converted to statements in dimension $d+1$; and these passages are often the key to understanding the overall strategy of the argument.

See also the blog post [**Go2008**] of Gowers for further discussion of this topic.

## 1.4. Stable implications

A large part of high school algebra is focused on establishing implications which are of the form "If $A = B$, then $C = D$", or some variant thereof. (Example: "If $x^2 - 5x + 6 = 0$, then $x = 2$ or $x = 3$.")

In analysis, though, one is often more interested in a *stability* version of such implications, e.g. "If $A$ is close to $B$, then $C$ is close to $D$". Furthermore, one often wants quantitative bounds on *how* close $C$ is to $D$, in terms

of how close $A$ is to $B$. (A typical example: if $|x^2 - 5x + 6| \leq \epsilon$, how close must $x$ be to 2 or 3?)

*Hilbert's nullstellensatz* (discussed for instance at [**Ta2008**, §1.15]) can be viewed as a guarantee that every algebraic implication has a stable version, though it does not provide a particularly efficient algorithm for locating that stable version.

One way to obtain stability results explicitly is to *deconstruct* the proof of the algebraic implication, and replace each step of that implication by a stable analogue. For instance, if at some point one used an implication such as

"If $A = B$, then $AC = BC$"

then one might instead use the stable analogue

$$|AC - BC| = |A - B||C|.$$

If one used an implication such as

"If $A = B$, then $f(A) = f(B)$"

then one might instead use a stable analogue such as

$$|f(A) - f(B)| \leq K|A - B|$$

where $K$ is the *Lipschitz constant* of $f$ (or perhaps one may use other stable analogues, such as the *mean-value theorem* or the *fundamental theorem of calculus*). And so forth.

A simple example of this occurs when trying to find a stable analogue of the obvious algebraic implication

"If $A_i = B_i$ for all $i = 1, \ldots, n$, then $A_1 \ldots A_n = B_1 \ldots B_n$,"

thus schematically one is looking for an implication of the form

"If $A_i \approx B_i$ for all $i = 1, \ldots, n$, then $A_1 \ldots A_n \approx B_1 \ldots B_n$."

To do this, we recall how the algebraic implication is proven, namely by successive substitution, i.e. by concatenating the $n$ identities

$$\begin{aligned} A_1 \ldots A_n &= B_1 A_2 \ldots A_n \\ B_1 A_2 \ldots A_n &= B_1 B_2 A_3 \ldots A_n \\ &\ldots \\ &\ldots B_1 \ldots B_{n-1} A_n \qquad\qquad = B_1 \ldots B_n. \end{aligned}$$

A stable version of these identities is given by the formula

$$B_1 \ldots B_{i-1} A_i \ldots A_n - B_1 \ldots B_i A_{i+1} \ldots A_n = B_1 \ldots B_{i-1} (A_i - B_i) A_{i+1} \ldots A_n$$

for $i = 1, \ldots, n$, and so by telescoping all of these identities together we obtain

$$A_1 \ldots A_n - B_1 \ldots B_n = \sum_{i=1}^{n} B_1 \ldots B_{i-1}(A_i - B_i)A_{i+1} \ldots A_n$$

which, when combined with tools such as the triangle inequality, gives a variety of stability results of the desired form (even in situations in which the $A$'s and $B$'s do not commute). Note that this identity is also the discrete form of the product rule

$$(A_1 \ldots A_n)' = \sum_{i=1}^{n} A_1 \ldots A_{i-1} A_i' A_{i+1} \ldots A_n$$

and in fact easily supplies a proof of that rule.

## 1.5. Notational conventions

Like any other human language, mathematical notation has a number of implicit conventions which are usually not made explicit in the formal descriptions of the language. These conventions serve a useful purpose by conveying additional contextual data beyond the formal logical content of the mathematical sentences.

For instance, while in principle any symbol can be used for any type of variable, in practice individual symbols have pre-existing connotations that make it more natural to assign them to specific variable types. For instance, one usually uses $x$ to denote a real number, $z$ to denote a complex number, and $n$ to denote a natural number; a mathematical argument involving a complex number $x$, a natural number $z$, and a real number $n$ would read very strangely. For similar reasons, $x \in X$ reads a lot better than $X \in x$; sets or classes tend to "want" to be represented by upper case letters (in Roman or Greek) or script letters, while objects should be lower case or upper case letters only. The most famous example of such "typecasting" is of course the epsilon symbol in analysis; an analytical argument involving a quantity epsilon which was very large or negative would cause a lot of unnecessary cognitive dissonance. In contrast, by sticking to the conventional roles that each symbol plays, the notational structure of the argument is reinforced and made easier to remember; a reader who has temporarily forgotten the definition of, say, "$z$" in an argument can at least guess that it should be a complex number, which can assist in recalling what that definition is.

As another example from analysis, when stating an inequality such as $X < Y$ or $X > Y$, it is customary that the left-hand side represents an "unknown" that one wishes to control, and the right-hand side represents a more "known" quantity that one is better able to control; thus for instance

$x < 5$ is preferable to $5 > x$, despite the logical equivalence of the two statements. This is why analysts make a significant distinction between "upper bounds" and "lower bounds"; the two are not symmetric, because in both cases is bounding an unknown quantity by a known quantity. In a similar spirit, another convention in analysis holds that it is preferable to bound non-negative quantities rather than non-positive ones.

Continuing the above example, if the known bound $Y$ is itself a sum of several terms, e.g. $Y_1 + Y_2 + Y_3$, then it is customary to put the "main term" first and the "error terms" later; thus for instance $x < 1 + \varepsilon$ is preferable to $x < \varepsilon + 1$. By adhering to this standard convention, one conveys useful cues as to which terms are considered main terms and which ones considered error terms.

## 1.6. Abstraction

It is somewhat unintuitive, but many fields of mathematics derive their power from strategically *ignoring* (or *abstracting* away) various aspects of the problems they study, in order to better focus on the key features of such problems. For instance:

- *Analysis* often ignores the exact value of numerical quantities, and instead focuses on their order of magnitude.
- *Geometry* often ignores explicit coordinate systems or other descriptions of spaces, and instead focuses on their intrinsic properties.
- Probability studies the effects of randomness, but deliberately ignores the mechanics of how random variables are actually generated. (This is in contrast to *measure theory*, which takes the complementary point of view; see [**Ta2011c**, §1.1] for further discussion.)
- *Algebra* often ignores how objects are constructed or what they are, but focus instead on what operations can be performed on them, and what identities these operations enjoy. (This is in contrast to *representation theory*, which takes the complementary point of view.)
- *Partial differential equations* often ignores the underlying physics (or other branches of science) that gives rise to various systems of interest, and instead only focuses on the differential equations and boundary conditions of that system itself. (This is in contrast to, well, physics.)

- Modern *algebraic geometry* and its relatives often ignore the individual points or elements of a space, and instead focus on the structures and maps that can be placed on top of such spaces.
- *Topology*, famously, ignores such distinctions as that between a doughnut and a coffee cup, instead focusing on those qualities of a space that are unaffected by continuous deformation or homeomorphism.

Sometimes it is not possible to ignore all but one aspect of a problem, but must instead deal with two or more aspects simultaneously. Such problems tend to require an interdisciplinary approach, blending methods from two or more subfields of mathematics.

Another major application of abstraction in mathematics is to build a variety of formal spaces, such as completions, compactifications, quotient spaces, limit spaces, universal objects, etc.. These abstract spaces are powerful because they *reify* (make real) various concepts which previously did not make rigorous sense in the existing, more concrete spaces. For instance:

- 2 has no square root in the rationals. No problem; we pass to the metric completion of the rationals (i.e. the real numbers), and now the square root of 2 exists.
- $-1$ has no square root in the reals. No problem; we pass to the algebraic completion of the reals (i.e. the complex numbers), and now the square root of $-1$ exists.
- A sequence $x_1, x_2, \ldots$ may not have a limit in the space (or spaces) that hold the points $x_n$ in this sequence. No problem; we pass to a larger space (e.g. replacing a function space by a space of distributions), or a limit space (e.g. an ultraproduct), or a completion or compactification.
- One wants to define how "twisted" a loop is in a topological space $X$, but the space is too complicated to define a notion of "winding number" or "degree" concretely. No problem: we look at the representative of that loop in the fundamental group $\pi_1(X)$ of that space, and that measures the twist. Similarly for any number of other "obstructions", which can be reified through "abstract nonsense" machinery such as homology and cohomology.

So by working in a sufficiently abstract framework, one can reify just about anything one wants; existential issues are largely eliminated. Of course, the difficulty is now pushed elsewhere; in order to get back from the abstract world to a concrete setting, one often has to do some nontrivial amount of work. For instance, it is often difficult to show that an

element that one has constructed in some abstract algebraic space is actually non-trivial, or that a solution to a PDE that one constructs in some abstract generalised sense is actually a classical solution. But at least one no longer has to deal with the problem that the objects one is dealing with don't exist at all.

## 1.7. Circular arguments

A circular argument such as

(1) $P$ is true because $Q$ is true.

(2) $Q$ is true because $P$ is true.

is useless as it stands. *However*, if the circular argument comes with a time delay, such as

(1) For any $n$, $P(n)$ is true because $Q(n)$ is true.

(2) For any $n$, $Q(n)$ is true because $P(n-1)$ is true.

and if one can verify a base case such as $P(0)$, then the argument becomes useful; this is essentially the *principle of mathematical induction*. There are also continuous versions of this argument, known as *continuity arguments*. For instance, if $A(t)$ and $B(t)$ are continuously varying quantities depending on a parameter $t \in \mathbf{R}$, and we know that

(1) For any $t$, if $A(t) < X$, then $B(t) < Y$.

(2) For any $t$, if $B(t) < Y$, then $A(t) < X - \varepsilon$, where $\varepsilon > 0$ is independent of $t$ (locally, at least).

Then (provided one has a base case such as $A(0) < X$), one can keep $A(t)$ bounded by $X$ and $B(t)$ bounded by $Y$ for all time. This is because the continuity of $A$ provides a time delay between being less than $X - \varepsilon$, and being greater than $X$.

A variant of the continuity argument is the *bootstrap argument* or *iterative argument*. These arguments reflect the fact that in analysis, it is not always necessary to obtain a 100% correct answer the first time around. In many cases, it is enough to get an answer that is 1% closer to the truth than your previous best guess... as long as you can iterate this process indefinitely, and then pass to the limit[3].

Examples of this principle include the contraction mapping theorem, Newton's method, inversion via Neumann series, the Picard existence theorem, the inverse function theorem, the method of parametrices, the open

[3]This assumes, of course, that your initial guess was a finite distance from the truth. It is also important that the 1% gain does not dwindle to 0% prematurely, but instead remains uniform so long as one is some distance away from the truth.

mapping theorem, the density increment argument (used for instance to establish *Roth's theorem on arithmetic progressions* [**Ro1953**]), and the energy increment argument (used for instance to establish *Szemerédi's regularity lemma for graphs* [**Sz1978**]).

## 1.8. The classical number systems

In the foundations of mathematics, the standard construction of the classical number systems (the *natural numbers* $\mathbf{N}$, the *integers* $\mathbf{Z}$, the *rationals* $\mathbf{Q}$, the *reals* $\mathbf{R}$, and the *complex numbers* $\mathbf{C}$) starting from the natural numbers $\mathbf{N}$ is conceptually simple: $\mathbf{Z}$ is the additive completion of $\mathbf{N}$, $\mathbf{Q}$ is the multiplicative completion of $\mathbf{Z}$, $\mathbf{R}$ is the metric completion of $\mathbf{Q}$, and $\mathbf{C}$ is the algebraic completion of $\mathbf{R}$. But the actual technical details of the construction are lengthy and somewhat inelegant. Here is a typical instance of this construction (as given for instance in [**Ta2006c**]):

- $\mathbf{Z}$ is constructed as the space of formal differences $a - b$ of natural numbers $a$, $b$, quotiented by additive equivalence (thus $a-b \sim c-d$ iff $a + d = b + c$), with the arithmetic operations extended in a manner consistent with the laws of algebra.
- $\mathbf{Q}$ is constructed as the space of formal quotients $a/b$ of an integer $a$ and a non-zero integer $b$, quotiented by multiplicative equivalence (thus $a/b \sim c/d$ iff $ad = bc$), with the arithmetic operations extended in a manner consistent with the laws of algebra.
- $\mathbf{R}$ is constructed as the space of formal limits $\lim_{n\to\infty} a_n$ of Cauchy sequences $a_n$ of rationals, quotiented by Cauchy equivalence (thus $\lim_{n\to\infty} a_n \sim \lim_{n\to\infty} b_n$ iff $a_n - b_n$ converges to zero as $n$ goes to infinity), with the arithmetic operations extended by continuity.
- $\mathbf{C}$ is constructed as the space of formal sums $a+bi$ of two reals $a, b$, with the arithmetic operations extended in a manner consistent with the laws of algebra and the identity $i^2 = -1$.

**Remark 1.8.1.** One can also perform these completions in a different order, leading to other important number systems such as the *positive rationals* $\mathbf{Q}^+$, the *positive reals* $\mathbf{R}^+$, the *Gaussian integers* $\mathbf{Z}[i]$, the *algebraic numbers* $\bar{\mathbf{Q}}$, or the *algebraic integers* $\mathcal{O}$.)

There is just one slight problem with all this: technically, with these constructions, the natural numbers are *not* a subset of the integers, the integers are *not* a subset of the rationals, the rationals are *not* a subset of the reals, and the reals are *not* a subset of the complex numbers! For instance, with the above definitions, an integer is an equivalence class of formal differences $a - b$ of natural numbers. A natural number such as

3 is not then an integer. Instead, there is a *canonical embedding* of the natural numbers into the integers, which for instance identifies 3 with the equivalence class

$$\{3-0, 4-1, 5-2, \ldots\}.$$

Similarly for the other number systems. So, rather than having a sequence of inclusions

$$\mathbf{N} \subset \mathbf{Z} \subset \mathbf{Q} \subset \mathbf{R} \subset \mathbf{C}$$

, what we have here is a sequence of canonical embeddings

$$\mathbf{N} \hookrightarrow \mathbf{Z} \hookrightarrow \mathbf{Q} \hookrightarrow \mathbf{R} \hookrightarrow \mathbf{C}.$$

In practice, of course, this is not a problem, because we simply *identify* a natural number with its integer counterpart, and similarly for the rest of the chain of embeddings. At an ontological level, this may seem a bit messy - the number 3, for instance is now simultaneously a natural number, an equivalence class of formal differences of natural numbers, and equivalence class of formal quotients of equivalence classes of formal differences of natural numbers, and so forth; but the beauty of the axiomatic approach to mathematics is that it is almost completely irrelevant exactly how one chooses to model a mathematical object such as 3, so long as all the relevant axioms concerning one's objects are verified, and so one can ignore such questions as what a number actually *is* once the foundations of one's mathematics have been completed.

**Remark 1.8.2.** As an alternative approach, one can carefully keep all the number systems disjoint by using distinct notation for each; for instance, one could distinguish between the natural number 3, the integer +3, the rational 3/1, the real number 3.0, and the complex number $3.0 + i0.0$. This type of distinction is useful in some situations, for instance when writing mathematical computer code, but in most cases it is more convenient to collapse all these distinctions and perform the identifications mentioned above.

Another way of thinking about this is to define a (classical) number to be an element not of any one of the above number systems *per se*, but rather of the *direct limit*

$$\varinjlim(\mathbf{N} \hookrightarrow \mathbf{Z} \hookrightarrow \mathbf{Q} \hookrightarrow \mathbf{R} \hookrightarrow \mathbf{C})$$

of the canonical embeddings. Recall that the *direct limit*

$$\varinjlim(\ldots \to A_{n-1} \to A_n \to A_{n+1} \to \ldots)$$

of a sequence of sets (or objects in set-like categories, e.g. groups, vector spaces, etc.) chained together by maps (or morphisms, for more general categories) $f_n : A_n \to A_{n+1}$ is the space of sequences $(a_{n_0}, a_{n_0+1}, \ldots)$ of elements of some terminal segment $A_{n_0} \to A_{n_0+1} \to \ldots$ of the sequence of

sets, such that the sequence of elements is compatible with the maps (i.e. $a_{n+1} = f_n(a_n)$ for all $n \geq n_0$), and then quotiented by tail equivalence: two sequences $(a_{n_0}, a_{n_0+1}, \ldots)$ and $(a_{1_0}, b_{n_1+1}, \ldots)$ are equivalent iff they eventually agree (i.e. $a_n = b_n$ for all sufficiently large $n$).

**Remark 1.8.3.** Direct limits also have an elegant category-theoretic definition; the direct limit $A$ of the above sequence can be defined (up to isomorphism) as a *universal object* for the commutative diagram

$$\ldots \to A_{n-1} \to A_n \to A_{n+1} \to \ldots \to A,$$

which means that every other competitor $B$ to the direct limit (i.e. any commutative diagram of the form

$$\ldots \to A_{n-1} \to A_n \to A_{n+1} \to \ldots \to B$$

factors uniquely through $A$.

There is also an important dual notion of a direct limit, namely the *inverse limit*

$$\varprojlim(\ldots \to A_{n-1} \to A_n \to A_{n+1} \to \ldots)$$

of a sequence, which is defined similarly to the direct limits but using initial segments of the sequence rather than terminal segments. Whereas direct limits seek to build a canonical space in which all the elements of the sequence embed, inverse limits seek to build a canonical space for which all the elements of the sequence are projections. A classic example of an inverse limit is the *p-adic number system* $\mathbf{Z}_p$, which is the inverse limit of the cyclic groups $\mathbf{Z}/p^n\mathbf{Z}$. Another example is the real number system $\mathbf{R}$, which can be viewed as the inverse limit of the finite-precision number systems $10^{-n}\mathbf{Z}$ (using arithmetic operations with rounding, and using rounding to map each finite precision number system to the next coarsest system); this is a different way to construct the real numbers than the one given above, but the two constructions can eventually be shown to be equivalent.

Direct limits and inverse limits can be generalised even further; in category theory, one can often take *limits* and *colimits* of more general diagrams than sequences. This gives a rich source of constructions of abstract spaces (e.g. direct sums or direct products) that are convenient places to do mathematics in, as they can connect to many otherwise distinct classes of mathematical structures simultaneously. For instance, the *adele ring*, which is the direct product of the reals and the p-adics, is a useful universal number system in algebraic number theory, which among other things can be used to greatly clarify the nature of the functional equation of the Riemann zeta function (see e.g. [**Ta2009**, §1.5] for further discussion).

## 1.9. Round numbers

It is a convention in popular culture to use round numbers as milestones in order to reflect on the progress of some statistic, such as when a major stock index passes, say, the $10,000$ level. People often celebrate their birthdays each year, and also use the new year to make resolutions; institutions similarly observe centenaries and other round number milestones.

Of course, thanks to the artificial nature of both our system of units, and also our decimal system to express numbers, such milestones have no particular *intrinsic* significance; a day in which the Dow Jones Industrial Average, for instance, crosses $10,000$ is not intrinsically much different from a day in which the Dow Jones crosses 10764, or $2^{13}$, or any other number.

However, there is some value in selecting *some* set of milestones of a given spacing (or log-spacing) in order to set up a periodic schedule in which to focus occasional attention on a topic. For instance, it is certainly useful to spend some time occasionally reflecting on one's past and making resolutions for one's future, but one should not spend every day of one's life doing so. Instead, the optimal fraction of time that one should invest in this is probably closer to $1/365$ than to $1/1$. As such, the convention to use the first of January of each year to devote to this is not such a bad one, though of course it is ultimately a somewhat arbitrary choice.

Similarly, for the majority of people who are not professional stock traders, the daily fluctuations of an index such as the Dow Jones are too noisy to be of much informational value; but if one only pays attention to this index when it crosses a multiple[4] of 1000, then this already gives a crude picture of the direction of the market that is sufficient for a first approximation, without requiring too much of one's time to be spent looking at this index.

At a somewhat less frivolous level, one advantage of selecting a conventional set of *preferred numbers* is that it allows for easier comparison and interchangeability between people, objects, and institutions. For instance, companies who report their financial results on a quarterly basis can be easily compared to each other, as opposed to companies who report at irregular or idiosyncratic schedules. In order to have interchangeability between resistors made by different manufacturers, the resistance is by convention set to lie in a discrete set of preferred numbers that are roughly equally spaced in log-scale (and which are famously colour-coded to indicate this number).

[4]Note though that if this index changed value by an order of magnitude or more, then one should presumably replace multiples of 1000 with an appropriately rescaled multiple. Ideally one should use milestones that are equally spaced in log-scale rather than in absolute scal, but with the decimal system the round numbers that do this (i.e. the powers of 10) are too far apart to be sufficiently useful.

Many other dimensions of standard objects (e.g. dimensions of a sheet of paper, of which $A4$ is the most well-known) use some sort of preferred number scheme, leading to a useful degree of interoperability[5].

## 1.10. The "no self-defeating object" argument, revisited

One notable feature of mathematical reasoning is the reliance on counterfactual thinking - taking a hypothesis (or set of hypotheses) which may or may not be true, and following it (or them) to its logical conclusion. For instance, most propositions in mathematics start with a set of hypotheses (e.g. "Let $n$ be a natural number such that ..."), which may or may not apply to the particular value of $n$ one may have in mind. Or, if one ever argues by dividing into separate cases (e.g. "Case 1: $n$ is even. ... Case 2: $n$ is odd. ..."), then for any given $n$, at most one of these cases would actually be applicable, with the other cases being counterfactual alternatives. But the purest example of counterfactual thinking in mathematics comes when one employs a proof by contradiction[6] (or *reductio ad absurdum*) - one introduces a hypothesis that in fact has no chance of being true at all (e.g. "Suppose for sake of contradiction that $\sqrt{2}$ is equal to the ratio $p/q$ of two natural numbers."), and proceeds to demonstrate this fact by showing that this hypothesis leads to absurdity.

Experienced mathematicians are so used to this type of counterfactual thinking that it is sometimes difficult for them to realise that it this type of thinking is not automatically intuitive for students or non-mathematicians, who can anchor their thinking on the single, "real" world to the extent that they cannot easily consider hypothetical alternatives. This can lead to confused exchanges such as the following:

> Lecturer: "Theorem. Let $p$ be a prime number. Then..."
> Student: "But how do you know that $p$ is a prime number? Couldn't it be composite?"

or

> Lecturer: "Now we see what the function $f$ does when we give it the input of $x + dx$ instead. ..."

[5] In contrast, items such as tupperware are usually not fixed to preferred dimensions, leading to a frustrating lack of compatibility between lids and containers from different manufacturers.

[6] Strictly speaking, there are two types of proofs by contradiction: genuine proofs by contradiction, which proves a statement $A$ is true by showing that the negation $\neg A$ leads to absurdity, and *proofs by negation*, which proofs a statement $A$ is false by showing that $A$ leads to absurdity. In classical logic, which enjoys the *law of excluded middle*, the two types of argument are logically equivalent, but in other logics, such as intuitionistic logic, the two types of arguments need to be carefully distinguished. However, this distinction is somewhat orthogonal to the discussion in this article. We thank Andrej Bauer for emphasising this point.

> Student: "But didn't you just say that the input was equal to $x$ just a moment ago?"

This is not to say that counterfactual thinking is not encountered at all outside of mathematics. For instance, an obvious source of counterfactual thinking occurs in fictional writing or film, particularly in speculative fiction such as science fiction, fantasy, or alternate history. Here, one can certainly take one or more counterfactual hypotheses (e.g. "what if magic really existed?") and follow them to see what conclusions would result. The analogy between this and mathematical counterfactual reasoning is not perfect, of course: in fiction, consequences are usually not logically entailed by their premises, but are instead driven by more contingent considerations, such as the need to advance the plot, to entertain or emotionally affect the reader, or to make some moral or ideological point, and these types of narrative elements are almost completely absent in mathematical writing. Nevertheless, the analogy can be somewhat helpful when one is first coming to terms with mathematical reasoning. For instance, the mathematical concept of a proof by contradiction can be viewed as roughly analogous in some ways to such literary concepts as satire, dark humour, or absurdist fiction, in which one takes a premise specifically with the intent to derive absurd consequences from it. And if the proof of (say) a lemma is analogous to a short story, then the statement of that lemma can be viewed as analogous to the moral of that story.

Another source of counterfactual thinking outside of mathematics comes from *simulation*, when one feeds some initial data or hypotheses (that may or may not correspond to what actually happens in the real world) into a simulated environment (e.g. a piece of computer software, a laboratory experiment, or even just a thought-experiment), and then runs the simulation to see what consequences result from these hypotheses. Here, proof by contradiction is roughly analogous to the "garbage in, garbage out" phenomenon that is familiar to anyone who has worked with computers: if one's initial inputs to a simulation are not consistent with the hypotheses of that simulation, or with each other, one can obtain bizarrely illogical (and sometimes unintentionally amusing) outputs as a result; and conversely, such outputs can be used to detect and diagnose problems with the data, hypotheses, or implementation of the simulation.

A final example of counterfactual thinking in everyday experience is that of law[7]; any case involving damages, for instance, will often need to consider a hypothetical world in which the criminal act did not occur, in order to

[7] I thank David Tweed for this example

compare the actual world against. In a similar spirit, an adversarial cross-examination, designed to poke holes in an alibi, can be viewed as roughly analogous to a proof by contradiction.

Despite the presence of these non-mathematical analogies, though, proofs by contradiction are still often viewed with suspicion and unease by many students of mathematics. Perhaps the quintessential example of this is the standard proof of Cantor's theorem that the set $\mathbf{R}$ of real numbers is uncountable. This is about as short and as elegant a proof by contradiction as one can have without being utterly trivial, and despite this (or perhaps because of this) it seems to offend the reason of many people when they are first exposed to it, to an extent far greater than most other results in mathematics[8].

In [**Ta2010b**, §1.15], I collected a family of well-known results in mathematics that were proven by contradiction, and specifically by a type of argument that I called the "no self-defeating object" argument; that any object that was so ridiculously overpowered that it could be used to "defeat" its own existence, could not actually exist. Many basic results in mathematics can be phrased in this manner: not only Cantor's theorem, but Euclid's theorem on the infinitude of primes, Gdel's incompleteness theorem, or the conclusion (from Russell's paradox) that the class of all sets cannot itself be a set.

In [**Ta2010b**, §1.15] each of these arguments was presented in the usual "proof by contradiction" manner; I made the counterfactual hypothesis that the impossibly overpowered object existed, and then used this to eventually derive a contradiction. Mathematically, there is nothing wrong with this reasoning, but because the argument spends almost its entire duration inside the bizarre counterfactual universe caused by an impossible hypothesis, readers who are not experienced with counterfactual thinking may view these arguments with unease.

It was pointed out to me, though (originally with regards to Euclid's theorem, but the same point in fact applies to the other results I presented) that one can pull a large fraction of each argument out of this counterfactual world, so that one can see most of the argument directly, without the need for any intrinsically impossible hypotheses. This is done by converting the "no self-defeating object" argument into a logically equivalent "any object can be defeated" argument, with the former then being viewed as an immediate corollary of the latter. This change is almost trivial to enact (it is often little more than just taking the contrapositive of the original statement),

[8]The only other two examples I know of that come close to doing this are the fact that the real number $0.999\ldots$ is equal to 1, and the solution to the blue-eyed islanders puzzle (see [**Ta2009**, §1.1]).

but it does offer a slightly different "non-counterfactual" (or more precisely, "not necessarily counterfactual") perspective on these arguments which may assist in understanding how they work.

For instance, consider the very first no-self-defeating result presented in [**Ta2010**, §1.15]:

**Proposition 1.10.1** (No largest natural number)**.** *There does not exist a natural number $N$ that is larger than all the other natural numbers.*

This is formulated in the "no self-defeating object" formulation. But it has a logically equivalent "any object can be defeated" form:

**Proposition 1.10.2.** *Given any natural number $N$, one can find another natural number $N'$ which is larger than $N$.*

**Proof.** Take $N' := N + 1$. □

While Proposition 1.10.1 and Proposition 1.10.2 are logically equivalent to each other, note one key difference: Proposition 1.10.2 can be illustrated with examples (e.g. take $N = 100$, so that the proof gives $N' = 101$ ), whilst Proposition 1.10.1 cannot (since there is, after all, no such thing as a largest natural number). So there is a sense in which Proposition 1.10.2 is more "constructive" or "non-counterfactual" than Proposition 1.10.1.

In a similar spirit, *Euclid's theorem,*

**Proposition 1.10.3** (Euclid's theorem)**.** *There are infinitely many primes.*

can be recast in "all objects can be defeated" form as

**Proposition 1.10.4.** *Let $p_1, \ldots, p_n$ be a collection of primes. Then there exists a prime $q$ which is distinct from any of the primes $p_1, \ldots, p_n$.*

**Proof.** Take $q$ to be any prime factor of $p_1 \ldots p_n + 1$ (for instance, one could take the smallest prime factor, if one wished to be completely concrete). Since $p_1 \ldots p_n + 1$ is not divisible by any of the primes $p_1, \ldots, p_n$, $q$ must be distinct from all of these primes. □

One could argue that there was a slight use of proof by contradiction in the proof of Proposition 1.10.4 (because one had to briefly entertain and then rule out the counterfactual possibility that $q$ was equal to one of the $p_1, \ldots, p_n$), but the proposition itself is not inherently counterfactual, as it does not make as patently impossible a hypothesis as a finite enumeration of the primes. Incidentally, it can be argued that the proof of Proposition 1.10.4 is closer in spirit to Euclid's original proof of his theorem, than the proof of Proposition 1.10.3 that is usually given today. Again, Proposition 1.10.4 is "constructive"; one can apply it to any finite list of primes, say

$2, 3, 5$, and it will actually exhibit a prime not in that list (in this case, 31). The same cannot be said of Proposition 1.10.3, despite the logical equivalence of the two statements.

**Remark 1.10.5.** It is best to avoid long proofs by contradiction which consist of many parts, each of which is different to convert to a non-counterfactual form. One sees this sometimes in attempts by amateurs to prove, say, the Riemann hypothesis; one starts with assuming a zero off of the critical line, and then derives a large number of random statements both using this fact, and not using this fact. At some point, one makes an error (e.g. division by zero), but one does not notice it until several pages later, when two of the equations derived disagree with each other. At this point, the author triumphantly declares victory.

On the other hand, there are many valid long proofs by contradiction in the literature. For instance, in PDE, a common and powerful way to show that a solution to an evolution equation exists for all time is to assume that it doesnt, and deduce that it must develop a singularity somewhere. One then applies an increasingly sophisticated sequence of analytical tools to control and understand this singularity, until one eventually is able to show that the singularity has an impossible nature (e.g. some limiting asymptotic profile of this singularity has a positive norm in one function space, and a zero norm in another). In many cases this is the only known way to obtain a global regularity result, but most of the proof is left in the counterfactual world where singularities exist. But, the use of contradiction is often "shallow", in that large parts of the proof can be converted into non-counterfactual form (and indeed, if one looks at the actual proof, it is usually the case that most of the key lemmas and sub-propositions in the proof are stated non-counterfactually). In fact, there are two closely related arguments in PDE, known as the "no minimal energy blowup solution" argument, and the "induction on energy" argument, which are related to each other in much the same way as the well-ordering principle is to the principle of mathematical induction (or the way that the "no self-defeating object" argument is related to the "every object can be defeated" argument); the former is counterfactual and significantly simpler, the latter is not but requires much lengthier and messier arguments. But it is generally accepted that the two methods are, on some level, equivalent. (See [**KiVa2008**] for further discussion of these arguments.)

As an analogy, one can think of a long proof as a long rope connecting point $A$ (the hypotheses) to point $B$ (the conclusion). This rope may be submerged in murky water (the counterfactual world) or held up above it (the non-counterfactual world). A proof by contradiction thus is like a rope that is almost completely submerged underwater, but as long as the rope is

only shallowly underwater, one can still see it well enough to conclude that it is unbroken. But if it sinks too far into the depths of the counterfactual world, then it becomes suspicious.

Finding a non-counterfactual formulation of an argument then resembles the action of lifting the rope up so that it is mostly above water (though, if either the hypothesis $A$ or conclusion $B$ are negative in nature (and thus underwater), one must still spend a tiny fraction of the argument at least in the counterfactual world; also, the proof of the non-counterfactual statement may still occasionally use contradiction, so the rope may still dip below the water now and then). This can make the argument clearer, but it is also a bit tiring to lift the entire rope up this way; if ones objective is simply to connect $A$ to $B$ in the quickest way possible, letting the rope slide underwater is often the simplest solution.

**1.10.1. Set theory.** Now we revisit examples of the no-self-defeating object in set theory. Take, for instance, *Cantor's theorem*:

**Proposition 1.10.6** (Cantor's theorem)**.** *The reals are uncountable.*

One can easily recast this in a "non-counterfactual" or "all objects can be defeated" form:

**Proposition 1.10.7.** *Let $x_1, x_2, x_3, \ldots$ be a countable sequence of real numbers (possibly with repetition). Then there exists a real number $y$ that is not equal to any of the $x_1, x_2, x_3, \ldots$.*

**Proof.** Set $y$ equal to $y = 0.a_1a_2a_3\ldots$, where

(1) $a_1$ is the smallest digit in $\{0, \ldots, 9\}$ that is not equal to the first digit past the decimal point of any[9] decimal representation of $x_1$;

(2) $a_2$ is the smallest digit in $\{0, \ldots, 9\}$ that is not equal to the second digit past the decimal point of any decimal representation of $x_2$;

(3) etc.

Note that any real number has at most two decimal representations, and there are ten digits available, so one can always find $a_1, a_2, \ldots$ with the desired properties. Then, by construction, the real number $y$ cannot equal $x_1$ (because it differs in the first digit from any of the decimal representations of $x_1$), it cannot equal $x_2$ (because it differs in the second digit), and so forth. □

[9]Here we write "any" decimal representation rather than "the" decimal representation to deal with the annoying $0.999\ldots = 1.000\ldots$ issue mentioned earlier. As with the proof of Euclid's theorem, there is nothing special about taking the smallest digit here; this is just for sake of concreteness.

Again, Proposition 1.10.7 is trivially equivalent to Proposition 1.10.6, and still briefly uses contradiction in its proof, but in this non-counterfactual form one can actually illustrate it with examples. For instance, one could start with a list of terminating decimals, starting with the single digit terminating decimals, the two-digit terminating decimals, and so forth:

$$\begin{aligned} x_1 &:= 0.1, \quad x_2 := 0.2, \quad x_3 := 0.3, \quad \dots \quad x_9 := 0.9, \\ x_{10} &:= 0.01, \quad x_{11} := 0.02, \quad \dots \quad x_{108} := 0.99, \\ x_{109} &:= 0.001, \quad x_{110} := 0.002, \quad \dots \quad x_{1007} := 0.999, \\ &\dots \end{aligned}$$

and one then sees that the construction will, in this case, give the number $y = 0.21111\dots$, which indeed does not occur on the above list.

It is instructive to try to "outrun" Proposition 1.10.7 by modifying the list to accommodate $0.2111\dots$ to the list. One cannot simply tack on this number "at the end" of this list, as the list is infinite and does not actually have an end. One can insert it at, say, the beginning of the list, and then move all the other numbers down one, but then Proposition 1.10.7 gives a new number not on the list (in this case, $0.0111\dots$). One can add that number to the list also, bumping everyone else down one, but then Proposition 1.10.7 gives yet another number not on the list (in this case, $0.10111\dots$). After doing this a few times, one can begin to appreciate how Proposition 1.10.7 always defeats any attempt to outrun it, much as one cannot obtain a largest natural number by continually adding $+1$ to one's previous proposed candidate.

It is also remarkable how inoffensive Proposition 1.10.7 and its proof is, when compared against the reaction one sometimes encounters to Proposition 1.10.6, which is logically equivalent. A single contraposition can dramatically change one's impression of a result.

In a similar spirit, the result

**Proposition 1.10.8** (No universal set)**.** *There does not exist a set that contains all sets (including itself).*

(which, of course, assumes one is working in something like the *Zermelo-Frankel axioms* of set theory) becomes

**Proposition 1.10.9.** *Given any set $A$, there exists a set $B$ which is not an element of $A$.*

**Proof.** Consider the set

$$B := \{C \in A : C \notin C\};$$

the existence of this set is guaranteed by the *axiom schema of specification*. If $B$ was an element of itself, then by construction we would have $B \notin B$, a contradiction. Thus we must have $B \notin B$. From construction, this forces $B \notin A$. □

In the usual axiomatic formulation of set theory, the axiom of foundation implies, among other things, that no set is an element of itself. With that axiom, the set $B$ given by Proposition 1.10.9 is nothing other than $A$ itself, which by the axiom of foundation is not an element of $A$. But since the axiom of foundation was not used in the proof of Proposition 1.10.9, one can also explore (counterfactually!) what happens in set theories in which one does not assume the axiom of foundation. Suppose, for instance, that one managed somehow to produce a set $A$ that contained itself[10] as its only element: $A = \{A\}$. Then the only element that $A$ has, namely $A$, is an element of itself, so the set $B$ produced by Proposition 1.10.9 is the empty set $B := \emptyset$, which is indeed not in $A$.

One can try outrunning Proposition 1.10.9 again to see what happens. For instance, let's add the empty set to the set $A$ produced earlier, to give the new set $A' := \{A, \emptyset\}$. The construction used to prove Proposition 1.10.9 then gives the set $B = \{\emptyset\}$, which is indeed not in $A'$. If we then try to add that set in to get a new set $A'' := \{A, \emptyset, \{\emptyset\}\}$, then one gets the set $B = \{\emptyset, \{\emptyset\}\}$, which is again not in $A''$. Iterating this, one in fact begins constructing the[11] *von Neumann ordinals*.

**1.10.2. Logic.** One can also convert the no-self-defeating arguments given in the logic section of the previous post into "every object can be defeated" forms, though these were more difficult for me to locate. We first turn to the result (essentially coming from the liar paradox) that the notion of truth cannot be captured by a predicate. We begin with the easier "self-referential case":

**Theorem 1.10.10** (Impredicativity of truth, self-referential case)**.** *(Informal statement) Let $L$ be a formal language that contains the concepts of predicates and allows self-reference, and let $M$ be an interpretation of that language (i.e. a way of consistently assigning values to every constant, ranges to every variable, and truth values to every sentence in that language, obeying all the axioms of that language). Then there does not exist a "truth predicate" $T(x)$ in $L$ that takes a sentence $x$ as input, with the property that for every sentence $x$ in $L$, that $T(x)$ is true (in $M$) if and only if $x$ is true (in $M$).*

[10] Informally, one could think of $A$ as an infinite nested chain, $A := \{\{\{\ldots\}\}\}$.

[11] Actually, the original set $A$ plays essentially no role in this construction; one could have started with the empty set and it would have generated the same sequence of ordinals.

Here is the non-counterfactual version:

**Theorem 1.10.11.** *(Informal statement) Let $L$ be a formal language that contains the concepts of predicates and strings and allows self-reference, and let $M$ be an interpretation of that language. Let $T()$ be a predicate in $L$ that takes sentences as input. Then there exists a sentence $G$ such that the truth value of $T(G)$ (in $M$) is different from the truth value of $G$ (in $M$).*

**Proof.** We define $G$ be the self-referential "liar sentence"

$$G := \text{“}T(G) \text{ is false”}.$$

Then, clearly, $G$ is true if and only if $T(G)$ is false, and the claim follows. □

Using the *Quining trick* to achieve indirect self-reference, one can remove the need for direct self-reference in the above argument:

**Theorem 1.10.12** (Impredicativity of truth)**.** *(Informal statement) Let $L$ be a formal language that contains the concepts of predicates and strings, and let $M$ be an interpretation of that language (i.e. a way of consistently assigning values to every constant, ranges to every variable, and truth values to every sentence in that language) that interprets strings in the standard manner (so in particular, every sentence or predicate in $L$ can also be viewed as a string constant in $M$). Then there does not exist a "truth predicate" $T(x)$ in $L$ that takes a string $x$ as input, with the property that for every sentence $x$ in $L$, that $T(x)$ is true (in $M$) if and only if $x$ is true (in $M$).*

**Remark 1.10.13.** A more formal version of the above theorem is given by *Tarski's undefinability theorem*, which can be found in any graduate text on logic.

Here is the non-counterfactual version:

**Theorem 1.10.14.** *(Informal statement) Let $L$ be a formal language containing the concepts of predicates and strings, and let $T(x)$ be a predicate on strings. Then there exists a sentence $G$ in $L$ with the property that, for any interpretation $M$ of $L$ that interprets strings in the standard manner,that the truth value of $T(G)$ in $M$ is different from the truth value of $G$ in $M$.*

**Proof.** (Sketch) We use the "quining" trick. Let $Q(x)$ be the predicate on strings defined by

$$Q(x) := \text{“}x \text{ is a predicate on strings, and } T(x(x)) \text{ is false”}$$

and let $G$ be the *Gödel sentence* $G := Q(Q)$. Then, by construction, $G$ is true in $M$ if and only if $T(G)$ is false in $M$, and the claim follows. □

Actually Theorem 1.10.14 is marginally stronger than Theorem 1.10.6 because it makes the sentence $G$ independent of the interpretation $M$, whereas Theorem 1.10.12 (when viewed in the contrapositive) allows $G$ to depend on $M$. This slight strengthening will be useful shortly.

An important special case of Theorem 1.10.14 is the *first incompleteness theorem*:

**Corollary 1.10.15** (Gödel's first incompleteness theorem)**.** *(Informal statement) Let $L$ be a formal language containing the concepts of predicates and strings that has at least one interpretation $M$ that gives the standard interpretation of strings (in particular, $L$ must be consistent). Then there exists a sentence $G$ in $L$ that is undecidable in $L$ (or more precisely, in a formal recursive proof system for $L$).*

**Proof.** (Sketch) A language $L$ that is powerful enough to contain predicates and strings will also be able to contain a provability predicate $P()$, so that a sentence $x$ in $L$ is provable in $L$'s proof system if and only if $P(x)$ is true in the standard interpretation $M$. Applying Theorem 1.10.14 to this predicate, we obtain a Gödel sentence $G$ such that the truth value of $G$ in $M$ differs from the truth value of $P(G)$ in $M$. If $P(G)$ is true in $M$, then $G$ must be true in $M$ also since $L$ is consistent, so the only remaining option is that $P(G)$ is false in $M$ and $G$ is true in $M$. Thus neither $G$ nor its negation can be provable, and hence $G$ is undecidable. □

Now we turn to the second incompleteness theorem:

**Theorem 1.10.16** (Gödel's second incompleteness theorem)**.** *(Informal statement) No consistent logical system which has the notion of a predicate and a string, can provide a proof of its own logical consistency.*

Here is the non-counterfactual version:

**Theorem 1.10.17.** *(Informal statement) Let $L, L'$ be consistent logical systems that have the notion of a predicate and a string, such that every sentence in $L'$ is also a sentence in $L$, and such that the consistency of $L'$ can be proven in $L$. Then there exists a sentence $G$ that lies in both $L$ and $L'$ that is provable in $L$ but is not provable in $L'$.*

**Proof.** (Sketch) In the common language of $L$ and $L'$, let $T()$ be the predicate

$$T(x) := \text{“}x \text{ is provable in } L'\text{”}.$$

Applying Theorem 1.10.14, we can find a sentence $G$ (common to both $L$ and $L'$) with the property that in any interpretation $M$ of either $L$ or $L'$, the truth value of $G$ and the truth value of $T(G)$ differ.

By Corollary 1.10.15 (or more precisely, the proof of that corollary), $G$ is not provable in $L'$. Now we show that $G$ is provable in $L$. Because $L$ can prove the consistency of $L'$, one can embed the *proof* of Corollary 1.10.15 inside the language $L$, and deduce that the sentence "$G$ is not provable in $L'$" is also provable in $L$. In other words, $L$ can prove that $T(G)$ is false. On the other hand, embedding the proof of Theorem 1.10.14 inside $L$, $L$ can also prove that the truth value of $G$ and $T(G)$ differ. Thus $L$ can prove that $G$ is true. □

The advantage of this formulation of the second incompleteness theorem, as opposed to the usual counterfactual one, is that one can actually trace through the argument with a concrete example. For instance, Zermelo-Frankel-Choice (ZFC) set theory can prove the consistency of Peano arithmetic (a result of Gentzen [**Ge1936**]), and so one can follow the above argument to show that the Gödel sentence of Peano arithmetic is provably true in ZFC, but not provable in Peano arithmetic.

**1.10.3. Computability.** By now, it should not be surprising that the no-self-defeating arguments in computability also have a non-counterfactual form, given how close they are to the analogous arguments in set theory and logic. For sake of completeness, we record this for *Turing's theorem*:

**Theorem 1.10.18** (Turing halting theorem)**.** *(Informal statement) There does not exist a program $P$ which takes a string $S$ as input, and determines in finite time whether $S$ is a program (with no input) that halts in finite time.*

Here is the non-counterfactual version:

**Theorem 1.10.19.** *(Informal statement) Let $P$ be a program that takes a string $S$ as input, returns a yes-no answer $P(S)$ as output, and which always halts in finite time. Then there exists a string $G$ that is a program with no input, such that if $P$ is given $G$ as input, then $P$ does not determine correctly whether $G$ halts in finite time.*

**Proof.** Define $Q()$ to be the program taking a string $R$ as input which does the following:

(1) If $R$ is not a program that takes a string as input, it halts.
(2) Otherwise, it runs $P$ with input $R(R)$ (which is a program with no input).
(3) If $P(R(R))$ returns "no", it halts, while if $P(R(R))$ returns "yes", it runs forever.

Now, let $G$ be the program $Q(Q)$. By construction, $G$ halts if and only if $P(G)$ returns "no", and the claim follows. □

One can apply Theorem 1.10.19 to various naive halting algorithms. For instance, let $P(S)$ be the program that simulates $S$ for (say) 1000 CPU cycles, and then returns "yes" if $S$ halted by that time, or "no" otherwise. Then the program $G$ generated by the above proof will take more than 1000 CPU cycles to execute, and so $P$ will determine incorrectly whether $G$ halted or not. (Notice the similarity here with Proposition 1.10.2.)

The same argument also gives a non-counterfactual version of the non-computability of the *busy beaver function*:

**Proposition 1.10.20.** *Let $f : \mathbf{N} \to \mathbf{N}$ be a computable function. Then there exists a natural number $n$ and a program $G$ of length $n$ (and taking no input) that halts in finite time, but requires more than $f(n)$ CPU cycles before it halts.*

**Proof.** Let $P(S)$ be the program that simulates $S$ for $f(n)$ CPU cycles, where $n$ is the length of $S$, and returns "yes" if $S$ halted by that time, or "no" otherwise. Then the program $G$ generated by Theorem 1.10.19 is such that $P$ does not correctly determine if $G$ halts. Since $P$ is always correct when it returns "yes", this means that $G$ does halt, but that $P(G)$ returned "no", which implies that $G$ takes more than $f(n)$ cycles to execute. □

Of course, once one has a program of length $n$ that runs for more than $f(n)$ CPU cycles, it is not hard to make a program of length a little bit larger than $n$ that outputs a number greater than $f(n)$, so that one can conclude as a corollary that the Busy Beaver function outgrows any computable function.

**1.10.4. Miscellaneous.** The *strategy stealing argument* in game theory is already more or less set up in non-counterfactual form: in any game that admits "harmless moves" (such as noughts and crosses), any strategy of the second player can be stolen to be defeated (or at least held to a draw) by the first player. Similarly for arbitrage strategies in finance (unless there are loopholes due to imperfect information or friction costs).

It is a bit more difficult to recast the no-self-defeating objects in physics in a non-counterfactual form, due to the large number of implicit physical assumptions in these arguments. I will present just one simple example of this, which is the grandfather paradox that asserts that controlled time travel is impossible because you could use such travel to go back in time to kill your grandfather before you were born. One can convert this to a slightly less counterfactual format:

**"Theorem" 1.10.21.** *(*Very *imprecisely stated!) Suppose that one has a mechanism in universe $U$ to travel back in time and arrive at universe $U'$. Then there can exist events in $U$ that occurred differently in universe $U'$.*

The "proof" is, of course, the same: starting from $U$, go back in time and kill your grandfather in universe $U'$. This version of the "theorem" (though not the precise "proof" given here) is of course invoked often in many science fiction stories involving time travel.

It seems possible to also cast the no-immovable-objects and no-controlled-and-detectable-tachyon-particles arguments from [**Ta2010b**, §1.15] in this form, but one would have to consider multiple universes to do this properly, and I will not attempt to do so here, as it appears to be rather complicated.

The *omnipotence paradox* in philosophy (can an omnipotent being create a stone so heavy that He cannot lift it?) can also be rephrased in a non-counterfactual form that does not require consideration of any omnipotent beings:

**"Theorem" 1.10.22.** *If $G$ is a being, then $G$ will be unable to do at least one of the following two tasks:*

(1) *Create a stone so heavy that $G$ cannot lift it.*

(2) *Be able to lift any possible stone.*

Of course, most beings will fail at both Task 1 and Task 2.

## 1.11. The "no self-defeating object" argument, and the vagueness paradox

We continue our discussion of the "no self-defeating object" argument in mathematics - a powerful and useful argument based on formalising the observation that any object or structure that is so powerful that it can "defeat" even itself, cannot actually exist. This argument is used to establish many basic impossibility results in mathematics, such as Gödel's theorem that it is impossible for any sufficiently sophisticated formal axiom system to prove its own consistency, Turing's theorem that it is impossible for any sufficiently sophisticated programming language to solve its own halting problem, or Cantor's theorem that it is impossible for any set to enumerate its own power set (and as a corollary, the natural numbers cannot enumerate the real numbers).

As remarked in the previous section, many people who encounter these theorems can feel uneasy about their conclusions, and their method of proof; this seems to be particularly the case with regard to Cantor's result that the reals are uncountable. In the previous post in this series, I focused on one particular aspect of the standard proofs which one might be uncomfortable with, namely their counterfactual nature, and observed that many of these proofs can be largely (though not completely) converted to non-counterfactual form. However, this does not fully dispel the sense that the

conclusions of these theorems - that the reals are not countable, that the class of all sets is not itself a set, that truth cannot be captured by a predicate, that consistency is not provable, etc. - are highly unintuitive, and even objectionable to "common sense" in some cases.

How can intuition lead one to doubt the conclusions of these mathematical results? I believe that one reason is because these results are sensitive to the amount of vagueness in one's mental model of mathematics. In the formal mathematical world, where every statement is either absolutely true or absolutely false with no middle ground, and all concepts require a precise definition (or at least a precise axiomatisation) before they can be used, then one can rigorously state and prove Cantor's theorem, Gdel's theorem, and all the other results mentioned in the previous posts without difficulty. However, in the vague and fuzzy world of mathematical intuition, in which one's impression of the truth or falsity of a statement may be influenced by recent mental reference points, definitions are malleable and blurry with no sharp dividing lines between what is and what is not covered by such definitions, and key mathematical objects may be incompletely specified and thus "moving targets" subject to interpretation, then one can argue with some degree of justification that the conclusions of the above results are incorrect; in the vague world, it seems quite plausible that one can always enumerate all the real numbers 'that one needs to", one can always justify the consistency of one's reasoning system, one can reason using truth as if it were a predicate, and so forth. The impossibility results only kick in once one tries to clear away the fog of vagueness and nail down all the definitions and mathematical statements precisely[12].

One can already see this with one of the most basic conclusions of the "no self-defeating object" argument, namely that the set of natural numbers is infinite. Let me rephrase this result in the following manner:

**Proposition 1.11.1.** *Let $A$ be a set of natural numbers with the following properties:*

(1) $0$ *lies in* $A$.

(2) *Whenever a natural number $n$ lies in $A$, then its successor $n+1$ also lies in $A$.*

*Then $A$ is infinite*[13].

[12]To put it another way, the no-self-defeating object argument relies very much on the disconnected, definite, and absolute nature of the boolean truth space $\{\text{true}, \text{false}\}$ in the rigorous mathematical world. If one works in a "fuzzier" model of truth, such as Bayesian probability (see Section 6.9), then it becomes possible for vaguely defined objects to exist, even when they would become self-defeating in a classical truth model.

[13]Here, infinite has its usual set theoretic meaning, i.e. the conclusion is that $A$ cannot be placed in bijection with a set of the form $\{1, \ldots, n\}$ for any natural number $n$.

Indeed, from the principle of mathematical induction, the hypotheses of Proposition 1 force $A$ to be the entire set of natural numbers, and so Proposition 1.11.1 is logically equivalent to the assertion that the set of natural numbers is infinite.

In the rigorous world of formal mathematics, Proposition 1.11.1 is of course uncontroversial, and is easily proven by a simple "no self-defeating object" argument:

**Proof.** Suppose for contradiction that $A$ was finite. As $A$ is non-empty (it contains 0), it must therefore have a largest element $n$ (as can be seen by a routine induction on the cardinality of $A$). But then by hypothesis, $n+1$ would also have to lie in $A$, and $n$ would thus need to be at least as large as $n+1$, a contradiction. □

But if one allows for vagueness in how one specifies the set $A$, then Proposition 1.11.1 can seem to be false, as observed by the ancient Greeks with the *sorites paradox*. To use their original example, consider the question of how many grains of sand are required to make a heap of sand. One or zero grains of sand are certainly not sufficient, but clearly if one places enough grains of sand together, one would make a heap. If one then "defines" $A$ to be the set of all natural numbers $n$ such that $n$ grains of sand are not sufficient to make a heap, then it is intuitively plausible that $A$ obeys both Hypothesis 1 and Hypothesis 2 of the above proposition, since it is intuitively clear that adding a single grain of sand to a non-heap cannot convert it to a heap. On the other hand, it is just as clear that the set $A$ is finite, thus providing a counterexample to Proposition 1.11.1.

The problem here is the vagueness inherent in the notion of a "heap". Given a pile of sand $P$, the question "Is $P$ a heap?" does not have an absolute truth value; what may seem like a heap to one observer may not be so to another. Furthermore, what may seem like a heap to one observer when presented in one context, may not look like a heap to the same observer when presented in another context; for instance, a large pile of sand that was slowly accumulated over time may not seem[14] as large as a pile of sand that suddenly appeared in front of the observer, even if both piles of sand were of identical size in absolute terms.

There are many modern variants of the sorites paradox that exploit vagueness of definition to obtain conclusions that apparently contradict rigorous statements such as Proposition 1.11.1. One of them is the *interesting*

[14]The well-known (though technically inaccurate) boiling frog metaphor is a particularly graphic way of depicting this phenomenon, which ultimately arises from the fact that people usually do not judge quantities in absolute terms, but instead by using relative measurements that compare that quantity to nearby reference quantities.

*number paradox*. Let $A$ be the set of all natural numbers that are "interesting", in the sense that they have some unusual defining feature (for instance, 561 is the first composite Carmichael number[15] and is thus presumably an interesting number). Then $A$ is presumably finite, but the first natural number that does not lie in $A$ is presumably also an interesting number, thus contradicting the definition of $A$. Here, the variant of Proposition 1.11.1 that is apparently being contradicted here is the well-ordering principle.

Of course, just as the sorites paradox can be diagnosed as arising from the vagueness of the term "heap", the interesting number paradox can be diagnosed as arising from the vagueness of the term "interesting". But one can create a functionally similar paradox that, at first glance, eliminates this vagueness, namely the *Berry paradox*. Let $A$ denote the set of all natural numbers that can be defined in fewer than sixteen words in the English language; thus, again 561 can be defined as "The first composite Carmichael number" and thus belongs in $A$. As there are only finitely many sentences that one can form with fewer than sixteen words of the English language, $A$ is clearly finite. But the first natural number that does not lie in $A$ can be described as "The first natural number not definable in fewer than sixteen words in the English language", which is a definition in fewer than sixteen words in the English language, and thus lies in $A$, again providing another seeming contradiction to the well-ordering principle.

Here, the vagueness is a bit harder to spot, and it comes from the use of "the English language", which is a vague and mutable concept that allows for self-reference and is in fact technically inconsistent if one tries to interpret it naively (as can be seen from any number of linguistic paradoxes, starting with the classic *liar paradox*). To make things more precise, one can try to replace the English language here by a formal mathematical language, e.g. the language of *true arithmetic* $\mathbf{N}$. Let us take one such formal language and call it $L$. Then one can certainly define the set $A_L$ of all natural numbers that can be definable in fewer than sixteen words in the language $L$, and one can form the natural number $n_L$, defined as "The first natural number not definable in fewer than sixteen words in the language $L$". This is a natural number that is definable in fewer than sixteen words - but not in the language $L$, but rather in a meta-language $L'$ that, among other things, is able to interpret what it means for a sentence in $L$ to define a natural number. This requires a truth predicate for $L$, and *Tarski's indefinability theorem* (cf. Theorem 1.10.12) asserts that such a predicate cannot exist inside $L$ itself. Indeed, one can interpret Berry's paradox as providing a proof of Tarski's theorem. Thus we see that a non-trivial mathematical theorem (in this case,

[15] A *Carmichael number* is a natural number $n$ such that $a^{n-1} = 1 \bmod n$ for all $a$ coprime to $n$; all prime numbers are Carmichael numbers by Fermat's little theorem, but not conversely.

Tarski's theorem) emerges when one finally clears away all the vagueness from a sorites-type paradox. Similarly, if instead one replaced the word "definable" by "provably definable", and used a formal axiom system $L$ as one's language, one could similarly deduce Gödel's incompleteness theorem (cf. Corollary 1.10.15) from this argument.

Similar paradoxes come up when trying to enumerate the real numbers, or subsets of real numbers. Here, the analogue of Proposition 1.11.1 is

**Proposition 1.11.2.** *Let $A$ be a set of real numbers with the following properties:*

(1) *$A$ is infinite.*

(2) *Given any sequence $x_1, x_2, x_3, \ldots$ of elements of $A$, one can find another element $y$ of $A$ that is not equal to any of the elements of the sequence (i.e. $y \neq x_n$ for every positive integer $n$).*

*Then $A$ is uncountable*[16]*.*

Again, in the rigorous world in which all terms are clearly defined, this proposition is easily proven:

**Proof.** Suppose for sake of contradiction that $A$ was countable. Then, being infinite, it could be enumerated by a sequence $x_1, x_2, x_3, \ldots$ of real numbers in $A$; but then by Hypothesis 2, there would then be another real number $y$ that was also in $A$, but distinct from all of the elements of the sequence, contradicting the fact that this was an enumeration. □

Since Cantor's diagonal argument demonstrates that the set of reals $\mathbf{R}$ itself obeys Hypothesis 2 (and also clearly obeys Hypothesis 1), we conclude as a corollary that the reals are uncountable.

However, one can find apparent counterexamples to Proposition 1.11.2 if one deploys enough vagueness. For instance, in the spirit of the Berry paradox, one could try setting $A$ to be the set of all "definable" real numbers - those numbers that can be defined using a finite sentence in the English language. As there are only countably many such sentences, $A$ would be countable; it is also clearly infinite. But, on the other hand, one could take any sequence of real numbers $x_1, x_2, x_3, \ldots$ in $A$ and apply the diagonal argument to define another real number that does not lie in that sequence.

Again, this apparent contradiction (known as *Richard's paradox*) becomes clarified if one removes the vagueness, by replacing "the English language" with a formal language $L$. For instance, one could take $L$ to be true arithmetic, and $A_L$ to be the set of all real numbers that are definable by

[16]Here uncountability has the usual set-theoretic definition, namely that the infinite set $A$ is not in one-to-one correspondence with the natural numbers.

a sentence in $L$. Then $A_L$ is indeed countable, but any enumeration of $A_L$ requires the truth predicate for $L$ and thus such an enumeration (as well as the Cantor diagonalisation thereof) cannot be defined in $L$, but only in a meta-language $L'$ - thus obtaining yet another proof of Tarski's indefinability theorem. Or, one could take $L$ to be a programming language, such as the language of Turing machines; then again the set $A_L$ of real numbers whose decimal expansion can be given as the output of an algorithm in $L$ is countable, but to enumerate it one would have to solve the halting problem for $L$, which requires a meta-language $L'$ that is distinct from $L$ itself; thus in this case the diagonalisation argument gives a proof of Turing's halting theorem.

It is also instructive to contrast the formal distinction between countable and uncountable sets with the sorites paradox distinction between a non-heap and a heap of sand. Intuitively, the act of adding one grain of sand to a non-heap cannot directly turn that non-heap into a heap, and yet Proposition 1.11.1 forces this to eventually be the case. Similarly, it is intuitively clear that act of producing a single real number not on one's countable enumeration of a set $A$ does not directly convert a countable set to become uncountable, and yet Proposition 1.11.2 forces this to eventually be the case. More formally, Proposition 1.11.1 is powered by the principle of mathematical induction, which allows one to iterate the $n \to n+1$ operation indefinitely (or more precisely, up to the first infinite ordinal $\omega$) to explicitly achieve infinite cardinality; in a similar spirit, one can interpret Proposition 1.11.2 as being powered by transfinite induction, in the sense that one can iterate the diagonalisation operation indefinitely (or more precisely, up to the first uncountable ordinal $\omega_1$) to explicitly achieve uncountable cardinality. The transition from countability to uncountability does not occur during the successor ordinal steps of this transfinite induction, but during the (final) limit ordinal step.

One can perform a similar analysis for the other results discussed in Section 1.10 (or [**Ta2010b**, §1.15]). For instance, *Russell's paradox* tells us, among other things, that (assuming the standard Zermelo-Frankel axioms of set theory) the class of all sets cannot itself be a set. Actually we have the following slightly more general statement, analogous to Proposition 1.11.1 or Proposition 1.11.2:

**Proposition 1.11.3.** *Let $\mathcal{C}$ be a class of sets with the following properties:*

(1) *The empty set $\emptyset$ belongs to $\mathcal{C}$.*

(2) *If a set $A$ belongs to $\mathcal{C}$, then the singleton set $\{A\}$ also belongs to $\mathcal{C}$.*

(3) *If any collection* $\{A_\beta : \beta \in B\}$ *of sets* $A_\beta$*, indexed by another set* $B$*, is such that each* $A_\beta$ *lies in* $\mathcal{C}$*, then their union* $\bigcup_{\beta \in B} A_\beta$ *also lies in* $\mathcal{C}$*.*

*Then* $\mathcal{C}$ *is not a set.*

**Remark 1.11.4.** Hypothesis 1 is actually redundant, being implied by the trivial case of Hypothesis 3 when $B$ is empty, but we keep it in order to emphasise the similarity with Propositions 1.11.1 and 1.11.2.) The Zermelo-Frankel axioms (as well as one's intuition from naive set theory) tell us that the class of all sets obeys Hypotheses 1, 2, 3, and so cannot[17] be a set. A key subtlety is that the set $\{A_\beta : \beta \in B\}$ itself is not required to lie in $\mathcal{C}$. If one imposes such a restriction, then $\mathcal{C}$ can become a set (and by adding a few additional axioms, such sets become the same concept as *Grothendieck universes*).

**Proof.** Suppose for contradiction that $\mathcal{C}$ is a set. Using the axiom (schema) of specification, the set $B := \{A \in \mathcal{C} : A \notin A\}$ is then a set. It is the union of all the singleton sets $\{A\}$ with $A \in B$, so by Hypotheses 2 and 3, $B \in \mathcal{C}$. But then we see that $B \in B$ if and only if $B \notin B$, which is absurd. □

Again, the hypotheses in this proposition seem innocuous, much like adding a single grain of sand to a non-heap of sand; but if one iterates them indefinitely then one eventually ends up with a class so large that it is no longer a set. Here, the transfinite induction is not up to the first infinite ordinal or the first uncountable ordinal, but rather across the class of *all* ordinals; the point then being that the class of all ordinals is itself not a set, a fact also known as the *Burali-Forti paradox*. Naive set theory intuition seems to be in contradiction to Proposition 1.11.3, but this is only due to the vagueness inherent in that concept of set theory. One can also try analysing Berry-type paradoxes in this setting, for instance working only with "constructible" sets (i.e. elements of Gödel's universe $L$); the consequence one gets from this is that the class of constructible sets is not itself constructible (in fact, it is not even a set, as it contains all the ordinals).

Proposition 1.11.3 may seem only of interest to set theorists and logicians, but if one makes a tiny modification of it, by replacing a class of sets with a *partially ordered* class, then one gets a very useful result:

**Lemma 1.11.5** (Zorn's lemma)**.** *Let* $\mathcal{C}$ *be a partially ordered class which obeys the following properties:*

(1) *At least one element belongs to* $\mathcal{C}$*.*

[17] In fact, the above proposition is essentially equivalent to the slightly stronger assertion that the class of all *well-founded* sets, also known as the *von Neumann universe* $V$, is not a set.

(2) *If $x$ is an element of $\mathcal{C}$, then there is another element $y$ of $\mathcal{C}$ such that $y > x$.*

(3) *If $(x_\beta)_{\beta \in B}$ is a totally ordered set in $\mathcal{C}$, indexed by another set $B$, then there exists an element $y \in \mathcal{C}$ such that $y \geq x_\beta$ for all $\beta \in B$.*

*Then $\mathcal{C}$ is not a set.*

This lemma (which, of course, requires the axiom of choice to prove, and is in fact equivalent to this axiom) is usually phrased in the contrapositive form: any non-empty set $\mathcal{C}$ for which every totally ordered set has an upper bound, has a maximal element. However when phrased in the above form, we see the close similarity between Zorn's lemma and Propositions 1.11.1-1.11.3. In this form, it can be used to demonstrate that many standard classes (e.g. the class of vector spaces, the class of groups, the class of ordinals, etc.) are not sets, despite the fact that each of the hypotheses in the lemma do not directly seem to take one from being a set to not being a set. This is only an apparent contradiction if one's notion of sets is vague enough to accommodate sorites-type paradoxes.

More generally, many of the objects demonstrated to be impossible in the previous posts in this series can appear possible as long as there is enough vagueness. For instance, one can certainly imagine an omnipotent being provided that there is enough vagueness in the concept of what "omnipotence" means; but if one tries to nail this concept down precisely, one gets hit by the omnipotence paradox. Similarly, one can imagine a foolproof strategy for beating the stock market (or some other zero sum game), as long as the strategy is vague enough that one cannot analyse what happens when that strategy ends up being used against itself. Or, one can imagine the possibility of time travel as long as it is left vague what would happen if one tried to trigger the grandfather paradox. And so forth. The "self-defeating" aspect of these impossibility results relies heavily on precision and definiteness, which is why they can seem so strange from the perspective of vague intuition.

## 1.12. A computational perspective on set theory

The standard modern foundation of mathematics is constructed using *set theory*. With these foundations, the mathematical universe of objects one studies contains not only[18] the "primitive" mathematical objects such as numbers and points, but also sets of these objects, sets of sets of objects, and so forth. One has to carefully impose a suitable collection of axioms

[18] In a *pure* set theory, the primitive objects would themselves be sets as well; this is useful for studying the foundations of mathematics, but for most mathematical purposes it is more convenient, and less conceptually confusing, to refrain from modeling primitive objects as sets.

on these sets, in order to avoid paradoxes such as *Russell's paradox*; but with a standard axiom system such as *Zermelo-Fraenkel-Choice* (ZFC), all actual paradoxes that we know of are eliminated. Still, one might be somewhat unnerved by the presence in set theory of statements which, while not genuinely paradoxical in a strict sense, are still highly unintuitive; *Cantor's theorem* on the uncountability of the reals, and the *Banach-Tarski paradox*, are perhaps the two most familiar examples of this. (Cantor's theorem is discussed below in Sections 1.10, 1.11; the Banach-Tarski paradox is discussed in [**Ta2010**, §2.2].)

One may suspect that the reason for this unintuitive behaviour is the presence of infinite sets in one's mathematical universe. After all, if one deals solely with finite sets, then there is no need to distinguish between countable and uncountable infinities, and Banach-Tarski type paradoxes cannot occur.

On the other hand, many statements in infinitary mathematics can be reformulated into equivalent statements in finitary mathematics (involving only finitely many points or numbers, etc.); see for instance [**Ta2008**, §1.3, §1.5], [**Ta2010b**, §2.11]. So, one may ask: what is the finitary analogue of statements such as Cantor's theorem or the Banach-Tarski paradox?

The finitary analogue of Cantor's theorem is well-known: it is the assertion that $2^n > n$ for every natural number $n$, or equivalently that the power set of a finite set $A$ of $n$ elements cannot be enumerated by $A$ itself. Though this is not quite the end of the story; after all, one also has $n + 1 > n$ for every natural number $n$, or equivalently that the union $A \cup \{a\}$ of a finite set $A$ and an additional element $a$ cannot be enumerated by $A$ itself, but the former statement extends to the infinite case, while the latter one does not. What causes these two outcomes to be distinct?

On the other hand, it is less obvious what the finitary version of the Banach-Tarski paradox is. Note that this paradox is available only in three and higher dimensions, but not in one or two dimensions; so presumably a finitary analogue of this paradox should also make the same distinction between low and high dimensions.

I therefore set myself the exercise of trying to phrase Cantor's theorem and the Banach-Tarski paradox in a more "finitary" language. It seems that the easiest way to accomplish this is to avoid the use of set theory, and replace sets by some other concept. Taking inspiration from theoretical computer science, I decided to replace concepts such as functions and sets by the concepts of *algorithms* and *oracles* instead, with various constructions in set theory being replaced instead by computer language *pseudocode*. The point of doing this is that one can now add a new parameter to the universe, namely the amount of computational resources one is willing to allow

one's algorithms to use. At one extreme, one can enforce a "strict *finitist*" viewpoint where the total computational resources available (time and memory) are bounded by some numerical constant, such as $10^{100}$; roughly speaking, this causes any mathematical construction to break down once its complexity exceeds this number. Or one can take the slightly more permissive "finitist" or "*constructivist*" viewpoint, where any finite amount of computational resource is permitted; or one can then move up to allowing any construction indexed by a countable *ordinal*, or the storage of any array of countable size. Finally one can allow constructions indexed by arbitrary ordinals (i.e. *transfinite induction*) and arrays of arbitrary infinite size, at which point the theory becomes more or less indistinguishable from standard set theory.

I describe this viewpoint, and how statements such as Cantor's theorem and Banach-Tarski are interpreted with this viewpoint, in the rest of this section. I should caution that this is a conceptual exercise rather than a rigorous one; I have not attempted to formalise these notions to the same extent that set theory is formalised. Thus, for instance, I have no explicit system of axioms that algorithms and oracles are supposed to obey. Of course, these formal issues have been explored in great depth by logicians over the past century or so, but I do not wish to focus on these topics in this post.

A second caveat is that the actual semantic content of this post is going to be extremely low. I am not going to provide any genuinely new proof of Cantor's theorem, or give a new construction of Banach-Tarski type; instead, I will be reformulating the standard proofs and constructions in a different language. Nevertheless I believe this viewpoint is somewhat clarifying as to the nature of these paradoxes, and as to how they are not as fundamentally tied to the nature of sets or the nature of infinity as one might first expect.

**1.12.1. A computational perspective on mathematics.** The great advantage of using set theory in mathematics is that all objects in a given set (e.g. all numbers in the real line $\mathbf{R}$) are available to you at all times; one can take one, many, or all objects in a set and manipulate them as one wishes (cf. the *axiom schema of replacement*); similarly, one can assign a truth value to a statement that quantifies over an arbitrary number of objects. If one removes sets from the picture, then one no longer has immediate access to arbitrary elements of a set, and one can no longer perform operations *en masse* on all the elements of a set at once; instead, one must use some

(possibly more restrictive) protocol[19] for manipulating objects in a class, or verifying whether a given statement is true or false.

For this, it is convenient to use the conceptual framework that is familiar to us through modern computer languages, such as $C$. In this paradigm, when dealing with a class of objects (e.g. integers), we do not get access to the entire set $\mathbf{Z}$ of integers directly. Instead, we have to declare a integer variable, such as $n$, and set it equal to some value, e.g. $n := 57$; or, if one is creating a routine that takes input, $n$ might be initialised to one of the unspecified inputs of that routine. Later on, we can use existing variables to define new ones, or to redefine existing ones, e.g. one might define $m$ to equal $n * n$, or perhaps one can increment $n$ to $n + 1$. One can then set up various loops and iterations to explore more of the parameter space; for instance, if countably infinite loops are permitted as a computational resource, then one can exhaust the positive integers by starting at $n = 1$ and incrementing $n$ by 1 indefinitely; one can similarly exhaust the negative integers, and by alternating between the two (and also passing through 0) one can exhaust the entire integers by a countably infinite loop. This of course is just the standard demonstration that the integers are countable.

Real-world computers, of course, have finite limits of precision; they cannot represent arbitrarily large integers, but only integers up to a certain size (e.g. $2^{16} - 1$ or $2^{32} - 1$). One could think about computational models with such a strict finitary limitation, but let us assume that we are in a more idealised setting in which there are no limitations on how large an integer one can store. Let us then make the even more idealised assumption that we can also store real numbers with unlimited precision; our computer never[20] makes any roundoff errors.

**Remark 1.12.1.** A technical point: it may be that the computational model of the real numbers is different from the standard real line $\mathbf{R}$; for instance, perhaps the computer only implements "constructible" real numbers, which is for instance the case in physical arbitrary precision computers. We will touch upon this point again later.

Note that if one were to expand out a given real number, say $\pi$, as a decimal expansion, then one would obtain an infinite string of digits. But, just as we do not have direct access to the set $\mathbf{Z}$ of all integers, we will not have direct access to the entire decimal expansion of $\pi$. If we have a natural

[19] In more philosophical terms, we are focusing more on an epistemological approach to mathematics, based on what we can measure and query, as opposed to an ontological approach, based on what we believe to exist.

[20] Note that there are indeed *arbitrary precision models of computation* that can do this, though the catch is that speed of computation depends heavily on how complicated it is to describe any given number.

number $n$, we are allowed to inspect the $n^{\text{th}}$ digit of $\pi$ by making a suitable function call (e.g. $\text{digit}(\pi, n)$), and if we are allowed to set up an infinite loop in which $n$ starts at 1 and increments indefinitely, one can exhaust all the digits of $\pi$, which is good enough for most "practical" mathematical purposes. For instance, if one were allowed to run programs of countable length using real arithmetic, one could make a program that determined whether the digits of $\pi$ were uniformly distributed or not, or to determine the truth of the Riemann hypothesis, or more generally to compute the truth-value of any first-order sentence in the theory of the real line.

We assume that the usual arithmetic operations can be performed on real numbers in reasonable amounts of time. For instance, given two real numbers $x, y$, one can determine whether they are equal or not in finite time (consulting some sort of "equality oracle" for the reals if necessary). Note that equality can be a subtle issue; if one thinks of real numbers as infinite strings of digits, their equality can only be verified directly by using a countable amount of computing power. But we will sidestep the issue of how exactly the reals are implemented by simply assuming that enough oracles exist to perform real arithmetic at an acceptable computational cost.

As already hinted at in the above discussion, we are assuming that our computer has access to a certain amount of computing resources (e.g. time, memory, random number generation, oracles). We will be rather vague on exactly how to formalise the concept of a resource, but basically the standard definitions used in computer science would be a good approximation here, at least when the resources are finite. But one can consider allowing for certain computational resources to be infinite in some carefully controlled manner; for instance, one could consider a situation in which countably infinite loops are permitted (provided that all variables in the loop that one wants to retain "converge" in some sense at the end of the loop), but for which uncountable loops are not allowed. We will not formalise such concepts here (but roughly speaking, they correspond to allowing transfinite induction up to some specified ordinal, and no further).

We will not be using sets or set-theoretic functions in this computer language. However, we will use as a substitute the concept of an *oracle* - a "black box" routine $f()$ that takes zero or more variables of a given class as input, and returns zero or more variables in various classes as output (usually we will have just a single output, but it will be convenient to allow multiple outputs). Being a black box, we do not know how the oracle obtains the output from the input, but we are able to use the oracle anyway. Let us assume that each invocation of an oracle takes some acceptable amount of time (e.g. bounded time when the computational resources are finite, or countably infinite time if countable time resources are allowed, etc.). All

our oracles are *consistent* in the sense that they always produce the same output for a fixed choice of input; thus, if one calls the oracle again at some later time with the same input, then the oracle will return the same output as it did before. It is important to note that consistency is a subtly weaker assumption than requiring the oracle is *non-adaptive*; we allow the oracle to "remember" previous queries, and to use that memory to formulate answers to later queries, as long as it does not contradict the outputs it gave previously.

We will be concerned primarily with *membership oracles* - an oracle $E()$ that takes a variable $x$ in a given class and returns an answer $E(x)$ that is either "Yes" or "No". Informally, the oracle $E$ is describing some subset of this class, and is answering questions regarding whether any given variable $x$ lies in this set or not. Note, however, that this set only exists in a "virtual" or "potential" sense; if the oracle is adaptive, the set of inputs that will give a "Yes" answer may not yet be fixed, but could depend on future queries to the oracle. If the class can be exhausted within the computational resources permitted (e.g. if the parameter space can be countably enumerated by the computer, and countably infinite loops are permitted), then one can query the oracle for every single element and thus pin down the set completely (though one may not be able to *store* this set if one does not have sufficient memory resources!), but if the parameter space is too large to be exhausted with the available resources, then the set that the oracle is describing will never be completely described.

To illustrate this, let us briefly return to the traditional language of set theory and recall the following textbook example of a non-measurable subset $E$ of the reals $\mathbf{R}$, which can be found for instance in [**Ta2011**, Proposition 1.2.18]. This set is constructed by partitioning $\mathbf{R}$ into cosets $x + \mathbf{Q}$ of the rationals $\mathbf{Q}$, and then using the axiom of choice to selecting a single representative of each coset; the set $E$ is the collection of all such representatives. Thus the rational translates $E + q$, $q \in \mathbf{Q}$ partition $\mathbf{R}$, and it is not hard to then deduce (using the properties of Lebesgue measure) that $E$ cannot be Lebesgue measurable.

We can simulate this non-measurable set by an *adaptive* oracle $E()$, which remembers all prior queries to itself, and works as follows:

- $E()$ takes a real number $x$ as input.
- If $E()$ has previously answered the question $E(x)$ of whether $x$ lies in $E$, repeat whatever answer was given previously.
- If $x$ has not been queried before, and if furthermore no rational translate $x + q$ of $x$ has been queried before either (i.e. $x$ differs

from all previous queries by an irrational number), then answer $E(x)$ with "yes".

- Finally, if $x$ has not been queried before, but $x+q$ has been queried before for some rational $q$, then answer $E(x)$ with "no".
- Store $x$ (and $E(x)$) in memory for use in future queries.

Assuming one has a rationality oracle $\mathbf{Q}()$ that can tell (in bounded time) whether a given real number $x$ is rational or not, then $E$ is a perfectly programmable oracle, which will run in finite time whenever one asks only a finite number[21] of queries of it. It is even completely deterministic - it requries no arbitrary choices on the part of the oracle. If one had the patience to query this oracle for every single real number $x$ (which would, of course, require an uncountable number of queries), the oracle would eventually describe completely the non-measurable set $E$. But if one is only permitted a finite or countable number of queries, then the non-measurable set only exists in a "virtual" sense.

More generally, non-adaptive oracles tend (as a rule of thumb) to generate measurable sets, while adaptive oracles are likely to generate non-measurable sets. So we see that non-measurability does not have to be viewed as a quirk arising from the nature of infinity, or from the axiom of choice; it can be viewed instead as the freedom to adapt to previous queries to membership of the set, which is a concept that makes sense even in a strictly finitist setting.

**Remark 1.12.2.** One can think of an non-adaptive oracle as being like a truthful observer, reporting on some objective set that exists independently of the queries, while an adaptive oracle is more like a pathological liar, inventing a previously non-existent set on the fly as needed in order to consistently answer the questions posed of it. It is then not so surprising that the set thus invented is likely to be non-measurable. We thus see that the ability of oracles to adapt is somewhat analogous to the role[22] the axiom of choice plays in traditional set theory.

We will discuss measurability and non-measurability in more detail a little later in this post.

We have seen that membership oracles are sort of like "virtual" sets. Many set operations can be simulated for membership oracles. For instance, given two membership oracles $E(), F()$ that apply to variables $x$ in some

[21] After querying the oracle an infinite number of times, though, it will require an infinite search loop in order to make sure that any subsequent answer it gives is consistent with all previous answers, unless one is allowed to use an infinite amount of memory, e.g. an array indexed by the quotient space $\mathbf{R}/\mathbf{Q}$.

[22] As pointed out to me by Chad Groft, adaptive oracles can also be used to interpret the method of *forcing* in model theory.

class (e.g. the real numbers), one can form the union oracle $(E \cup F)()$, which works by querying both $E(x)$ and $F(x)$ and returning "Yes" if at least one of $E(x)$ and $F(x)$ was "Yes", and "No" otherwise. More generally, any finite boolean operation of membership oracles gives another membership oracle, and (if countable computational resources are available) the same is true for countable boolean operations also. As one increases the amount of computational resources available, more and more set-theoretic operations become available, and when one allows unlimited resources (or more precisely, transfinite induction up to any ordinal, and storage of arbitrarily sized infinite sets), then all the standard operations in set theory (e.g. invocations of the *axiom schema of replacement*, the *power set axiom*, the *axiom of choice*, etc.) become available.

**Remark 1.12.3.** Membership oracles $E()$ only describe a raw, unstructured set. If one wants to place additional structure on a set (e.g. measure structure, topological structure, smooth structure, etc.) then additional oracles would be needed. Of course, the same is true in traditional set theory; for instance, to place a topology on a set $E$ one also needs to specify a collection $\mathcal{F}$ of open sets on $E$ obeying the axioms of a topology, and similarly for other structures one can place on sets. We will see an example of these additional oracles later in this section, when we revisit the concept of measurability.

**Remark 1.12.4.** Membership oracles are weaker than sets in many ways. One of these comes from a breakdown of the law of the excluded middle. In set theory, a statement about a set $E$ is either true or false; for instance, $E$ is either finite or infinite. If one is instead given an adaptive membership oracle $E()$, questions such as whether $E()$ describes a finite set or not are undecidable if one only has finite computational resources. However, one can imagine strengthening the membership oracle $E()$ to a oracle that, in addition to answering questions about membership of individual elements, will also answer more general questions about $E$ (such as whether $E$ is finite), in such a fashion that all the answers given are consistent with each other. In such a way, the law of the excluded middle can be restored; but then programming an adaptive oracle in a way that keeps all the answers consistent becomes quite a challenge[23].

**1.12.2. Cantor's theorem.** *Cantor's theorem* (and its proof) transfers easily enough from sets to oracles. The analogue of a (proposed) enumeration of the real numbers is an "enumeration oracle" $f()$ that takes a natural number $n$ as input, and returns a real number $f(n)$ as output; we allow repetitions. The Cantor diagonal argument shows that given any such putative enumeration $f$, and given access to a countable amount of computing

[23]Note though that the Gödel completeness theorem asserts, in some sense, that this is always possible, provided that one's initial constraints on $E$ are originally consistent.

resources, one can construct a real number $x$ which is guaranteed not to be covered by the enumeration oracle; for instance, one can construct $x$ to be a string of decimals $0.x_1x_2\ldots = \sum_{j=1}^{\infty} \frac{x_j}{10^j}$, with $x_1$ chosen to be the first digit in[24] $\{1,\ldots,8\}$ not equal to the first digit of $f(1)$, $x_2$ chosen to be the first digit of $\{1,\ldots,8\}$ not equal to the second digit of $f(2)$, and so forth.

This is of course virtually identical to the usual proof of Cantor's theorem. But the proof highlights that in order to exhibit a counterexample to the claim that $f$ enumerates the reals, one needs a countably infinite amount of computational resources. And indeed if one works in a finitary computational model in which one is only allowed to run programs that are guaranteed to halt, and if one limits one's real number system to the "finitely constructible" reals - those reals that can be generated by halting programs - then one can indeed enumerate the "real numbers" in this system, by enumerating all the possible programs to generate real numbers as $P_1, P_2, \ldots$, and computing $f(n)$ to be the output of $P_m$, where $1 \leq m \leq n$ is the first integer such that $P_m$ halts in less than $n$ steps, while outputting a number different from $f(1),\ldots,f(n-1)$ (setting $f(n) = 0$ if no such integer exists). In this model, the real number $x$ given by Cantor's argument is not finitely constructible, but only countably constructible. It is only because one has access to countably infinite computational resources (in particular, the ability to sum convergent infinite series such as $\sum_{j=1}^{\infty} \frac{x_j}{10^j}$), that the reals are no longer countable.

Let us now consider a slightly different version of Cantor's theorem, which in the language of set theory asserts that for a given set $A$ (e.g. the natural numbers $\mathbf{N}$), that one cannot enumerate the power set $2^A$ by $A$ itself. In order to phrase this using the language of oracles rather than sets, one needs to be able to treat oracles themselves as variables; note that many modern computer languages (particularly functional languages such as *Lisp* or *Haskell*) already do this. In particular, we allow for the existence of oracles that generate further oracles as output, or themselves take oracles as input.

A proposed enumeration of the power set of the natural numbers (say) by the natural numbers themselves can then be viewed as an enumeration oracle $f()$ which takes a natural number $n$ as input, and returns a membership oracle $f(n)()$ in the natural numbers as output; thus $f(n)$ itself is able to accept a natural number $m$ as input and return a yes-or-no answer $f(n)(m)$ as output.

Cantor's diagonal argument then asserts that given any proposed enumeration oracle $f()$ of the above form, one can always find a membership

[24]Here, we have excluded the digits 0 and 9 to avoid the technical and irrelevant $0.999\ldots = 1.000\ldots$ issue.

oracle $E()$ which is not enumerated by $f$. Indeed, one simply sets $E(n)$ to be the negation of $f(n)(n)$ for any given natural number input $n$.

With this version of Cantor's argument, we no longer need a countably infinite amount of computational resources; the above argument is valid even in the finitary computational model. In that model, the argument shows that the class of oracles that terminate in finite time cannot itself be enumerated by an oracle that terminates in finite time; this is essentially *Turing's halting theorem*. Thus we see that Turing's theorem and Cantor's theorem are simply different cases[25] of a single general theorem, the former in the case of finitary computational resources and the latter in the case of unlimited computational resources.

Also observe that this version of Cantor's argument works if the natural numbers are replaced by any other class of variable; for instance, in the finite class of integers between 1 and $n$, the argument demonstrates that the membership oracles in this class cannot be enumerated by the numbers from 1 to $n$ itself, thus $2^n > n$.

Let us now discuss the analogous situation with the inequality $n+1 > n$. The claim is now that if $A$ is a finite class, and $A'$ is a strictly larger class (containing at least one additional element $a$ outside of $A$), and one is given an enumeration oracle $f()$ that takes an variable $x$ in $A$ as input and returns a variable $f(x)$ in $A'$ as output, then one should be able to find a variable $y$ in $A'$ or equal to $a$ which is not covered by the enumeration oracle.

One way to find such a $y$ is by the following algorithm, which is just a painfully disguised version of the proof that $n + 1 > n$ using induction. Let $a$ be an element in $A'$ that is not in $A$. We first search through all the elements $x$ of $A$ to see if there is a solution to the equation $f(x) = a$. If no solution exists, then we are done; we can just take $y = a$. So suppose instead that we have some found some $x_0$ for which $f(x_0) = a$. Then we can delete $x_0$, and all other solutions to $f(x) = a$, from the domain $A$ of the oracle $f$, and also delete $a$ from the output $A'$ of the oracle, giving rise to a modified oracle $f'$ which takes as input a variable $x$ with $f(x) \neq a$, and returns the output $f'(x) := f(x)$. Note that if we ever find a variable $y$ in the range of $f'$ that is not enumerated by $f'$, then $y$ will not be enumerated by $f$ either. So we can descend from $f$ to the oracle $f'$, which has a smaller domain and range. Also note that the range remains strictly larger than the domain, as $x_0$ lies in the range but not in the domain. We can thus keep iterating this procedure; since we cannot have an *infinite descent*, the algorithm must eventually terminate. Unpacking the termination condition, we have indeed produced an element $y$ in the range of $f$ which was not enumerated by $f$.

[25] See also [**Ta2010b**, §1.15] and Sections 1.10, 1.11 for further discussion of these "no self-defeating object" arguments.

We observe that this algorithm only halts due to the principle of infinite descent, which is only valid when the initial class one starts with was finite. For infinite classes, which admit infinite descents, one can of course find surjective enumerations between such classes and strictly larger classes; for instance, the enumeration oracle that sends a natural number $n$ to $n-1$ is an enumeration of $\mathbf{N} \cup \{-1\}$ by $\mathbf{N}$. In contrast, the proof of Cantor's theorem does not rely on facts specific to finite classes, such as the principle of infinite descent, and is thus valid in arbitrary classes.

**1.12.3. Measurability revisited.** In Section 1.12.1, we gave an example of a membership oracle $E$ in the real line which was non-measurable, in the sense that the set that it was (virtually) describing necessarily failed to be Lebesgue measurable. But the concept of Lebesgue measurability was still defined in the context of set theory, and not in a purely oracle-theoretic language. It is then natural to ask whether one can define Lebesgue measurability purely in the context of computational models, and in particular whether one can discuss this concept in a finitary computational model.

For simplicity let us now work in a bounded subset of $\mathbf{R}$, such as the unit interval $[0,1]$, so that we can work with a finite measure rather than a $\sigma$-finite measure.

From classical measure theory, we recall the following characterisation of Lebesgue measurability: a subset $E$ of the unit interval is Lebesgue measurable if for every $\varepsilon > 0$, one can find an elementary subset $E_\varepsilon$ of the unit interval (i.e. a finite union of open intervals) whose set-theoretic difference $E \Delta E_\varepsilon$ with $E$ has *outer measure* less than $\varepsilon$, or equivalently that it can be covered by a countable collection of intervals $I_{\varepsilon,1}, I_{\varepsilon,2}, \ldots$ whose length adds up to less than $\varepsilon$.

Thus, if one wants a membership oracle $E()$ to "certify" its measurability, one way to do so is to provide an additional "measurability oracle" $M()$, which takes a positive real number $\varepsilon > 0$ as input, and returns three outputs:

- A description of an elementary set $E_\varepsilon$ (which can be done in a finite amount of time and space, by specifying the number of intervals used to form $E_\varepsilon$, together with their endpoints);
- An interval oracle $I()$, that for each natural number input $n$ returns an interval $I_{\varepsilon,n}$;
- A covering[26] oracle $N()$, which for each real number $x$ in $E \Delta E_\varepsilon$, returns a natural number $n$ for which $x \in I_{\varepsilon,n}$.

[26] Actually, this oracle is redundant, as it can be simulated from the previous two outputs by a finite brute force search; but we include it for conceptual clarity.

With these oracles, one can then verify the measurability of $E$ by selecting an $\varepsilon > 0$, and verifying firstly that after selecting any natural number $N$, the sum of the lengths of $I_{\varepsilon,1}, \ldots, 1_{\varepsilon,N}$ remains less than $\varepsilon$, and secondly that after selecting any real number $x$, that if $x$ lies in $E$ but not in $E_\varepsilon$ or vice versa, that $x$ indeed lies in $I_{\varepsilon,N(x)}$. In principle, if one performed this verification procedure an uncountable number of times (once for each choice of $\varepsilon, N, x$) one would fully demonstrate the measurability of $E$; but if one instead only had access to finite or countable computational resources, then one could only verify measurability on an "as needed" basis.

So, in the oracle model, a measurable set is not simply a membership oracle $E()$; it must also be supplemented with an additional measurability oracle $M()$ that "witnesses" the measurability. This is analogous to how sets must be augmented with (say) topological structure if one wants to perform topology on that set, or algebraic structure if one wants to perform algebra on that set, etc.

If one possesses a measurability oracle $M()$ for a set $E$ (or more precisely, a membership oracle $E()$), then can estimate the Lebesgue measure of $E$ to within accuracy $\varepsilon$ by calling $M(\varepsilon)$ to obtain an approximant $E_\varepsilon$, and then computing the measure $|E_\varepsilon|$ of $E_\varepsilon$ (which can be done in a finite amount of time, as $E_\varepsilon$ is simply a finite union of intervals). A key fact (which, not coincidentally, is crucial in the standard construction of Lebesgue measure) is that these approximations to the Lebesgue measure of $E$ are compatible with each other in the following sense: if one calls one measurability oracle $M(\varepsilon)$ for $E()$ at accuracy $\varepsilon > 0$ to get one estimate $|E_\varepsilon|$ for the Lebesgue measure of $E()$, and if one then calls another (possibly different) measurability oracle $M'(\varepsilon')$ for the same set $E()$ at another accuracy $\varepsilon' > 0$ to get another estimate $|E'_{\varepsilon'}|$ for $E()$, then these two estimates can only differ by at most $\varepsilon + \varepsilon'$; in particular, sending $\varepsilon \to 0$, one obtains a Cauchy sequence and (after a countable number of operations) one can then compute the Lebesgue measure of $E$ to infinite precision.

This key fact boils down (after some standard manipulations) to the fact that an interval such as $[a, b]$ has outer measure at least $b - a$; in our oracle based model, this means that if one is given an interval oracle $I()$ that generates open intervals $I(n)$ for each natural number $n$, in such a way that the total length $\sum_n |I(n)|$ of these intervals is less than $b - a$, then one can find a point in $[a, b]$ which is not covered by any of the $I(n)$.

This can be done using a countable amount of computational power (basically, the ability to run a single infinite loop; this is roughly equivalent to the theory $\mathrm{RCA}_0$ tht is used in in reverse mathematics). The point is that for each finite $N$, the set $S_N := [a, b] \backslash \bigcup_{n=1}^N I(n)$ is a computable non-empty finite union of closed intervals in $[a, b]$, which decreases in $N$. The infimum

$\inf(S_N)$ can be computed infinite time for each $N$, and increases in $N$; the limit $\lim_{N\to\infty} \inf(S_N)$ is then an element in $\bigcap_N S_N$ that lies in $[a,b]$ but is outside all of the $I(n)$. Thus, given a countable amount of computational power, one can consistently define Lebesgue measure of a measurable set, and verify its basic properties.

It is instructive to apply the above discussion to the non-measurable membership oracle $E()$ given in previous sections (trivially modified to lie in $[0,1]$ rather than in $\mathbf{R}$). If one is given a purported measurability oracle $M()$ for this oracle $E()$, one can eventually show that this oracle $M()$ does not actually certify the measurability of $E()$ as claimed, but this requires a countably infinite amount of computation to establish. (Basically, there are two cases, based on whether $M()$ asserts that $E()$ has positive Lebesgue measure or not (which can be decided after a countable amount of computation). If it has positive measure, then by invoking $M(\varepsilon)$ with $\varepsilon$ less than half this measure, one can soon find an interval $I$ in which $E()$ has density greater than $1/2$ (or more precisely, the complement of $E$ in $I$ has outer measure strictly less than $|I|/2$) and then one can run a variant of the above Bolzano-Weierstrass argument to find two points $x, y$ in $E()$ and in $I$ which differ by a rational, contradicting the construction of $E$. If instead $M()$ asserts that $E()$ has zero measure, then $M()$ can cover $E()$ by intervals of arbitrarily small total measure, and then $M()$ can do the same for the union of all the rational shifts of $E()$, and one can then find an element $x \in [0,1]$ such that no rational shift $x+q$ of $x$ lies in $E$.)

On the other hand, if one is only allowed to query $E()$ and $M()$ finitely many times, then one can show that one can adaptively build $M()$ and $E()$ in response in these queries in such a way that one never obtains a contradiction, while retaining the properties that the shifts of $E()$ by the rationals partition $\mathbf{R}$. So a pathological liar can build a non-measurable set but claim that it is measurable; the deception can sustain any finite number of queries about the set and its measurability, but given a countable amount of queries one can eventually demonstrate that there is an inconsistency (at least for non-measurable sets coming from the coset partition of $\mathbf{R}$ by the rationals).

**Remark 1.12.5.** Interestingly, the type of adaptive oracles one uses to create non-measurable sets are not compatible with random number generation. If one has access to a source of random real numbers (say in $[0,1]$), then in principle one can (almost surely) compute the Lebesgue measure in countable time of any subset $E$ of $[0,1]$ accessed through a random oracle $E()$ by the *Monte Carlo method*: randomly sampling $N$ points in $[0,1]$, counting the proportion that lie in $E$, and then sending $N \to \infty$. However this only

works properly if the oracle $E()$ does not adapt to the various random inputs it is given, and is instead independent of these inputs. For instance, if one were to use Monte Carlo methods to measure the non-measurable set $E$ described above, one would almost surely get that $E$ has measure 1, as each random number is almost certain to lie in a different coset of $\mathbf{Q}$!.

**1.12.4. The Banach-Tarski paradox.** We now turn to the *Banach-Tarski paradox*. The usual formulation of this paradox involves a partition of the unit ball into pieces that can be rearranged (after rotations and translation) to form two copies of the ball. To avoid some minor technicalities, we will work instead on the unit sphere $S^2$ with an explicit countable set $\Sigma$ removed, and establish the following version of the paradox:

**Proposition 1.12.6** (Banach-Tarski paradox, reduced version)**.** *There exists a countable subset $\Sigma$ of $S^2$ and partition of $S^2\backslash\Sigma$ into four disjoint pieces $E_1 \cup E_2 \cup E_3 \cup E_4$, such that $E_1$ and $E_2$ can be rotated to cover $S^2\backslash\Sigma$, and $E_3$ and $E_4$ can also be rotated to cover $S^2\backslash\Sigma$.*

Of course, from the rotation-invariant nature of Lebesgue measure on the sphere, such a partition can only occur if at least one of $E_1, E_2, E_3, E_4$ are not Lebesgue measurable.

We return briefly to set theory and give the standard proof of this proposition. The first step is to locate two rotations $a, b$ in the orthogonal group $SO(3)$ which generate the free group $\langle a, b\rangle$. This can be done explicitly; for instance one can take

$$a := \begin{pmatrix} 3/5 & 4/5 & 0 \\ -4/5 & 3/5 & 0 \\ 0 & 0 & 1 \end{pmatrix}; \quad b := \begin{pmatrix} 1 & 0 & 0 \\ 0 & 3/5 & -4/5 \\ 0 & 4/5 & 3/5 \end{pmatrix}.$$

(See [**Ta2010**, §2.2] for a verification (using the *ping-pong lemma*) that $a, b$ do indeed generate the free group. Each rotation in $\langle a, b\rangle$ has two fixed antipodal points in $S^2$; we let $\Sigma$ be the union of all these points. Then the group $\langle a, b\rangle$ acts freely on the remainder $S^2\backslash\Sigma$.

Using the *axiom of choice*, we can then build a (non-measurable) subset $E$ of $S^2\backslash\Sigma$ which consists of a single point from each orbit of $\langle a, b\rangle$. For each $i = 1, 2, 3, 4$, we then define $E_i$ to be the set of all points of the form $wx$, where $x \in E$ and $w \in \langle a, b\rangle$ is a word such that

- $w$ is the identity, or begins with $a$ (if $i = 1$);
- $w$ begins with $a^{-1}$ (if $i = 2$);
- $w$ begins with $b$ (if $i = 3$);
- $w$ begins with $b^{-1}$ (if $i = 4$).

It is then clear that $E_1, E_2, E_3, E_4$ partition $S^2\backslash\Sigma$, while $a^{-1}E_1 \cup aE_2$ and $b^{-1}E_3 \cup bE_4$ both cover $S^2\backslash\Sigma$, as claimed.

Now let us interpret this example using oracles. The free group $\langle a, b\rangle$ is countably enumerable, and so with a countable amount of time and memory, one can construct and store $\Sigma$ without difficulty. A membership oracle $E()$ for $E$ can then be constructed by an adaptive oracle much as in the previous sections; $E(x)$ returns yes if $x$ is the first point in its orbit that is queried, and returns no otherwise. The oracles $E_i(x)$ can then be defined by querying $E$ for all the points in $x$'s orbit in some arbitrary order until one finds the point $w^{-1}x$ which lies in $E$, and then deciding membership in $E_i(x)$ depending on the first symbol of $w$. (For instance, if $x$'s orbit has not been queried before, the first point in the orbit that one queries $E()$ for will lie in $E$; thus one sees that the order in which one decides to search through the orbit will in fact influence quite strongly what $E$ and the $E_i$ will look like.)

It is then not hard to show that $E_1(), E_2(), E_3(), E_4()$ do indeed partition $S^2\backslash\Sigma$ in the sense that for each $x \in S^2\backslash\Sigma$, exactly one of $E_1(x), E_2(x), E_3(x), E_4(x)$ will return "yes". Also, for any $x \in S^2\backslash\Sigma$, at least one of $E_1(ax), E_2(a^{-1}x)$ will return "yes", and at least one of $E_3(bx), E_4(b^{-1}x)$ will return "yes". Thus, after performing an uncountable number of queries to fully complete all of these sets, we obtain sets obeying the properties in Proposition 1.12.6; but the oracles that do this are perfectly deterministic, and run in only a countable amount of time, at least for the first few queries (and one can probably even trim it down to a finite amount of time, if one has an efficient way of deciding whether two points lie in the same orbit of $\langle a, b\rangle$).

Now we discuss how a Banach-Tarski type construction does not work in one or two dimensions. Let us just consider the one-dimensional case:

**Proposition 1.12.7.** *There does not exist a decomposition of the unit interval $[0, 1)$ into finitely many sets $E_1, \ldots, E_k$ such that some translates $E_1 + x_1, \ldots, E_k + x_k$ of these sets cover $[0, 2)$.*

We briefly review the proof of this proposition. Suppose for contradiction that we have such a decomposition. We can colour the lattice $\mathbf{Z}^k$ in $k$ colours $1, \ldots, k$, by giving each $k$-tuple $(n_1, \ldots, n_k)$ the colour $i$ if $n_1x_1 + \ldots + n_kx_k \bmod 1 \in E_i$. This is clearly a $k$-colouring of $\mathbf{Z}^k$. Furthermore, for every $k$-tuple $\vec{n}$, we see that at $\vec{n} - e_i$ has colour $i$ for at least two values of $i = 1, \ldots, k$, where $e_1, \ldots, e_k$ is the standard basis for $\mathbf{Z}^k$. If one then uses *double-counting* to get two different estimates for the number of coloured points in a box $\{1, \ldots, N\}^k$ for a large enough $N$, one obtains a contradiction.

Note that this proof is quite finitary; given some real numbers $x_1, \ldots, x_k$ and some membership oracles $E_1(), \ldots, E_k()$, one could convert the above

argument into an algorithm that would be able to demonstrate in finite time that either the oracles $E_1(), \ldots, E_k()$ fail to partition $[0,1)$, or that the translates $E_1() + x_1, \ldots, E_k() + x_k$ fail to cover $[0,2)$; we leave this as an exercise to the reader.

**Remark 1.12.8.** The key distinction here between the low and high dimensional cases is that the free group $\langle a, b \rangle$ is not *amenable*, whereas $\mathbf{Z}^k$ is amenable. See [**Ta2010**, §2.2, §2.8] for further discussion.

**1.12.5. Summary.** The above discussion suggests that it is possible to retain much of the essential mathematical content of set theory without the need for explicitly dealing with large sets (such as uncountable sets), but there is a significant price to pay in doing so, namely that one has to deal with sets on a "virtual" or "incomplete" basis, rather than with the "completed infinities" that one is accustomed to in the standard modern framework of mathematics. Conceptually, this marks quite a different approach to mathematical objects, and assertions about such objects; such assertions are not simply true or false, but instead require a certain computational cost to be paid before their truth can be ascertained. This approach makes the mathematical reasoning process look rather strange compared to how it is usually presented, but I believe it is still a worthwhile exercise to try to translate mathematical arguments into this computational framework, as it illustrates how some parts of mathematics are in some sense "more infinitary" than others, in that they require a more infinite amount of computational power in order to model in this fashion. It also illustrates why we adopt the conveniences of infinite set theory in the first place; while it is technically possible to do mathematics without infinite sets, it can be significantly more tedious and painful to do so.

*Chapter 2*

# Group theory

## 2.1. Torsors

Given a (multiplicative) group $G$, a (left) *$G$-space* is a space $X$ of states, together with an *action* of the group $G$ that allows each group element $g \in G$ to transform any given state $x \in X$ to another state $gx \in X$, in a manner compatible with the group law (in particular, $ex = x$ for the group identity $e$, and $(gh)x = g(hx)$ for group elements $g, h$). One often also imposes additional compatibility conditions with other structures on the space (e.g. topological, differential, or algebraic structure).

A special case of a $G$-space is a *principal $G$-homogeneous space* or a *$G$-torsor*, defined as a $G$-space which is *uniquely transitive*, i.e. given any two states $x, y \in X$ there is a unique group element $g \in G$ such that $gx = y$; inspired by this, one can write $g$ as $y/x$. A $G$-torsor can be viewed as a copy of the original group $G$, but one that does not necessarily have a preferred identity element[1]

Many natural concepts in mathematics and physics are more naturally torsor elements than groups. Consider for instance the concept of *length*. In mathematics, one often ignores issues of units, and regards the length of a line segment as taking values in the non-negative real line $\mathbf{R}^+$; but in the absence of a preferred unit length scale, it is actually more natural to view length as taking values in some $\mathbf{R}^+$-torsor, say $\mathcal{L}$. To extract a non-negative real number for the length $|AB|$ of a line segment $AB$, one has to divide by some unit length $U \in \mathcal{L}$, such as a unit foot or a unit yard. For instance, if $AB$ is 30 feet long, and $U$ is a unit foot, then $|AB|/U = 30$.

[1]If there is a preferred identity or origin element $O$, then one can place the $G$-torsor in one-to-one correspondence with $G$ by identifying $gO$ with $g$ for every group element $g$.

Observe that changing units is a passive transformation (see Section 2.2 below) rather than an active one, and as such behaves in the inverse manner to what one might naively expect. For instance, if one changes the unit of length $U$ from feet to yards, which is a unit that is three times larger, then the numerical length $|AB|/U$ of $AB$ *shrinks* by a factor of 3: $AB$ is 30 feet long, but is only 10 yards long. Thus, while a unit yard is three times longer than a unit foot, the yard coordinate (the dual coordinate to the unit yard, which converts lengths to positive real numbers) is one third of the foot coordinate. (See Section 6.3 for further discussion.)

More generally, one can use torsors to rigorously set up the physical concepts of *units* and *dimensional analysis*. The product of two lengths in $\mathcal{L}$ is not another length, but instead takes values in another torsor, the torsor $\mathcal{L}^2 = \mathcal{L} \otimes_{\mathbf{R}^+} \mathcal{L}$ of areas. One can use the square $U^2$ of the unit length as a unit area. The assertion that physical laws have to be dimensionally consistent is then equivalent to the assertion that they are invariant with respect to the passive transformation of changing the units.

Much as dimensionful units such as length or mass are torsors for the non-negative reals, points in space are torsors for the translation group $\mathbf{R}^3$, and (oriented) spatial coordinate frames are torsors for the linear group[2] $SL_3(\mathbf{R})$ (if the origin is fixed) or $SL_3(\mathbf{R}) \ltimes \mathbf{R}^3$ (otherwise). And so forth. Indeed, one can view basis vectors and coordinate systems as higher-dimensional analogues of units and unit measurement coordinates respectively.

If one works with spacetime coordinate frames rather than spatial coordinate frames, then the situation is similar, but the structure group will be different (e.g. the Galilean group for Galilean relativity, the Poincare or Lorentz group for special relativity, or the diffeomorphism group for general relativity).

Viewing group elements as quotients of torsors is sometimes helpful when trying to visualise operations such as conjugation $h \to ghg^{-1}$; one can interpret this operation as that of moving both the observer and the object by g. For instance, consider the *lamplighter group* $\mathbf{Z}/2\mathbf{Z} \wr \mathbf{Z}$, the *wreath product* of $\mathbf{Z}/2\mathbf{Z}$ with $\mathbf{Z}$. One can define this group by using as a state space $X$ the configuration space of a doubly infinite sequence of lamps (indexed by the integers), with each lamp being either "on" or "off", and with at most finitely many of the lamps being "on", together with the position of a lamplighter, located at one of the lamps; more formally, we have $X := (\mathbf{Z}/2\mathbf{Z})_0^{\mathbf{Z}} \times \mathbf{Z}$, where $(\mathbf{Z}/2\mathbf{Z})_0^{\mathbf{Z}}$ is the space of compactly supported sequences from $\mathbf{Z}$ to $\mathbf{Z}/2\mathbf{Z}$. The lamplighter has the ability to toggle the lamp on and off at his or her current location, and also has the ability to move left or right. The

[2]If one insists on the coordinate frames being orthogonal, then the relevant group becomes $SO_3(\mathbf{R})$ or the Euclidean group $SO_3(\mathbf{R}) \ltimes R^3$, as appropriate.

lamplighter group $G$ is then the group of transformations on the state space $X$ that is generated by the following operations:

- $e$: Move the lamplighter one unit to the right.
- $e^{-1}$: Move the lamplighter one unit to the left.
- $f$: Toggle the lamp at the current location of the lamplighter.

It is not hard to show that $X$ then becomes a $G$-torsor.

One way to describe a group element of $G$ is then to describe an initial state $A$ in $X$ and a final state $B$ in $X$, and then define $B/A$ to be the unique group element that transforms $A$ to $B$; one can view $B/A$ as a "program" (e.g. made up of a string of $e$'s, $e^{-1}$'s, and $f$'s, or perhaps expressed in a more "high-level" language) that one could give to a lamplighter that is currently in the system $A$ that would then transform it to $B$. Note that multiple programs can give the same group element, for instance $fefe^{-1}$ is the same element of $G$ as $efe^{-1}f$. Also, multiple pairs $A$, $B$ can give rise to the same element[3] $B/A$.

One can express any such "program" $B/A$ in a canonical form as that of "change the set $S$ of lights, as described using one's current location, and then move $n$ steps to the right (or left)". This expresses the lamplighter group as a *semi-direct product* of $(\mathbf{Z}/2\mathbf{Z})_0^{\mathbf{Z}}$ and $\mathbf{Z}$. If the lamplighter position does not change between $A$ and $B$, then the program is simply that of changing a set of lights, and $B/A$ now lives in the abelian subgroup $(\mathbf{Z}/2\mathbf{Z})_0^{\mathbf{Z}}$ of the lamplighter group.

If one conjugates a group element $B/A$ by another group element $g$, one obtains the new group element $g(B/A)g^{-1} = (gB)/(gA)$. A little thought then reveals that the program needed to execute $(gB)/(gA)$ is similar to that for $B/A$, except that the set of lights $S$ that one needs to change has been modified. As such, we see that the commutator $[g, B/A] = ((gB)/(gA))/(B/A)$ is an element of the abelian subgroup $(Z/2Z)_0^Z$, making the lamplighter group *metabelian* and thus *solvable*.

I found this sort of torsor-oriented perspective useful when thinking about such concepts as that of a harmonic function on a group G (something that comes up, for instance, in modern proofs of Gromov's theorem regarding groups of polynomial growth, see Section 2.5). One can instead think about a harmonic function on a $G$-torsor $X$, defined as an "energy functional" on such a space with the property that the energy of any state is equal to the average energy of the neighbouring states (at least if the group $G$ is discrete; for continuous groups, one has to neglect higher order terms). If the group $G$ is not commutative, then actively transforming the

[3]This perspective is a generalisation of the standard way of visualising a spatial vector as an arrow from one spatial point $A$ to another spatial point $B$.

states can destroy the harmonicity property; but passively transforming the states does not. It is because of this that the space of (right)-harmonic functions still has a left $G$-action, and vice versa.

## 2.2. Active and passive transformations

Consider the following (somewhat informally posed) questions:

**Question 2.2.1.** *Let $T$ be equilateral triangle in a plane whose vertices are labeled 1, 2, 3 in clockwise order. Define the following two operations on this triangle:*

- *$F$: Flip the triangle to swap the vertices 2 and 3, while keeping 1 fixed.*
- *$R$: Rotate the triangle clockwise in the plane by 120 degrees.*

*Do the operations $F$ and $R$ commute, i.e. does $F \circ R = R \circ F$?*

**Question 2.2.2.** *Suppose one is viewing some text on a computer screen. The text is so long that one cannot display all of it at once on a screen: currently, only a middle portion of the text is visible. To see more of the text, we press the "up" arrow key (or click the "up" button). When one does so, does the text on the screen move up or down?*

We discuss Question 2.2.2 first. The answer depends on the user interface model. Most such models are *passive transformations*; the "up" command moves the *observer* up, and one's view of the text then moves *down* as a consequence. A minority of models (such as "hand" tools in various pieces of software) are instead *active transformations*; dragging a hand tool upwards causes the text to move upward (keeping the observer position fixed).

In some cases, the model used may be ambiguous at first. If one is viewing a map, does the "+" key cause the map image on the screen to enlarge, or shrink? Somewhat confusingly, two different pieces of mapping software can respond in opposite ways to such a command; some use active transformation models ("+" makes the world bigger, so that less of the world remains inside the viewscreen), and others use passive transformation models ("+" makes the *observer* bigger, so that more of the world can now fit inside the viewscreen).

This question is a special case of the double action of a group $G$ on itself (or more generally, on a left $G$-torsor, as discussed in Section 2.1). Imagine that inside a group $G$ (or left $G$-torsor) one has an observer $O$ and an object $X$; there exists a unique group element $g = X/O$ such that $X = gO$, and in that case we say that $X$ has an *apparent position* of $g$ from the perspective of the observer $O$.

We can then change this apparent position in two different ways. Firstly, we may apply an active transformation, and shift the object $X$ by a group element $h$ to move it to $hX = hgO$; the apparent position of the object then shifts from $g$ to $hg$, and so the active transformation corresponds to the *left* action of the group $G$ on itself.

Or, we may apply a passive transformation, and shift the *observer* $O$ by a group element $h$ to move it to $hO$; since $X = gO = gh^{-1}(hO)$, we see that the apparent position shifts from $g$ to $gh^{-1}$. Thus the passive transformation corresponds to the *right* action of $G$ on itself.

Note the presence of the inverse in the passive transformation; it is this inverse which is the source of the confusions mentioned above.

The distinction between active and passive transformations also arises when trying to direct the motion of another person. "Move left! No, your *other* left!"

Even if the group $G$ is non-commutative, the left-action of $G$ and the right-action of $G$ will still commute with each other, as moving an object and moving the observer can be done in either order without any difference to the final state of the system.

And now we can answer Question 2.2.1. The answer is that it depends on whether one interprets the operations $F$ and $R$ as active operations or passive operations; the phrasing of the question makes it ambiguous.

For instance, we can treat the flip operation $F$ as an active transformation, by viewing the labels 1, 2, 3 as being attached[4] to the triangle object being manipulated. As one applies $F$, the labels 1 and 2 physically change locations.

Or, one can view the flip operation $F$ as a passive transformation, caused by motion of the observer rather than the object. Here, the labels 1, 2, 3 are attached to the observer rather than to the object (and are usually displayed *outside* the triangle). The operation $F$ flips the triangle, but the labels 1, 2, 3 remain where they are.

The difference between the two interpretations of $F$ becomes apparent once the object's value of 3 moves away from the observer's value of 3, as they are then flipping the triangle across different axes.

Similarly, one can view the rotation operation $R$ as an active rotation, in which the triangle is physically rotated in the direction which is clockwise in its own orientation, or as a passive rotation in which the triangle is rotated in the direction which is clockwise in the observer's orientation (or equivalently, the observer is rotated in a *counter-clockwise* direction).

[4]This is usually drawn by placing the labels 1, 2, 3 *inside* the triangle.

The difference between the two interpretations of $R$ becomes apparent once the triangle is flipped over, so that its orientation is the opposite of that of the observer.

If $F$ is active and $R$ is passive (or vice versa), then the transformations commute. But if $F$ and $R$ are both active or both passive, then the transformations do not commute.

In order to fully remove all ambiguity from the system, one needs to label the vertices of the triangle *twice*: first by an "external" labeling (e.g. $A$, $B$, $C$) which is not affected by any of the transformations, and secondly by an "internal" labeling (e.g. 1, 2, 3) which moves with the operations being applied. Traditionally, the external labels are displayed outside the triangle, and the internal vertices are displayed inside the triangle. Similarly, one needs to display an external orientation (e.g. a counterclockwise arrow, displayed outside the triangle) that is not affected by the operations being applied, and also an internal orientation (e.g. another counterclockwise arrow, displayed inside the triangle) that can get flipped over by the operations being applied. There are then four operations of interest:

(1) Active-$F$: Flip the triangle across the axis given by the vertex internally labeled 3, thus swapping the vertices internally labeled 1, 2, and also reversing the internal orientation.

(2) Passive-$F$: Flip the triangle across the axis given by the vertex externally labeled $C$, thus swapping the internal labels of the vertices that are externally labeled $A$, $B$, and also reversing the internal orientation.

(3) Active-$R$: Rotate the triangle by 120 degrees in the internally clockwise direction, moving the internal labels appropriately.

(4) Passive-$R$: Rotate the triangle by 120 degrees in the externally clockwise direction, moving the internal labels appropriately.

Then the two passive transformations commute with the two active transformations, but the two passive transformations do not commute with each other, and neither do the two active transformations. (It is instructive to work this out on paper, physically cutting out a triangle if necessary.)

## 2.3. Cayley graphs and the geometry of groups

In most undergraduate courses, groups are first introduced as a primarily *algebraic* concept - a set equipped with a number of algebraic operations (group multiplication, multiplicative inverse, and multiplicative identity) and obeying a number of rules of algebra (most notably the associative law). It is only somewhat later that one learns that groups are not solely an algebraic object, but can also be equipped with the structure of a manifold

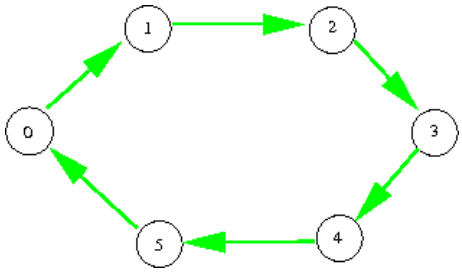


**Figure 1.** Cayley graph of $\mathbf{Z}/6\mathbf{Z}$ with generator 1 (in green).

(giving rise to *Lie groups*) or a topological space (giving rise to *topological groups*). (See also [**Ta2010b**, §1.14] for a number of other ways to think about groups.)

Another important way to enrich the structure of a group $G$ is to give it some *geometry*. A fundamental way to provide such a geometric structure is to specify a list of generators $S$ of the group $G$. Let us call such a pair $(G, S)$ a *generated group*; in many important cases the set of generators $S$ is finite, leading to a *finitely generated group*. A generated group $(G, S)$ gives rise to the *word metric* $d : G \times G \to \mathbf{N}$ on $G$, defined to be the maximal metric for which $d(x, sx) \leq 1$ for all $x \in G$ and $s \in S$ (or more explicitly, $d(x, y)$ is the least $m$ for which $y = s_1^{\epsilon_1} \dots s_m^{\epsilon_m} x$ for some $s_1, \dots, s_m \in S$ and $\epsilon_1, \dots, \epsilon_m \in \{-1, +1\}$). This metric then generates the balls $B_S(R) := \{x \in G : d(x, \mathrm{id}) \leq R\}$. In the finitely generated case, the $B_S(R)$ are finite sets, and the rate at which the cardinality of these sets grow in $R$ is an important topic in the field of *geometric group theory*. The idea of studying a finitely generated group via the geometry of its metric goes back at least to the work of Dehn [**De1912**].

One way to visualise the geometry of a generated group is to look at the (labeled) *Cayley colour graph* (or *Cayley graph*, for short) of the generated group $(G, S)$. This is a directed coloured graph, with edges coloured by the elements of $S$, and vertices labeled by elements of $G$, with a directed edge of colour $s$ from $x$ to $sx$ for each $x \in G$ and $s \in S$. The word metric then corresponds to the graph metric of the Cayley graph. See for instance Figure 1 and Figure 2.

We can thus see that the same group can have somewhat different geometry if one changes the set of generators. For instance, in a large cyclic group $\mathbf{Z}/N\mathbf{Z}$, with a single generator $S = \{1\}$ the Cayley graph "looks one-dimensional", and balls $B_S(R)$ grow linearly in $R$ until they saturate the entire group, whereas with two generators $S = \{s_1, s_2\}$ chosen at random, the Cayley graph "looks two-dimensional", and the balls $B_S(R)$ typically grow quadratically until they saturate the entire group.

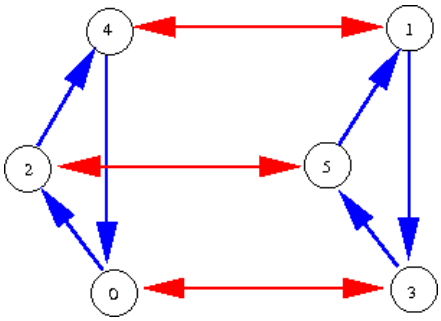


**Figure 2.** Cayley graph of $\mathbf{Z}/6\mathbf{Z}$ with generators 2 (in blue) and 3 (in red).

Cayley graphs have three distinguishing properties:

- (Regularity) For each colour $s \in S$, every vertex $x$ has a single $s$-edge leading out of $x$, and a single $s$-edge leading into $x$.
- (Connectedness) The graph is connected.
- (Homogeneity) For every pair of vertices $x, y$, there is a unique coloured graph isomorphism that maps $x$ to $y$.

It is easy to verify that a directed coloured graph is a Cayley graph (up to relabeling) if and only if it obeys the above three properties. Indeed, given a graph $(V, E)$ with the above properties, one sets $G$ to equal the (coloured) automorphism group of the graph $(V, E)$; arbitrarily designating one of the vertices of $V$ to be the identity element id, we can then identify all the other vertices in $V$ with a group element. One then identifies each colour $s \in S$ with the vertex that one reaches from id by an $s$-coloured edge. Conversely, every Cayley graph of a generated group $(G, S)$ is clearly regular, is connected because $S$ generates $G$, and has isomorphisms given by right multiplication $x \mapsto xg$ for all $g \in G$. (The regularity and connectedness properties already ensure the uniqueness component of the homogeneity property.)

From the above equivalence, we see that we do not really need the vertex labels on the Cayley graph in order to describe a generated group, and so we will now drop these labels and work solely with *unlabeled* Cayley graphs, in which the vertex set is not already identified with the group. As we saw above, one just needs to designate a marked vertex of the graph as the "identity" or "origin" in order to turn an unlabeled Cayley graph into a labeled Cayley graph; but from homogeneity we see that all vertices of an unlabeled Cayley graph "look the same" and there is no canonical preference for choosing one vertex as the identity over another. I prefer here to keep the graphs unlabeled to emphasise the homogeneous nature of the graph.

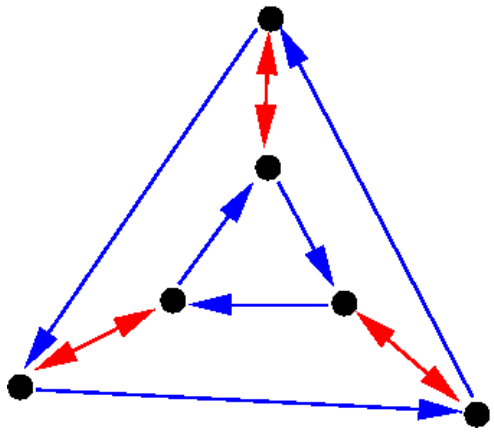

**Figure 3.** Cayley graph of $S_3$.

It is instructive to revisit the basic concepts of group theory using the language of (unlabeled) Cayley graphs, and to see how geometric many of these concepts are. In order to facilitate the drawing of pictures, I work here only with small finite groups (or Cayley graphs), but the discussion certainly is applicable to large or infinite groups (or Cayley graphs) also.

For instance, in this setting, the concept of *abelianness* is analogous to that of a *flat* (zero-curvature) geometry: given any two colours $s_1, s_2$, a directed path with colours $s_1, s_2, s_1^{-1}, s_2^{-1}$ (adopting the obvious convention that the reversal of an $s$-coloured directed edge is considered an $s^{-1}$-coloured directed edge) returns to where it started[5]. Thus, for instance, the two depictions of $\mathbf{Z}/6\mathbf{Z}$ above are abelian, whereas the group $S_3$, which is also the dihedral group of the triangle, and thus admits the Cayley graph depicted in Figure 3, is not abelian.

A subgroup $(G', S')$ of a generated group $(G, S)$ can be easily described in Cayley graph language if the generators $S'$ of $G'$ happen to be a subset of the generators $S$ of $G$. In that case, if one begins with the Cayley graph of $(G, S)$ and erases all colours except for those colours in $S'$, then the graph *foliates* into connected components, each of which is isomorphic to the Cayley graph of $(G', S')$. For instance, in the above Cayley graph depiction of $S_3$, erasing the blue colour leads to three copies of the red Cayley graph (which has $\mathbf{Z}/2\mathbf{Z}$ as its structure group), while erasing the red colour leads to two copies of the blue Cayley graph (which as $A_3 \equiv \mathbf{Z}/3\mathbf{Z}$ as its structure group). If $S'$ is not contained in $S$, then one has to first "change basis" and add or remove some coloured edges to the original Cayley graph before one can obtain this formulation (thus for instance $S_3$ contains two

[5]Note that a generated group $(G, S)$ is abelian if and only if the generators in $S$ pairwise commute with each other.

more subgroups of order two that are not immediately apparent with this choice of generators). Nevertheless the geometric intuition that subgroups are analogous to foliations is still quite a good one.

We saw that a subgroup $(G', S')$ of a generated group $(G, S)$ with $S' \subset S$ foliates the larger Cayley graph into $S'$-connected components, each of which is a copy of the smaller Cayley graph. The remaining colours in $S$ then join those $S'$-components to each other. In some cases, each colour $s \in S \backslash S'$ will connect a $S'$-component to exactly one other $S'$-component; this is the case for instance when one splits $S_3$ into two blue components. In other cases, a colour $s$ can connect a $S'$-component to multiple $S'$-components; this is the case for instance when one splits $S_3$ into three red components. The former case occurs precisely when[6] the subgroup $G'$ is *normal*. We can then *quotient out* the $(G', S')$ Cayley graph from $(G, S)$, leading to a quotient Cayley graph $(G/G', S \backslash S')$ whose vertices are the $S'$-connected components of $(G, S)$, and the edges are projected from $(G, S)$ in the obvious manner. We can then view the original graph $(G, S)$ as a *bundle* of $(G', S')$-graphs over a base $(G/G', S \backslash S')$-graph (or equivalently, an *extension* of the base graph $(G/G', S \backslash S')$ by the fibre graph $(G', S')$); for instance $S_3$ can be viewed as a bundle of the blue graph $A_3$ over the red graph $\mathbf{Z}/2\mathbf{Z}$, but not conversely. We thus see that the geometric analogue of the concept of a normal subgroup is that of a *bundle*. The generators in $S \backslash S'$ can be viewed as describing a *connection* on that bundle.

Note, though, that the structure group of this connection is not simply $G'$, unless $G'$ is a *central* subgroup; instead, it is the larger group $G' \rtimes \mathrm{Aut}(G')$, the semi-direct product of $G'$ with its automorphism group. This is because a non-central subgroup $G'$ can be "twisted around" by operations such as conjugation $g' \mapsto sg's^{-1}$ by a generator $s \in S$. So central subgroups are analogous to the geometric notion of a *principal bundle*. For instance, Figure 4 depicts the Heisenberg group

$$\begin{pmatrix} 1 & \mathbf{F}_2 & \mathbf{F}_2 \\ 0 & 1 & \mathbf{F}_2 \\ 0 & 0 & 1 \end{pmatrix}$$

over the field $\mathbf{F}_2$ of two elements, which one can view as a central extension of $\mathbf{F}_2^2$ (the blue and green edges, after quotienting) by $\mathbf{F}_2$ (the red edges). Note how close this group is to being abelian; more generally, one can think of nilpotent groups as being a slight perturbation of abelian groups.

In the case of $S_3$ (viewed as a bundle of the blue graph $A_3$ over the red graph $\mathbf{Z}/2\mathbf{Z}$), the base graph $\mathbf{Z}/2\mathbf{Z}$ is in fact embedded (three times) into the large graph $S_3$. More generally, the base graph $(G/G', S \backslash S')$ can be lifted

[6]Note that a subgroup $G'$ of a generated group $(G, S)$ is normal if and only if left-multiplication by a generator of $S$ maps right-cosets of $G'$ to right-cosets of $G'$.

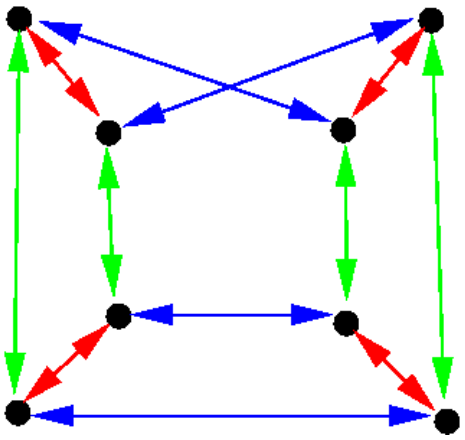

**Figure 4.** The Heisenberg group over $\mathbf{F}_2$.

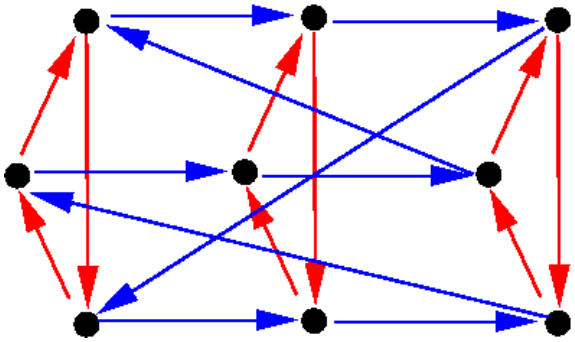

**Figure 5.** The group $\mathbf{Z}/9\mathbf{Z}$, with generators 1 (in blue) and 3 (in red).

back into the extension $(G, S)$ if and only if the short exact sequence $0 \to G' \to G \to G/G' \to 0$ *splits*, in which case $G$ becomes a *semidirect product* $G \equiv G' \rtimes H$ of $G'$ and a lifted copy $H$ of $G/G'$. Not all bundles can be split in this fashion. For instance, consider the group $\mathbf{Z}/9\mathbf{Z}$ depicted in Figure 5. This is a $\mathbf{Z}/3\mathbf{Z}$-bundle over $\mathbf{Z}/3\mathbf{Z}$ that does not split; the blue Cayley graph of $\mathbf{Z}/3\mathbf{Z}$ is not visible in the $\mathbf{Z}/9\mathbf{Z}$ graph directly, but only after one quotients out the red fibre subgraph. The notion of a splitting in group theory is analogous to the geometric notion of a *global gauge* (see [**Ta2009b**, §1.4]). The existence of such a splitting or gauge, and the relationship between two such splittings or gauges, are controlled by the *group cohomology* of the sequence $0 \to G' \to G \to G/G' \to 0$.

Even when one has a splitting, the bundle need not be completely trivial, because the bundle is not principal, and the connection can still twist the fibres around. For instance, $S_3$ when viewed as a bundle over $\mathbf{Z}/2\mathbf{Z}$ with

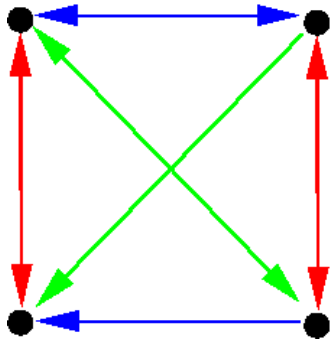

**Figure 6.** The Klein four-group $\mathbf{Z}/2\mathbf{Z} \times \mathbf{Z}/2\mathbf{Z}$..

fibres $A_3$ splits, but observe that if one uses the red generator of this splitting to move from one copy of the blue $A_3$ graph to the other, that the orientation of the graph changes. The bundle is trivialisable if and only if $G'$ is a *direct summand* of $G$, i.e. $G$ splits as a direct product $G = G' \times H$ of a lifted copy $H$ of $G/G'$. Thus we see that the geometric analogue of a direct summand is that of a trivialisable bundle (and that trivial bundles are then the analogue of direct products). Note that there can be more than one way to trivialise a bundle. For instance, with the Klein four-group $\mathbf{Z}/2\mathbf{Z} \times \mathbf{Z}/2\mathbf{Z}$ depicted in Figure 6, the red fibre $\mathbf{Z}/2\mathbf{Z}$ is a direct summand, but one can use either the blue lift of $\mathbf{Z}/2\mathbf{Z}$ or the green lift of $\mathbf{Z}/2\mathbf{Z}$ as the complementary factor.

## 2.4. Group extensions

In mathematics, one frequently starts with some space $X$ and wishes to *extend* it to a larger space $Y$. Generally speaking, there are two ways in which one can extend a space $X$:

- By *embedding* $X$ into a space $Y$ that has $X$ (or at least an isomorphic copy of $X$) as a *subspace*.
- By *covering* $X$ by a space $Y$ that has $X$ (or an isomorphic copy thereof) as a *quotient*.

For many important categories of interest (such as *abelian categories*), the former type of extension can be represented by the *exact sequence*,

$$0 \to X \to Y$$

and the latter type of extension be represented by the exact sequence

$$Y \to X \to 0.$$

In some cases, $X$ can be both embedded in, and covered by, $Y$, in a consistent fashion; in such cases we sometimes say that the above exact sequences *split*.

An analogy would be to that of digital images. When a computer represents an image, it is limited both by the scope of the image (what it is picturing), and by the resolution of an image (how much physical space is represented by a given pixel). To make the image "larger", one could either *embed* the image in an image of larger scope but equal resolution (e.g. embedding a picture of a $200 \times 200$ pixel image of person's face into a $800 \times 800$ pixel image that covers a region of space that is four times larger in both dimensions, e.g. the person's upper body) or *cover* the image with an image of higher resolution but of equal scope (e.g. enhancing a $200 \times 200$ pixel picture of a face to a $800 \times 800$ pixel of the same face). In the former case, the original image is a *sub-image* (or *cropped image*) of the extension, but in the latter case the original image is a *quotient* (or a *pixelation*) of the extension. In the former case, each pixel in the original image can be identified with a pixel in the extension, but not every pixel in the extension is covered. In the latter case, every pixel in the original image is *covered* by several pixels in the extension, but the pixel in the original image is not canonically identified with any particular pixel in the extension that covers it; it "loses its identity" by dispersing into higher resolution pixels.

**Remark 2.4.1.** Note that "zooming in" the visual representation of an image by making each pixel occupy a larger region of the screen neither increases the scope or the resolution; in this language, a zoomed-in version of an image is merely an *isomorphic copy* of the original image; it carries the same amount of information as the original image, but has been represented in a new *coordinate system* which may make it easier to view.

In the study of a given category of spaces (e.g. topological spaces, manifolds, groups, fields, etc.), embedding and coverings are both important; this is particularly true in the more topological areas of mathematics, such as manifold theory. But typically, the term *extension* is reserved for just one of these two operations. For instance, in the category of fields, coverings are quite trivial; if one covers a field $k$ by a field $l$, the kernel of the covering map $\pi : l \to k$ is necessarily trivial and so $k, l$ are in fact isomorphic. So in field theory, a *field extension* refers to an embedding of a field, rather than a covering of a field. Similarly, in the theory of metric spaces, there are no non-trivial isometric coverings of a metric space, and so the only useful notion of an extension of a metric space is the one given by embedding the original space in the extension.

On the other hand, in group theory (and in group-like theories, such as the theory of dynamical systems, which studies group actions), the term "extension" is reserved for coverings, rather than for embeddings. I think one of the main reasons for this is that coverings of groups automatically generate a special type of embedding (a *normal* embedding), whereas most

embeddings don't generate coverings. More precisely, given a group extension $G$ of a base group $H$,

$$G \to H \to 0,$$

one can form the *kernel* $K = \ker(\phi)$ of the covering map $\pi : G \to H$, which is a normal subgroup of $G$, and we thus can extend the above sequence canonically to a *short exact sequence*

$$0 \to K \to G \to H \to 0.$$

On the other hand, an embedding of $K$ into $G$,

$$0 \to K \to G$$

does not similarly extend to a short exact sequence unless the the embedding is normal.

Another reason for the notion of extension varying between embeddings and coverings from subject to subject is that there are various natural *duality operations* (and more generally, *contravariant functors*) which turn embeddings into coverings and vice versa. For instance, an embedding of one vector space $V$ into another $W$ induces a covering of the *dual space* $V^*$ by the dual space $W^*$, and conversely; similarly, an embedding of a locally compact abelian group $H$ in another $G$ induces a covering of the *Pontryagin dual* $\hat{H}$ by the Pontryagin dual $\hat{G}$. In the language of images, embedding an image in an image of larger scope is largely equivalent to covering the Fourier transform of that image by a transform of higher resolution, and conversely; this is ultimately a manifestation of the basic fact that frequency is inversely proportional to wavelength.

Similarly, a common duality operation arises in many areas of mathematics by starting with a space $X$ and then considering a space $C(X)$ of functions on that space (e.g. continuous real-valued functions, if $X$ was a topological space, or in more algebraic settings one could consider homomorphisms from $X$ to some fixed space). Embedding $X$ into $Y$ then induces a covering of $C(X)$ by $C(Y)$, and conversely, a covering of $X$ by $Y$ induces an embedding of $C(X)$ into $C(Y)$. Returning again to the analogy with images, if one looks at the collection of *all* images of a fixed scope and resolution, rather than just a single image, then increasing the available resolution causes an *embedding* of the space of low-resolution images into the space of high-resolution images (since of course every low-resolution image is an example of a high-resolution image), whereas increasing the available scope causes a *covering* of the space of narrow-scope images by the space of wide-scope images (since every wide-scope image can be *cropped* into a narrow-scope image). Note in the case of images, that these extensions can be split: not only can a low-resolution image be viewed as a special case of a high-resolution image, but any high-resolution image can be *pixelated*

into a low-resolution one. Similarly, not only can any wide-scope image be cropped into a narrow-scope one, a narrow-scope image can be extended to a wide-scope one simply by filling in all the new areas of scope with black (or by using more advanced image processing tools to create a more visually pleasing extension). In the category of sets, the statement that every covering can be split is precisely the *axiom of choice*.

I've recently found myself having to deal quite a bit with group extensions in my research, so I have decided to make some notes on the basic theory of such extensions. This is utterly elementary material for a group theorist, but I found this task useful for organising my own thoughts on this topic, and also in pinning down some of the jargon in this field.

### 2.4.1. Basic concepts.

**Definition 2.4.2** (Group extension)**.** An *extension* of a group $H$ is a group $G$, together with a surjective *projection map* (or *covering map*) $\pi : G \to H$. If the kernel of $\pi$ can be identified with (i.e. is isomorphic to) a group $K$, we say that $G$ is an *extension* of $H$ by $K$, and we have the short exact sequence

$$0 \to K \to G \to H \to 0.$$

If the group $K$ has some property $\mathcal{P}$, we say that $G$ is a $\mathcal{P}$ extension of $H$. Thus for instance, if $K$ is abelian, $G$ is an abelian extension of $H$; if $K$ is central (in $G$), $G$ is a central extension of $H$; and so forth. We refer to $H$ as the *base* of the extension, and $K$ as the *fibre*, and refer to $H$ and $K$ collectively as *factors* of $G$.

If $K$ has some property $\mathcal{P}$, and $H$ has some property $\mathcal{Q}$, then we say that[7] $G$ is $\mathcal{P}$-*by*-$\mathcal{Q}$. Thus, for instance, $G$ is abelian-by-finite if $K$ is abelian and $H$ is finite, but finite-by-abelian if $K$ is finite and $H$ is abelian.

One can think of a $K$-by-$H$ group as a group that looks like $H$ "at large scales" and like $K$ "at small scales"; one can also view this group as a principal $K$-bundle over $H$.

There are several ways to generate a group extension $G \to H \to 0$. Firstly, given any homomorphism $\pi : G \to G'$ from one group $G$ to another, the *homomorphism theorem* tells us that $G$ is an extension of the image $\pi(G)$, with kernel $\ker(\pi)$:

$$0 \to \ker(\pi) \to G \to \pi(G) \to 0.$$

Conversely, every group extension arises in this manner.

[7] I have no idea why the order is traditionally arranged in this way; I would have thought that extending a $\mathcal{Q}$ group by a $\mathcal{P}$ group would give a $\mathcal{P}$-by-$\mathcal{Q}$ group, rather than the other way around; perhaps at one point the idea of a normal embedding was considered more important than a group extension. Nevertheless, the notation seems to be entrenched by now.

A group extension $\pi : G \to H$ *splits* if there is a homomorphism $\phi : H \to G$ such that $\pi(\phi(h)) = h$ for all $h \in H$. In this case, $H$ acts on the kernel $K$ by conjugation (after identifying $H$ with $\phi(H)$); denoting this action by $\rho$ (thus $\rho(h)k := \phi(h)k\phi(h)^{-1}$), we can then canonically identify $G$ with the *semi-direct product* $K \rtimes_\rho H$, defined as the set of pairs $(k, h)$ with $k \in K$, $h \in H$, with the group law $(k, h)(k', h') := (k\rho(h)(k'), hh')$, by identifying $(k, h)$ with $k\phi(h)$. Conversely, every semi-direct product $K \rtimes_\rho H$ is a group extension of $H$ by $K$ which splits. If the conjugation action $\rho$ is trivial, then the semi-direct product simplifies to the *direct product* $K \times H$. In particular, any semi-direct product which is a central extension is of this form.

Note that, in general, an extension of $H$ by $K$ is a different concept from an extension of $K$ by $H$, because one can have $H$ as a normal subgroup but not as a quotient, or vice versa. For instance, $S_3$ has $A_3$ as a normal subgroup, but not as a quotient; $S_3$ is an extension of $\mathbf{Z}/2\mathbf{Z}$ by $A_3$, but not vice versa. To put it another way, the operator "-by-" is not commutative: $H$-by-$K$ is a different concept from $K$-by-$H$.

A subgroup $L$ of an $K$-by-$H$ group $G$ is automatically an $K'$-by-$H'$ group for some subgroups $H', K'$ of $H, K$ respectively; this is essentially *Goursat's lemma*. Indeed, one can take $K' := K \cap L$ and $H' := \pi(L)$, where $\pi : G \to H$ is the projection map. Furthermore, the index of the subgroup is the product of the index of $H'$ in $H$, and the index of $K'$ in $K$.

Some standard notions in group theory can be defined using group extensions:

(1) A *metabelian group* is the same thing as an abelian-by-abelian group, i.e. an abelian extension of an abelian group.

(2) A *metacyclic group* is the same thing as an cyclic-by-cyclic group, i.e. a cyclic extension of a cyclic group.

(3) A *polycyclic group* is defined recursively by declaring the trivial group to be polycyclic of length 0, and defining a polycyclic group of length $l$ to be an extension of a cyclic group by a polycyclic group of length $l - 1$. Thus polycyclic groups are polycyclic-by-cyclic, where the polycyclic factor has a shorter length than the full group.

(4) A *supersolvable group* is defined recursively by declaring the trivial group to be supersolvable of length 0, and defining a supersolvable group of length $l$ to be a cyclic extension supersolvable group of length $l-1$. Thus supersolvable groups are cyclic-by-supersolvable, where the supersolvable factor has a shorter length that the full group. In other words, supersolvable groups are towers of cyclic extensions.

(5) A *solvable group* is defined recursively by declaring the trivial group to be solvable of length 0, and defining a solvable group of length $l$ to be an extension of an abelian group by a solvable group of length $l-1$. Thus solvable groups are solvable-by-abelian, where the solvable factor has a shorter length. One can equivalently define solvable groups as abelian-by-solvable, where the solvable factor again has a shorter length (because the final term in the *derived series* is abelian and normal). In other words, a solvable group is a tower of abelian extensions.

(6) A *nilpotent group* is defined recursively by declaring the trivial group to be nilpotent of step 0, and defining a nilpotent group of step $s$ to be a central extension of a nilpotent group of step $s-1$, thus nilpotent groups are central-by-nilpotent. In other words, a nilpotent group is a tower of central extensions.

**Remark 2.4.3.** The inclusions here are: cyclic implies abelian implies metabelian implies solvable, cyclic implies metacyclic implies supersolvable implies polycyclic implies solvable, metacyclic implies metabelian, and abelian implies nilpotent implies solvable.

The trivial group is the identity for the "-by-" operator: trivial-by-$\mathcal{P}$ or $\mathcal{P}$-by-trivial is the same thing as $\mathcal{P}$.

Now we comment on the associativity of the "-by-" operator. If $N, H, K$ are groups, observe that an $N$-by-($H$-by-$K$) group (i.e. an extension of an $H$-by-$K$ group by $N$) is automatically an ($N$-by-$H$)-by-$K$ group (i.e. an extension of $K$ by an $N$-by-$H$ group), since if we denote $G$ by the $N$-by-($H$-by-$K$) group, and $\pi$ the quotient map from $G$ to the $H$-by-$K$ group, then $\pi^{-1}(H)$ is a $N$-by-$H$ normal subgroup of $G$ whose quotient is $K$. Thus, for instance, every cyclic-by-metacyclic group is metacyclic-by-cyclic, and more generally every supersolvable group is polycyclic.

On the other hand, the converse is not true: not every ($N$-by-$H$)-by-$K$ group is an $N$-by-($H$-by-$K$) group. The problem is that $N$ is normal in the $N$-by-$H$ group, but need not be normal in the ($N$-by-$H$)-by-$K$ group. For instance, the semi-direct product $\mathbf{Z}^2 \rtimes SL_2(\mathbf{Z})$ is ($\mathbf{Z}$-by-$\mathbf{Z}$)-by-$SL_2(\mathbf{Z})$ but not $\mathbf{Z}$-by-($\mathbf{Z}$-by-$SL_2(\mathbf{Z})$). So the "-by-" operation is not associative in general (for instance, there are polycyclic groups that are not supersolvable). However, if $N$ is not just normal in the $N$-by-$H$ group, but is *characteristic* in that group (i.e. invariant under all (outer) automorphisms of that group), then it is automatically normal in the larger ($N$-by-$H$)-by-$K$ group, and then one can interpret the ($N$-by-$H$)-by-$K$ group as an $N$-by-($H$-by-$K$) group. So one recovers associativity when the first factor is characteristic. This explains why solvable groups can be recursively expressed both

as abelian-by-solvable, and equivalently as solvable-by-abelian; this is ultimately because the *commutator subgroup* $[G,G]$ is a characteristic subgroup of $G$. An easy but useful related observation is that solvable-by-solvable groups are again solvable (with the length of the product being bounded by the sum of the length of the factors).

Given a group property $\mathcal{P}$, a group $G$ is said to be *virtually* $\mathcal{P}$ if it has a finite index subgroup with the property $\mathcal{P}$; thus for instance a virtually abelian group is one with a finite index abelian subgroup, and so forth. As another example, "finite" is the same as "virtually trivial". The property of being virtually $\mathcal{P}$ is not directly expressible in terms of group extensions for arbitrary properties $\mathcal{P}$; however, if the group property $\mathcal{P}$ is *hereditary* in the sense that subgroups of a $\mathcal{P}$ group are also $\mathcal{P}$, then a virtually $\mathcal{P}$ group is the same concept as a $\mathcal{P}$-by-finite group. This is because every finite index subgroup $H$ of a group $G$ automatically contains[8] a finite index *normal* subgroup of $G$.

One also observes that if $\mathcal{P}$, $\mathcal{Q}$ are hereditary properties, then the property of $\mathcal{P}$-by-$\mathcal{Q}$ is hereditary also; if $0 \to P \to G \to Q \to 0$ is a $\mathcal{P}$-by-$\mathcal{Q}$ group, and $G'$ is a subgroup of $G$, then the short exact sequence

$$0 \to (P \cap G') \to G' \to \pi(G') \to 0,$$

where $\pi : G \to Q$ is a projection map, demonstrates that $G'$ is also a $\mathcal{P}$-by-$\mathcal{Q}$ group. Thus for instance the properties of being metabelian, metacyclic, polycyclic, supersolvable, solvable, or nilpotent, are hereditary. As a consequence, virtually nilpotent is the same as nilpotent-by-finite, etc.

We saw for hereditary properties $\mathcal{P}$ that "$\mathcal{P}$-by-finite" was the same concept as "virtually $\mathcal{P}$". It is natural to ask whether the same is true for "finite-by-$\mathcal{P}$". The answer is no; for instance, one can extend the an infinite vector space $V$ over a finite field $F$ by $F$ (using some non-degenerate bilinear anti-symmetric form $\omega : V \times V \to F$, and defining $(v,f)(w,g) = (v+w, f+g+\omega(v,w))$ for $v,w \in V$ and $f,g \in F$) to create a nilpotent group which is finite-by-abelian, but not virtually abelian. Conversely, the semidirect product $\mathbf{Z} \rtimes \mathbf{Z}/2\mathbf{Z}$ (where $\mathbf{Z}/2\mathbf{Z}$ acts on $\mathbf{Z}$ by reflection) is virtually abelian, but not finite-by-abelian. On the other hand, for hereditary $\mathcal{P}$, a finite-by-$\mathcal{P}$ group is virtually (central finite)-by-$\mathcal{P}$. This is because if $G$ is an extension of a $\mathcal{P}$ group $P$ by a finite group $F$, then $G$ acts by conjugation on the finite group $F$; the stabiliser $G'$ of this action is then a finite index subgroup, whose intersection of $F$ is then central in $G'$. The projection of $G'$ onto $P$ is also a $\mathcal{P}$ group by the hereditary nature of $\mathcal{P}$, and the claim follows.

[8]Proof: $G$ acts on the finite quotient space $G/H$ by left multiplication, hence the *stabiliser* of $G/H$ has finite index in $G$. But this stabliser is also normal in $G$ and contained in $H$.

**Remark 2.4.4.** There is a variant of the above result which is also useful. Suppose one has an $H$-by-$K$ group $G$ in which the action of $K$ on $H$ is virtually trivial (i.e. there are only a finite number of distinct automorphisms of $H$ induces by $K$). Then $G$ is virtually a central $H'$-by-$K'$ group for some finite index subgroups $H', K'$ of $H, K$.

One can phrase various results in group theory in a succinct form using this notation. For instance, one of the basic facts about (discrete) *amenable* groups is that amenable-by-amenable groups are amenable; see [**Ta2010**, §2.8]. As another example, the main result of a well-known paper of Larsen and Pink [**LaPi2011**] is a classification of finite linear groups over a field of characteristic $p$, namely that such groups are virtually ($p$-group by abelian) by (semisimple of Lie type), where one has bounds on the index of the "virtually" and on the type of the semisimple group.

## 2.5. A proof of Gromov's theorem

A celebrated theorem of Gromov [**Gr1981**] reads:

**Theorem 2.5.1** (Gromov's theorem)**.** *Every finitely generated group of polynomial growth is virtually nilpotent.*

The original proof of Gromov's theorem was quite non-elementary, using an infinitary limit and exploiting the work surrounding the solution to *Hilbert's fifth problem*. More recently, Kleiner [**Kl2010**] provided a proof which was more elementary (based in large part on an earlier paper of Colding and Minicozzi [**CoMi1997**]), though still not entirely so, relying in part on (a weak form of the) Tits alternative [**Ti1972**] and also on an ultrafilter argument of Korevaar-Schoen [**KoSc1997**] and Mok [**Mo1995**]. Kleiner's argument is discussed further in [**Ta2009**, §1.2].

Recently, Yehuda Shalom and I [**ShTa2010**] established a quantitative version of Gromov's theorem by making every component of Kleiner's argument finitary. Technically, this provides a fully elementary proof of Gromov's theorem (we do use one infinitary limit to simplify the argument a little bit, but this is not truly necessary); however, because we were trying to quantify as much of the result as possible, the argument became quite lengthy.

In this note I want to record a short version of the argument of Yehuda and myself which is not quantitative, but gives a self-contained and largely elementary proof of Gromov's theorem. The argument is not too far from the Kleiner argument, but incorporates a number of simplifications. In a number of places, there was a choice to take between a short argument that was "inefficient" in the sense that it did not lead to a good quantitative bound, and a lengthier argument which led to better quantitative bounds. I have opted for the former in all such cases.

**2.5.1. Overview of argument.** The argument requires four separate ingredients. The first is the existence of non-trivial Lipschitz harmonic functions $f : G \to \mathbf{R}$:

**Theorem 2.5.2** (Existence of non-trivial Lipschitz harmonic functions)**.** *Let $G$ be an infinite group generated by a finite symmetric set $S$. Then there exists a non-constant function $f : G \to \mathbf{R}$ which is harmonic in the sense that*

$$f(x) = \frac{1}{|S|} \sum_{s \in S} f(xs)$$

*for all $x \in G$, and Lipschitz in the sense that*

$$|f(x) - f(sx)| \leq C$$

*for all $x \in G$ and $s \in S$, and some $C < \infty$.*

The second is that there are not *too* many such harmonic functions:

**Theorem 2.5.3** (Kleiner's theorem)**.** *Let $G$ be a group of polynomial growth generated by a finite symmetric set $S$ of generators. Then the vector space $V$ of Lipschitz harmonic functions is finite-dimensional.*

The third ingredient is that Gromov's theorem is true in the compact linear group case:

**Theorem 2.5.4** (Gromov's theorem in the compact linear case)**.** *Let $G$ be a finitely generated subgroup of a compact linear group $H \subset GL_n(\mathbf{C})$ of polynomial growth. Then $G$ is virtually abelian.*

The final ingredient is that Gromov's theorem is inductively true once one can locate an infinite cyclic quotient:

**Theorem 2.5.5** (Gromov's theorem with an cyclic quotient)**.** *Let $G$ be a finitely generated group which has polynomial growth of exponent at most $d$ (i.e. the volume of a ball $B_S(r)$ grows like $O(r^d)$ for any fixed set of generators $S$). Suppose inductively that Gromov's theorem is already known for groups of polynomial growth of exponent at most $d-1$, and suppose that $G$ contains a finite index subgroup $G'$ which can be mapped homomorphically onto an infinite cyclic group. Then $G$ is virtually nilpotent.*

We prove these four facts in later sections. For now, let us see how they combine to establish Gromov's theorem in full generality.

We assume that $G$ has polynomial growth of order $d$, and assume inductively that Gromov's theorem has already been established for growth of order $d-1$ or less. We fix a symmetric set $S$ of generators.

We may assume that $G$ is infinite otherwise we are already done. So by Theorem 2.5.2, the space $V$ of (complex) Lipschitz harmonic functions

consists of more than just the constants $\mathbf{R}$. In particular, setting $W := V/\mathbf{C}$, we have a non-trivial short exact sequence

$$0 \to \mathbf{C} \to V \to W \to 0.$$

The left translation action of $G$ preserves the space of Lipschitz harmonic functions, and is thus an action of $G$ on $V$. Since $G$ preserves constants, it is also an action of $G$ on $W$. Now, on $W$, the homogeneous Lipschitz norm is a genuine norm, and is preserved by the action of $G$. Since all norms are equivalent on a finite-dimensional space, we can place an arbitrary Euclidean structure on $W$ and conclude that this structure is preserved up to constants by $G$. So, the image of the action of $G$ on $W$ is precompact, and thus its closure is a compact linear group. By Theorem 2.5.4, this image is virtually abelian. If it is infinite, then we thus see that a finite index subgroup of $G$ has an infinite abelian image, and thus has a surjective homomorphism onto the integers, and we are done by Theorem 2.5.5. So we may assume that this image is finite; thus there is a finite index subgroup $G'$ of $G$ that is trivial on $W$. The action of $G'$ on $V$ then collapses to the form $gf = f + \lambda_g(f)$ for some linear functional $\lambda_g \in V^*$ (in fact $\lambda_g$ annihilates 1 and so comes from $W^*$). Note that $\lambda$ is then an additive representation of $G$. If the image of this representation is infinite, then we are again done by Theorem 2.5.5, so we may assume that it is finite; thus there is a finite index subgroup $G''$ of $G'$ that is trivial on $V$. In other words, all Lipschitz harmonic functions are $G''$-invariant, and thus take only finitely many values. But looking at the maximum such value and using harmonicity (i.e. using the *maximum principle*) we conclude that all Lipschitz harmonic functions are constant, a contradiction.

**2.5.2. Building a Lipschitz harmonic function.** Now we prove Theorem 2.5.2. We introduce the function

$$\mu := \frac{1}{|S|} \sum_{s \in S} \delta_s$$

where $\delta_s$ is the Kronecker delta function. The property of a function $f : G \to \mathbf{C}$ being harmonic is then simply that $f * \mu = f$, using the discrete convolution structure on the group.

To build such a function, we consider the functions

$$f_n := \frac{1}{n} \sum_{m=1}^{n} \mu^{(m)}$$

where $\mu^{(m)} := \mu * \ldots * \mu$ is the convolution of $m$ copies of $\mu$. This sequence of functions is "asymptotically harmonic" in the sense that

$$\|f_n\|_{\ell^1(G)} = 1$$

but

$$\|f_n - f_n * \mu\|_{\ell^1(G)} = O(1/n)$$

(we allow implied constants to depend on $S$).

There are now two cases. The first case is the **non-amenable case**, when we have

$$\|f_n - f_n * \delta_s\|_{\ell^1(G)} > \varepsilon > 0$$

for some $s \in S$, some $\varepsilon > 0$, and infinitely many $n$; informally, this means that the averaged iterated convolutions $f_n$ are not getting smoother as $n \to \infty$. By the duality of $\ell^1(G)$ and $\ell^\infty(G)$, we see that for each such $n$ we can find $H_n$ with $\|H_n\|_{\ell^\infty(G)} = 1$ such that

$$|H_n * f_n(\mathrm{id}) - H_n * f_n(s)| > \varepsilon.$$

But Young's inequality, $H_n * f_n$ has $\ell^\infty(G)$ norm of at most 1, and

$$\|H_n * f_n - H_n * f_n * \mu\|_{L^\infty(G)} = O(1/n).$$

Using the sequential Banach-Alaoglu theorem we may take a subsequence limit and obtain a non-trivial bounded harmonic function. Since bounded functions are automatically Lipschitz, and the claim follows.

The second case is the **amenable case**, when we have

$$\|f_n - f_n * \delta_s\|_{\ell^1(G)} \to 0$$

as $n \to \infty$ for each $s \in S$. Setting $F_n := f_n^{1/2}$, one soon verifies that

$$\|F_n\|_{\ell^2(G)} = 1$$

and

$$\|F_n - F_n * \delta_s\|_{\ell^2(G)} = o(1)$$

In particular

$$\|F_n - F_n * \mu\|_{\ell^2(G)} = o(1).$$

From this and the spectral theorem, we see that the positive-definite Laplacian operator $\Delta : \ell^2(G) \to \ell^2(G)$ defined by the formula

$$\Delta F := F - F * \mu$$

has non-trivial spectrum at the origin. On the other hand, as $G$ is infinite, there are no non-trivial harmonic functions in $\ell^2(G)$ (as can be seen from the maximum principle), and so the spectrum at the origin is not coming from a zero eigenfunction. From this and the spectral theorem (taking spectral projections to $[0, \varepsilon]$ for small $\varepsilon$), one can find a sequence $G_n \in \ell^2(G)$ of functions such that

$$\sum_{g \in G} G_n(g) \Delta G_n(g) = 1$$

but

$$\|\Delta G_n\|_{\ell^2(G)} \to 0$$

as $n \to \infty$.

A summation by parts gives the Dirichlet energy identity

$$\sum_{g \in G} G_n(g) \Delta G_n(g) = \frac{1}{2|S|} \sum_{s \in S} \|G_n - G_n * \delta_s\|^2_{\ell^2(G)}$$

and thus

$$\|G_n - G_n * \delta_s\|_{\ell^2(G)} = O(1),$$

and also there exists $s_0 \in S$ such that

$$\|G_n - G_n * \delta_{s_0}\|_{\ell^2(G)} \gg 1$$

for infinitely many $n$. By the self-duality of $\ell^2(G)$, we may thus find a sequence $H_n \in \ell^2(G)$ with $\|H_n\|_{\ell^2(G)} = 1$ such that

$$|H_n * G_n(\mathrm{id}) - H_n * G_n(s_0)| \gg 1$$

for infinitely many $n$. From Young's inequality we also see that

$$\|H_n * G_n - H_n * G_n * \delta_s\|_{\ell^\infty(G)} = O(1)$$

(so $H_n * G_n$ is uniformly Lipschitz) and

$$\|\Delta(H_n * G_n)\|_{\ell^\infty(G)} \to 0$$

as $n \to \infty$, thus $H_n * G_n$ is asymptotically harmonic. Using the Arzelá-Ascoli theorem to take another subsequence limit (after first subtracting a constant to normalise $H_n * G_n$ to be zero at the identity, so that $H_n * G_n$ becomes locally bounded by the uniform Lipschitz property) we obtain the required non-trivial Lipschitz harmonic function.

**Remark 2.5.6.** In the case of groups of polynomial growth, one can verify that one is always in the "amenable" case. In the non-amenable case, the theory of Poisson boundaries gives a plentiful supply of *bounded* Lipschitz harmonic functions (in fact, there is an infinite-dimensional space of such).

**2.5.3. Kleiner's theorem.** We now prove Theorem 2.5.3. Our proof will basically repeat those in Kleiner's original paper [**Kl2010**]. For simplicity, let us assume a stronger condition than polynomial growth, namely *bounded doubling*

$$|B_S(2R)| \leq C|B_S(R)|$$

for some fixed constant $C$ and all $R > 0$. In general, polynomial growth does not obviously imply bounded doubling at all scales, but there is a simple pigeonhole argument that gives bounded doubling on *most* scales, and this turns out to be enough to run the argument below. But in order not to deal with the (minor) technicalities arising from exceptional scales in which bounded doubling fails, I will assume bounded doubling at all scales. The full proof in the general case can, of course, be found in Kleiner's paper

(which in turn was based upon an earlier argument of Colding and Minicozzi [**CoMi1997**]).

Let $\varepsilon > 0$ be a small parameter. The key lemma is

**Lemma 2.5.7** (Elliptic regularity)**.** *Cover $B_S(4R)$ by balls $B$ of radius $\varepsilon R$. Suppose that a harmonic function $f : G \to \mathbf{R}$ has mean zero on every such ball. Then one has*

$$\|f\|_{\ell^2(B_S(R))} \ll \varepsilon \|f\|_{\ell^2(B_S(4R))}.$$

Let's see how this lemma establishes the theorem. Consider some Lipschitz harmonic functions $u_1, \ldots, u_D$, which we normalise to all vanish at the identity. Let $V$ be the space spanned by $u_1, \ldots, u_D$. For each $R$, the $L^2(B_S(R))$ inner product gives a quadratic form $Q_R$ on $V$. Using this quadratic form, we can build a Gram matrix determinant

$$\det(Q_R(u_i, u_j))_{1 \leq i,j \leq D}.$$

From the Lipschitz nature of the harmonic functions, we have a bound of the form

$$\det(Q_R(u_i, u_j))_{1 \leq i,j \leq D} \ll R^D \tag{2.1}$$

as $R \to \infty$. On the other hand, we also have the monotonicity property

$$\det(Q_R(u_i, u_j))_{1 \leq i,j \leq D} \leq \det(Q_{4R}(u_i, u_j))_{1 \leq i,j \leq D}.$$

Now by bounded doubling, we can cover $B_S(4R)$ by $O_\varepsilon(1)$ balls of radius $B$. This creates a codimension $O_\varepsilon(1)$ subspace of $V$ on which $Q_R$ is bounded by $O(\varepsilon)$ times $Q_{4R}$. Using this, we obtain the improved bound

$$\det(Q_R(u_i, u_j))_{1 \leq i,j \leq D} \leq O(\varepsilon)^{D - O_\varepsilon(1)} \det(Q_{2R}(u_i, u_j))_{1 \leq i,j \leq D}.$$

For $\varepsilon$ small enough and $D$ large enough, the rate of growth $O(\varepsilon)^{D-O_\varepsilon(1)}$ is strictly less than $4^{-D}$. Iterating this estimate by doubling $R$ off to infinity, and comparing against (2.1), we conclude in the limit that

$$\det(Q_R(u_i, u_j))_{1 \leq i,j \leq D} = 0$$

for all $R$, and so $u_1, \ldots, u_D$ cannot be linearly independent. This implies that the space of Lipschitz harmonic functions has dimension at most $D+1$, and the claim follows.

It remains to prove the lemma. Fix the harmonic function $f$.

There are two basic ingredients here. The first is the reverse Poincaré inequality[9]

$$\sum_{x \in B_S(2R)} |\nabla f(x)|^2 \ll R^{-2} \sum_{x \in B(x_0, 4R)} |f(x)|^2$$

[9]This inequality is in the general spirit of the philosophy that functions that are harmonic on a ball, should be smooth that ball.

where

$$|\nabla f(x)|^2 := \sum_{s \in S} |f(x) - f(xs)|^2.$$

This claim (which heavily exploits the harmonicity of $f$) is proven by writing $|f|^2$ as $f(f * \mu)$, multiplying by a suitable cutoff function adapted to $B(x_0, 2r)$ and equalling one on $B(x_0, r)$, and summing by parts; we omit the standard details.

The second claim is the Poincaré inequality

$$\sum_{x,y \in B(x_0,r)} |f(x) - f(y)|^2 \ll r^2 |B_S(r)| \sum_{x \in B(x_0,3r)} |\nabla f(x)|^2,$$

which does not require harmonicity. To prove this claim, observe that the left-hand side can be bounded by

$$\sum_{g \in B_S(2r)} \sum_{x \in B(x_0,r)} |f(x) - f(xg)|^2.$$

But by expanding each $g \in B_S(2r)$ as a word of length most $2r$ and using the triangle inequality in $\ell^2$ and Cauchy-Schwarz, we have

$$\sum_{x \in B(x_0,r)} |f(x) - f(xg)|^2 \ll r^2 \sum_{x \in B(x_0,3r)} |\nabla f(x)|^2$$

and the claim follows.

If $f$ has mean zero on $B(x_0, r)$, the Poincaré inequality implies that

$$\sum_{x \in B(x_0,r)} |f(x)|^2 \ll r^2 \sum_{x \in B(x_0,3r)} |\nabla f(x)|^2. \tag{2.2}$$

To prove the lemma, we first use bounded doubling to refine the family of balls $B = B(x_i, \varepsilon R)$ so that the triples $3B = B(x_i, 3\varepsilon R)$ have bounded overlap. Applying (2.2) for each such ball and summing we obtain the claim.

**2.5.4. The compact linear case.** Now we prove Theorem 2.5.4. It is a classical fact that all compact linear groups $H$ are isomorphic to a subgroup[10] of a unitary group $U(n)$; indeed, if one takes the standard inner product on $\mathbf{C}^n$ and averages it by the Haar measure of $H$, one obtains an inner product which is $H$-invariant, and so $H$ can be embedded inside the unitary group associated to this group. Thus it suffices to prove the claim when $H = U(n)$.

A key observation is that if two unitary elements $g, h$ are close to the identity, then their commutator $[g, h] = ghg^{-1}h^{-1}$ is even closer to the

[10]Indeed, thanks to a theorem of Cartan, $H$ is isomorphic to a *Lie* subgroup of $U(n)$, i.e. an analytic submanifold of $U(n)$ that is also a subgroup; but we will not need this fact here.

identity. Indeed, since multiplication on the left or right by unitary elements does not affect the operator norm, we have

$$\|[g,h]-1\|_{op} = \|gh-hg\|_{op}$$
$$= \|(g-1)(h-1)-(h-1)(g-1)\|_{op}$$

and so by the triangle inequality

$$\|[g,h]-1\|_{op} \leq 2\|g-1\|_{op}\|h-1\|_{op}. \tag{2.3}$$

We now need to exploit (2.3) to prove Theorem 2.5.4. As a warm-up, we first prove the following slightly easier classical result:

**Theorem 2.5.8** (Jordan's theorem)**.** *Let $G$ be a finite subgroup of $U(n)$. Then $G$ contains an abelian subgroup of index $O_n(1)$ (i.e. at most $C_n$, where $C_n$ depends only on $n$).*

And indeed, the proof of the two results are very similar. Let us first prove Jordan's theorem. We do this by induction on $n$, the case $n=1$ being trivial. Suppose first that $G$ contains a *central element* $g$ (i.e. an element that commutes with all elements of $G$) which is not a multiple of the identity. Then, by definition, $G$ is contained in the *centraliser* $Z(g) := \{a \in U(n) : ag = ga\}$ of $g$, which by the spectral theorem is isomorphic to a product $U(n_1) \times \ldots \times U(n_k)$ of smaller unitary groups. Projecting $G$ to each of these factor groups and applying the induction hypothesis, we obtain the claim.

Thus we may assume that $G$ contains no central elements other than multiples of the identity. Now pick a small $\varepsilon > 0$ (one could take $\varepsilon = 1/10$ in fact) and consider the subgroup $G'$ of $G$ generated by those elements of $G$ that are within $\varepsilon$ of the identity (in the operator norm). By considering a maximal $\varepsilon$-net of $G$ we see that $G'$ has index at most $O_{n,\varepsilon}(1)$ in $G$. By arguing as before, we may assume that $G'$ has no central elements other than multiples of the identity.

If $G'$ consists only of multiples of the identity, then we are done. If not, take an element $g$ of $G'$ that is not a multiple of the identity, and which is as close as possible to the identity (here is where we use that $G$ is finite). By (2.3), we see that if $\varepsilon$ is sufficiently small depending on $n$, and if $h$ is one of the generators of $G'$, then $[g,h]$ lies in $G'$ and is closer to the identity than $g$, and is thus a multiple of the identity. On the other hand, $[g,h]$ has determinant 1. Given that it is so close to the identity, it must therefore be the identity (if $\varepsilon$ is small enough). In other words, $g$ is central in $G'$, and is thus a multiple of the identity. But this contradicts the hypothesis that there are no central elements other than multiples of the identity, and we are done.

The proof of Theorem 2.5.4 is analogous. Again, we pick a small $\varepsilon > 0$, and define $G'$ as before. If $G'$ has a central element that is not a multiple of

the identity, then we can again argue via induction, so suppose that there are no such elements.

Being finitely generated, it is not difficult to show that $G'$ can be generated by a finite set $S$ of generators within distance $\varepsilon$ of the identity. Now pick an element $h_1 \in S$ which is not a multiple of the identity, and is at a distance $\delta_1$ from the identity for some $0 < \delta_1 \leq \varepsilon$. We look at all the commutators $[g, h_1]$ where $g \in S$. By (2.3), they are all at distance $O(\varepsilon\delta_1)$ from the identity, and have determinant 1. If they are all constant multiples of the identity, then by arguing as before we see that $h_1$ is central in $G'$, a contradiction, so we can find an element $h_2 := [g_1, h_1]$ for some $g_1 \in S$ which is a distance $\delta_2 = O_n(\varepsilon\delta_1)$ from the origin and is not a multiple of the identity. Continuing this, we can construct $h_3 := [g_2, h_2]$, etc., where each $h_n$ is a distance $0 < \delta_n = O(\varepsilon\delta_{n-1})$ from the identity, and is a commutator of $h_{n-1}$ with a generator.

Because of the lacunary nature of the distances of $h_1, h_2, h_3, \ldots$, we easily see that the words $h_1^{i_1} \ldots h_m^{i_m}$ with $0 \leq i_1, \ldots, i_m \leq c\varepsilon^{-1}$ are distinct for some small $c > 0$. On the other hand, all of these words lie in the ball of radius $O(m\varepsilon^{-1}2^m)$ generated by $S$. This contradicts the polynomial growth hypothesis for $\varepsilon$ taken small enough and $m$ large enough.

**Remark 2.5.9.** Theorem 2.5.4 can be deduced as a corollary of Gromov's theorem, though we do not do so here as this would be circular. Indeed, it is not hard to see that the image of a torsion-free nilpotent group in a unitary group must be abelian.

**2.5.5. The case of an infinite abelian quotient.** Now we prove Theorem 2.5.5 (which was already observed in Gromov's original paper [**Gr1981**], and also closely related to earlier work of Milnor [**Mi1968**] and of Wolf [**Wo1968**]).

Since $G$ is finitely generated and has polynomial growth of order $d$, the finite index subgroup $G'$ is also finitely generated of growth $d$. By hypothesis, there is a non-trivial homomorphism $\phi : G' \to \mathbf{Z}$. Using the Euclidean algorithm, one can move the generators $e_1, \ldots, e_m$ of $G'$ around so that all but one of them, say $e_1, \ldots, e_{m-1}$, lie in the kernel $\ker(\phi)$ of $\phi$; we thus see that this kernel must then be generated by $e_1, \ldots, e_{m-1}$ and their conjugates $e_m^k e_i e_m^{-k}$ by powers of $e_m$.

Let $S_k$ be the set of $e_m^{k'} e_i e_m^{-k'}$ for $1 \leq i \leq m-1$ and $|k'| \leq k$, and let $B_k$ be the words of length at most $k$ generated by elements of $S_k$. Observe that if at least the elements in $S_{k+1}$ is not contained in $B_k \cdot B_k^{-1}$, then $B_{k+1}$ is at least twice as big as $B_k$. Because of polynomial growth, this implies that $S_{k+1} \subset B_k \cdot B_k^{-1}$ for some $k \geq 1$, which implies that $\ker(\phi)$ is generated by $S_k$.

Observe that the ball of radius $R$ generated by $S_k$ is at least $R/2$ times as large as the ball of radius $R/2$ generated by $e_1, \ldots, e_{m-1}$. Since $G'$ has growth $d$, we conclude that $\ker(\phi)$ has growth at most $d-1$, and is thus virtually nilpotent by hypothesis.

We have just seen that the kernel $\ker(\phi)$ contains a nilpotent subgroup $N$ of some finite index $M$; it is thus finitely generated. From Lagrange's theorem, we see that the group $N'$ generated by the powers $g^M$ with $g \in \ker(\phi)$ is then contained in $N$ and is therefore nilpotent. $N'$ is clearly a *characteristic subgroup* of $\ker(\phi)$ (i.e. preserved under all outer automorphisms), and is thus normal in $N$. The group $N/N'$ is nilpotent and finitely generated with every element being of order $M$, and is thus finite; thus $N'$ is finite index in $\ker(\phi)$. Since it is characteristic, it is in particular invariant under conjugation by $e_m$. If one lets $G'' = \mathbf{Z} \ltimes_{e_m} N'$ be the group generated by $N'$ and $e_m$, we see that $G''$ is a finite index subgroup[11] of $G$. In particular, it has polynomial growth.

To conclude, we need to show that $G''$ is virtually nilpotent. It will suffice to show that the conjugation action of $e_m^a$ on $N'$ acts unipotently on $N'$ for some finite $a > 0$. We can induct on the step of the nilpotent group $N'$, assuming that the claim has already been proven for the quotient group $N'/Z(N')$ (where $Z(N')$ is the centre of $N'$), which has one lower step on $N'$. Thus it suffices to prove unipotence on just the center $Z(N')$, which is a finitely generated abelian group and thus isomorphic to some $\mathbf{Z}^d \times H$ for some finite group $H$. The torsion group $H$ must be preserved by this action. By *Lagrange's theorem*, the action on $H$ becomes trivial after raising $e_m$ to a suitable power, so we only need to consider the action on $\mathbf{Z}^d$. In this case the conjugation action can be viewed as a matrix $A$ in $SL_d(\mathbf{Z})$. Because $G''$ has polynomial growth, the powers $A^n$ of $A$ for $n \in \mathbf{Z}$ can only grow polynomially; in other words, all the eigenvalues of $A$ have unit magnitude. On the other hand, these eigenvalues consist of Galois conjugacy classes of algebraic integers. But it is a classical result of Kronecker that the only algebraic integers $\alpha$ whose Galois conjugacy classes all have unit magnitude are the roots of unity[12]. We conclude that all the eigenvalues of $A$ are roots of unity, i.e. some power of $A$ is unipotent, and the claim follows.

[11] Note that as $e_m$ is not annihilated by $\phi$, it will have infinite torsion even after quotienting out by $N'$.

[12] Proof: the action of the $\alpha^n$ on the ring $\mathbf{Z}[\alpha]$ are uniformly bounded in $n$ and must thus repeat itself due to the finite-dimensional nature of $\mathbf{Z}[\alpha]$.

*Chapter 3*

# Analysis

## 3.1. Orders of magnitude, and tropical geometry

In analysis, it is often the case that one does not need to know the exact value of the numerical quantities that one is manipulating, but only their order of magnitude. For instance, if one knows that one number $A$ is on the order of $10^6$, and another number $B$ is on the order of $10^3$, then this should be enough (given a sufficiently precise quantification of "on the order of") to ensure that $B$ is significantly smaller than $A$, and can thus be viewed as a "lower order term".

Orders of magnitude can be made more precise by working asymptotically (in which there is a parameter $n$ going off to infinity) or via nonstandard analysis (in which there is a parameter $n$ that is an unbounded number). For instance, if $A$ is comparable to $n^6$, and B is comparable to $n^3$ then for sufficiently large $n$, $B$ will be smaller than $A$, and $A + B$ and $A$ will be asymptotically equal as $n$ goes to infinity (in the sense that the ratio between $A + B$ and $A$ goes to 1 as $n \to \infty$). In particular, $A + B$ will also be comparable to $n^6$.

One reason for working with orders of magnitude is that it has a simpler arithmetic than the arithmetic of numbers. For instance, if $A$ is a positive quantity comparable to $n^a$, and $B$ is a positive quantity comparable to $n^b$, then A+B is comparable to $n^{\max(a,b)}$, and $AB$ is comparable to $n^{a+b}$, and $A/B$ is comparable to $n^{a-b}$. Thus we see that by passing from numbers to orders of magnitude, the addition operation has been transformed into the simpler max operation, while the multiplication and division operations have been transformed into addition and subtraction operations. To put it another way, the map from numbers (or more precisely, positive numbers

depending in an approximately polynomial fashion on $n$) to orders of magnitude is a homomorphism from the former *semiring* to the latter, where we give the orders of magnitude the *tropical semiring* structure given by the *max-plus algebra*.

To phrase this equivalently in the context of nonstandard analysis: if $n$ is an unbounded positive nonstandard number, then the map $x \mapsto \mathrm{st}(\log_n x)$ is a semiring homomorphism from the semiring $n^{O(1)}$ of positive numbers of polynomial size, to the real numbers with the max-plus algebra.

If one does not work asymptotically (or with nonstandard analysis), but works with finite orders of magnitude, then the relationship between ordinary arithmetic and tropical arithmetic is only approximate rather than exact. For instance, if $A = 10^6$ and $B = 10^3$, then $A + B \approx 10^{\max(6,3)}$. If we replace the base 10 with a larger base, then the error in the exponent here goes to zero as the base goes to infinity, and we recover the asymptotic homomorphism.

This illustrates that tropical arithmetic is a degenerate limit of ordinary arithmetic, which explains why so many algebraic and geometric facts in ordinary arithmetic and geometry have analogues in tropical arithmetic and tropical geometry. In particular, the analogue of an algberaic variety is the *spine* of the associated *amoebae*. For instance, consider the plane

$$\{(x, y, z) \in \mathbf{R} : x + y + z = 0\}$$

in classical geometry. Then, by the triangle inequality, the largest of the three magnitudes $|x|, |y|, |z|$ cannot exceed twice the second largest of the magnitudes. In particular, if $|x|, |y|, |z|$ are comparable with $n^a, n^b, n^c$ respectively, then the largest value of $a$, $b$, $c$ must equal the second largest value; or equivalently (in the max-plus algebra), one has

$$\max(a, b) = \max(b, c) = \max(c, a).$$

Geometrically, this asserts that $(a, b, c)$ lies in a certain two-dimensional $Y$-shaped object (three half-planes glued along a common axis). The decomposition of the tropical analogue of the plane $x + y + z = 0$ into these three half-planes is an important decomposition in analysis; for instance, in Littlewood-Paley theory, it is known as the *Littlewood-Paley trichotomy*, dividing all frequency interactions into high-low, low-high, and high-high frequency interactions.

One can also use the relationship between tropical geometry and classical geometry in the reverse direction, viewing various concepts from combinatorial optimisation as degenerations of ones from algebraic geometry (or conversely, viewing the latter as relaxations of the former). For instance, a

metric on $n$ points can be viewed as[1] the tropical analogue of an idempotent real symmetric $n \times n$ matrix, because the triangle inequality for metrics is the tropical analogue of the idempotency relation $P^2 = P$, after identifying a metric with its adjacency matrix.

## 3.2. Descriptive set theory vs. Lebesgue set theory

The set of reals is uncountable, and the set of all subsets of the real line is even more uncountable, with plenty of room for all sorts of pathological sets to lurk. There are basically two major ways we use to handle this vast wilderness. One is to start classifying sets by how "nice" they are (e.g. open, closed, $G_\delta$, Borel, analytic, Baire class, etc. ); this leads to the subject of *descriptive set theory*, which is a subtle subject that can be sensitive to the choice of axioms of set theory that one wishes to use. The other approach, which one might dub the "Lebesgue" approach, is to

- restrict attention to Lebesgue measurable sets; and
- ignore anything that happens on a set of measure zero.

This latter approach is very suitable for applications to analysis (in many applications, particularly those involving integration or averaging, we are willing to lose control on a small set, and particularly on sets of measure zero), and vastly cuts down the complexity of the sets one has to deal with; every Lebesgue measurable set is (locally) an elementary set (a finite union of intervals) outside of a set of arbitrarily small measure, and is hence (locally) a pointwise limit of elementary sets outside of a set of zero measure. As such, most of the delicate hierarchies of classes of sets in descriptive set theory are not needed in the Lebesgue world.

In descriptive set theory, the concept of (everywhere) pointwise convergence is of basic importance. In the Lebesgue worldview, the concept of *almost everywhere* pointwise convergence takes its place. The two look very similar, but the former is much stronger than the latter in some ways. Consider for instance the following two classical results:

**Lemma 3.2.1.** *Let $f$ be an everywhere pointwise limit of continuous functions $f_n$ on $\mathbf{R}$ (i.e. a* Baire class 1 *function). Then the set of points of discontinuity of $f$ is of the first category (i.e. the countable union of nowhere dense sets). In particular (by the* Baire category theorem*), the set of points where $f$ is continuous must be dense.*

**Lemma 3.2.2.** *Every Lebesgue measurable function that is finite a.e., is the almost everywhere limit of continuous functions on $\mathbf{R}$. In particular there*

[1] I learned this example from Berned Sturmfels; see [**HeJeKoSt2009**] for further discussion.

*exist nowhere continuous functions that are the a.e. limit of continuous functions.*

To prove Lemma 3.2.1, it suffices to prove that for every (rational) $a < b$, the set of points $x$ where

$$\liminf_{y \to x} f(y) < a < b < \limsup_{y \to x} f(y)$$

is nowhere dense. Suppose for contradiction that this set was dense in an interval $I$; then $f$ exceeds $b$ on a dense subset of $I$ and dips below a on a dense subset of $I$ also. But then the sets where $f(x) = \limsup_{n\to\infty} f_n(x) > b$ and $f(x) = \liminf_{n\to\infty} f_n(x) < a$ are the countable intersection of open dense subsets in $I$, and thus must have a point in common by the Baire category theorem, a contradiction.

Lemma 3.2.2 can be proven by a variety of truncation, regularisation, and approximation methods (e.g. *Lusin's theorem* will work).

Note how Lemma 3.2.1 is powered by the Baire category theorem, which is a basic tool in descriptive set theory. But because sets of the first category can have full measure (since open dense sets can have arbitrarily small measure), the Baire category theorem becomes useless once one is allowed to ignore sets of measure zero, which is why Lemma 3.2.1 fails so dramatically in the Lebesgue world.

The strength of everywhere pointwise convergence, as compared against almost everywhere pointwise convergence, can also be seen by noting that there are a variety of useful and general tools by which one can establish almost everywhere pointwise convergence (e.g. by converging sufficiently fast in measure or in an $L^p$ sense), but very few ways to establish *everywhere* pointwise convergence without also giving the significantly stronger property of (local) uniform convergence.

## 3.3. Complex analysis vs. real analysis

The various fields of analysis in mathematics typically take place in a domain over some field or algebra, which is either one-dimensional or higher dimensional (though one occasionally also sees fractional dimensional domains). Thus for instance we have real analysis in one dimension or in higher dimensions, complex analysis in one variable and several complex variables, quaternionic or Clifford analysis in one variable or in several variables, and so forth[2].

[2]In theoretical computer science, one also sees plenty of analysis over finite fields such as $\mathbf{F}_2$, e.g. using *hypercontractivity*. Analysis over the $p$-adics $\mathbf{Z}_p$ or adeles $\mathbf{A}$ is also of some use in number theory. One now also increasingly sees analysis on objects such as graphs, in which there is no underlying algebra at all, although paths in a graph can be sort of viewed as curves over $\mathbf{Z}$ to some extent.

Higher dimensional analysis is more difficult than one-dimensional analysis for many reasons, but one main one is that one-dimensional domains tend to be 1"flat" or otherwise have an uninteresting geometry (at least when there are no topological obstructions). For instance, a one-dimensional simple connected real curve is always (intrinsically) isometric to an interval; and a one-dimensional simply connected complex domain is always conformal to a disk (Riemann mapping theorem). Another key advantage of the one-dimensional setting is that power series are indexed by a single parameter rather than by multiple parameters, making it easier to manipulate them by peeling off one monomial at a time.

These are some of the reasons why complex analysis in one variable is significantly more powerful than real analysis in two variables, despite the fact that the complex line $\mathbf{C}$ has the structure of the real plane $\mathbf{R}^2$ after one forgets the complex structure. But another crucial reason for this is that in the complex domain, there exist non-trivial closed contours, whereas in the real domain, all closed contours are degenerate. Thus the *fundamental theorem of calculus* in real analysis gets augmented to the significantly more powerful *Cauchy theorem* in complex analysis, which is the basis for all the power of contour integration methods. By exploiting the additional dimension available in the complex setting, one can avoid being blocked by singularities or other obstructions, for instance by shifting a contour to go *around* a singularity rather than *through* it.

Another example of this comes from spectral theory. Suppose for instance that one wants to understand the distribution of the (real) eigenvalues $\lambda_1 < \ldots < \lambda_n$ of an $n \times n$ Hermitian matrix $A$. A popular real variable approach is the *moment method*, which proceeds by computing the moments

$$\mathrm{tr}(A^k) = \sum_i \lambda_i^k$$

for $k = 1, 2, 3, \ldots$. In principle, these moments determine the spectrum; and they are particularly useful for detecting the *edges* of the spectrum (the most extreme eigenvalues $\lambda_1$ and $\lambda_n$), as this is where the function $x^k$ is largest. However, it takes a nontrivial amount of effort[3] to use the moments to control the *bulk* of the spectrum (one has to combine many moments together in order to create a polynomial that is localised in whatever portion of the bulk one wants to inspect).

[3]Indeed, some of the most powerful ways to solve the inverse moment problem proceed via complex-variable methods.

On the other hand, one can proceed by the complex-variable method of the *Stieltjes transform*

$$\mathrm{tr}((A - zI)^{-1}) = \sum_i \frac{1}{\lambda_i - z},$$

where $z$ is a complex number. This transform is well-defined for all $z$ outside of the spectrum of $A$, and in particular for all complex $z$ with non-zero imaginary parts. To understand the distribution of the spectrum near some value $x$, it suffices to control the Stieltjes transform for complex numbers near $x$, such as $x + i\varepsilon$. The point is that even if $x$ is deep in the bulk of the spectrum and thus not easily accessed by real-variable techniques, one can go *around* the spectrum and get arbitrarily close to $x$ by complex-variable techniques.

A comparison of the real-variable moment method and the complex-variable Stieltjes transform method in the case of establishing the semicircular law for Wigner random matrices is given in [**Ta2011c**].

In view of all this, it is natural to then revisit, say, real analysis in four variables, and try to recast it as quarternionic analysis in one variable. While this can be done to some extent (and is part of the field of *Clifford analysis*), there is however a significant obstruction to transferring the most powerful components of complex analysis to non-commutative algebras: whereas the product of two differentiable functions over a commutative algebra remains differentiable, the product of two differentiable functions over a non-commutative algebra is usually not differentiable over that algebra. So while the space of real differentiable or complex differentiable functions form a commutative algebra, the quarternionic differentiable functions do not form either a commutative algebra or a non-commutative algebra, but simply fail to be an algebra altogether.

In real and complex analysis, the algebra property (when combined with tools such as *Neumann series* or the *Weierstrass approximation* theorem, together with the ability to take limits) leads to many useful ways to manipulate differentiable functions, and in particular to invert them when they are non-zero. The loss of the algebra property when the underlying algebra is non-commutative is thus a serious setback, that has rendered the advantages of Clifford analysis over real analysis in several variables somewhat modest in nature.

## 3.4. Sharp inequalities

Much of "hard" analysis relies extensively on inequalities that control the magnitude of one expression $X$ by another $Y$. One can loosely divide analysis into various regimes, depending on how precise one wants this control to be. In descending order of precision, one has:

(1) **Sharp inequalities**. These are exact inequalities such as $X \leq Y$ without any unspecified constants or error terms (or $X \leq CY$ with an extremely explicit $C$, typically involving special constants such as $\pi$ or $e$). Typical tools used to establish sharp inequalities include convexity, calculus of variations (or just plain old calculus), rearrangement, gradient flows, monotonicity formulae and induction. On the other hand, one cannot lose any constant factors or error terms in one's arguments, no matter how small, unless they are somehow compensated for by a larger factor with a beneficial sign. A proof of a sharp inequality $X \leq Y$ typically also comes automatically with a good description of the cases in which equality $X = Y$ occurs.

(2) **Almost sharp inequalities**. These are inequalities such as $X \leq (1 + O(\varepsilon))Y$, where $\varepsilon$ is small in some sense. So one cannot afford to lose a constant factor in the main term $Y$ on the right-hand side, but can accept these sorts of losses in the error term. Typical tools used here include Taylor expansion in $\varepsilon$, subtracting off the main term, and various "defect" or "stability" versions of sharp inequalities.

(3) **Quasi-inequalities**. These are inequalities such as $X \leq CY$, where $C$ is an unspecified constant that varies from line to line. One is now prepared to lose such constants in arguments, which opens up tools such as the triangle inequality (or more generally, divide-and-conquer strategies), equivalences of function space norms, use of bump functions and mollifiers, and asymptotic notation (such as big-$O$ notation).

(4) **Quasi-inequalities with logarithmic losses**. These are estimates such as $X \leq C(\log^C n)Y$, or perhaps $X \leq n^{o(1)}Y$, where $n$ is some basic entropy parameter (e.g. the ratio between the largest scale and the smallest scale in the problem, raised to the power of the ambient dimension); thus one can accept logarithmic losses in the entropy, or polynomial losses in the dimension. Such losses are often acceptable when one is ultimately only interested in dimensional or exponent information, or if one expects to eventually gain a power saving $n^{-c}$ that will overcome all logarithmic losses.

Important techniques that become available in this setting include dyadic pigeonholing, chaining, and concentration of measure.

(5) **Quasi-inequalities with polynomial losses**. These are estimates such as $X \leq Cn^C Y$ that are polynomial in the entropy (or exponential in the dimension). Such estimates tend to be very easy to obtain by crude methods, but are sometimes needed as an initial inequality that one will subsequently amplify to a stronger estimate with a reduced loss.

(6) **Coarse or qualitative inequalities**. These are estimates such as $X \leq C_n Y$, or even $X \leq C(n, Y)$, where the dependence of $C$ on $n$ or $Y$ is unspecified[4]. Basically, one may as well work with a fixed n here and not worry any further about dependence of constants on $n$. One key tool that can be used in this regime is the use of various compactness arguments or correspondence principles to convert this "hard analysis" problem into a "soft analysis" one in an infinitary setting, so that some sophisticated machinery from infinitary analysis (e.g. Lie group theory, ergodic theory, measure theory, etc.) can be applied.

As indicated above, most techniques in hard analysis to prove inequalities tend to sit most naturally in just one of the above categories, and are more difficult to apply outside of that category[5]. But there are also some powerful tricks available to move *between* categories. In one direction, it is sometimes possible to prove a coarser inequality (such as a quasi-inequality) by guessing what a more precise version of that inequality (such as a sharp inequality) would be, and proving that more precise version using methods adapted to that category (e.g. induction). Such a proof might not be immediately obvious in the original category. In the other direction, one can sometimes prove a precise inequality by first proving an ostensibly weaker inequality in a coarser category, and then amplifying it by some trick (e.g. a *tensor power trick*, see [**Ta2008**, §1.9]).

Coarser inequalities, being logically weaker than more precise inequalities, are generally easier to prove and to generalise. So progress in hard analysis often proceeds in stages, in which coarse inequalities are established first for various problems of interest in an area, which then suggest the way forward for a subsequent wave of more precise inequalities.

[4]This is also the regime for classical "for every $\varepsilon$ there exists a $\delta$" type analysis.

[5]For instance, *Calderon-Zygmund theory*, a key foundation of harmonic analysis, works best in the quasi-inequality category, but becomes mostly trivial (and thus useless) once one is willing to tolerate logarithmic losses.

## 3.5. Implied constants and asymptotic notation

A basic but incredibly useful notational convention in analysis, which I believe was first promoted by Hardy, is to use a symbol (usually $C$) to denote an unspecified constant whose value varies from one line to the next.

But having a symbol for a constant whose exact value one does not particularly care about is also useful in algebra as well. For instance, I was recently performing a lengthy algebraic computation to determine exactly a certain explicit probability distribution. During this computation, various normalisation constants kept emerging. But it was not strictly necessary to keep track of their exact values; instead, one could call them all "$C$", and at the end of the day, one could recover the final combination of all the normalisation constants by using the fact that probability distributions have total mass one.

In a similar spirit, one can often use "$C$" to denote the exponents of various normalising factors, if one is able to recover these exponents at the end of the day by such tools as dimensional analysis (or testing against key examples).

On the other hand, working out the constants $C$ exactly (and making sure they match what one can obtain *a posteriori* via the above shortcuts) is sometimes a beneficial "checksum" to help catch any errors in one's argument.

The standard algebraic trick of passing to a quotient space (e.g. passing to a projective space) can be viewed as a way to systematically ignore things such as normalisation constants.

The use of the *big-O notation* $O()$ to hide the implicit constants $C$ is very useful in analysis. One does have to be a little careful, though, when combining the big-$O$ notation with an iteration argument, such as an induction argument. Indeed, consider the following "proof" that all natural numbers are bounded, i.e. $n = O(1)$ for all $n$: clearly $0 = O(1)$, and if $n = O(1)$ then $n + 1 = O(1)$ as well, and so the claim follows by induction.

The problem here is that the statement "$n = O(1)$" is not really a single statement, but is one of a family of statements "$|n| \leq C$", and one cannot induct on a family of statements, only on a single statement.

But it is safe to combine induction with $O()$ notation so long as one makes the constants in the $O()$ notation depend on the iteration stage, as one can then move the quantification over the implied constant $C$ inside the iteration, thus avoiding the previous problem. For instance, the previous induction gives a perfectly valid proof that $n = O_n(1)$, i.e. $n$ is bounded in magnitude by a constant $C_n$ depending on $n$. In this case, of course, the conclusion is not particularly strong, but there are other situations in

which one does not mind such dependencies on the implied constants. For instance, in analysis one is often willing to lose constants that depend on the dimension $d$, which allows one to use induction-on-dimension arguments using $O()$ (or more precisely, $O_d()$) notation.

## 3.6. Brownian snowflakes

(A side of) the *Koch snowflake* is a famous example of a *self-similar fractal* - a non-smooth curve, consisting of parts which are similar to the whole. It can be constructed recursively, by starting with a line segment and continually introducing "kinks".

*Brownian motion* (in, say, the plane) is another fractal, but it does not look self-similar in the way the Koch snowflake does. Nevertheless, one can think of every Brownian motion trajectory $(B_t)_{t>0}$ as a (normalised random gaussian) *projection* of a self-similar fractal $(c_t)_{t>0}$ in an infinite dimensional Hilbert space, defined by requiring the increments $c_t - c_s$ to have length $|t-s|^{1/2}$ for every $t > s$ and to be orthogonal when the intervals $[s,t]$ are disjoint. This fractal can be constructed by a recursive construction extremely similar to that of the snowflake[6].

So, Brownian motions can be thought of as the shadows of an infinite-dimensional snowflake.

## 3.7. The Euler-Maclaurin formula, Bernoulli numbers, the zeta function, and real-variable analytic continuation

The *Riemann zeta function* $\zeta(s)$ is defined in the region $\mathrm{Re}(s) > 1$ by the absolutely convergent series

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s} = 1 + \frac{1}{2^s} + \frac{1}{3^s} + \ldots. \tag{3.1}$$

Thus, for instance, it is known that $\zeta(2) = \pi^2/6$, and so

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = 1 + \frac{1}{4} + \frac{1}{9} + \ldots = \frac{\pi^2}{6}. \tag{3.2}$$

For $\mathrm{Re}(s) \leq 1$, the series on the right-hand side of (3.1) is no longer absolutely convergent, or even conditionally convergent. Nevertheless, the $\zeta$ function can be extended to this region (with a pole at $s = 1$) by *analytic*

[6]It can also be constructed concretely as the indicator functions $c_t = 1_{[0,t]}$ in $L^2(\mathbf{R})$, but this makes the curve look deceptively linear.

*continuation*. For instance, it can be shown that after analytic continuation, one has $\zeta(0) = -1/2$, $\zeta(-1) = -1/12$, and $\zeta(-2) = 0$, and more generally

$$\zeta(-s) = -\frac{B_{s+1}}{s+1} \tag{3.3}$$

for $s = 1, 2, \ldots$, where $B_n$ are the *Bernoulli numbers*. If one *formally* applies (3.1) at these values of $s$, one obtains the somewhat bizarre formulae

$$\sum_{n=1}^{\infty} 1 = 1 + 1 + 1 + \ldots$$
$$= -1/2; \tag{3.4}$$

$$\sum_{n=1}^{\infty} n = 1 + 2 + 3 + \ldots$$
$$= -1/12; \tag{3.5}$$

$$\sum_{n=1}^{\infty} n^2 = 1 + 4 + 9 + \ldots$$
$$= 0; \tag{3.6}$$

and more generally

$$\sum_{n=1}^{\infty} n^s = 1 + 2^s + 3^s + \ldots = -\frac{B_{s+1}}{s+1}. \tag{3.7}$$

Clearly, these formulae do not make sense if one stays within the traditional way to evaluate infinite series, and so it seems that one is forced to use the somewhat unintuitive analytic continuation interpretation of such sums to make these formulae rigorous. But as it stands, the formulae look "wrong" for several reasons. Most obviously, the summands on the left are all positive, but the right-hand sides can be zero or negative. A little more subtly, the identities do not appear to be consistent with each other. For instance, if one adds (3.4) to (3.5), one obtains

$$\sum_{n=1}^{\infty} (n+1) = 2 + 3 + 4 + \ldots = -7/12 \tag{3.8}$$

whereas if one subtracts 1 from (3.5) one obtains instead

$$\sum_{n=2}^{\infty} n = 0 + 2 + 3 + 4 + \ldots = -13/12 \tag{3.9}$$

which looks inconsistent with (3.8).

However, it is possible to interpret (3.4), (3.5), (3.6) by purely real-variable methods, without recourse to complex analysis methods such as analytic continuation, thus giving an "elementary" interpretation of these

sums that only requires undergraduate calculus; we will later also explain how this interpretation deals with the apparent inconsistencies pointed out above.

To see this, let us first consider a convergent sum such as (3.2). The classical interpretation of this formula is the assertion that the partial sums

$$\sum_{n=1}^{N} \frac{1}{n^2} = 1 + \frac{1}{4} + \frac{1}{9} + \ldots + \frac{1}{N^2}$$

converge to $\pi^2/6$ as $N \to \infty$, or in other words that

$$\sum_{n=1}^{N} \frac{1}{n^2} = \frac{\pi^2}{6} + o(1)$$

where $o(1)$ denotes a quantity that goes to zero as $N \to \infty$. Actually, by using the *integral test* estimate

$$\sum_{n=N+1}^{\infty} \frac{1}{n^2} \leq \int_N^\infty \frac{dx}{x^2} = \frac{1}{N}$$

we have the sharper result

$$\sum_{n=1}^{N} \frac{1}{n^2} = \frac{\pi^2}{6} + O(\frac{1}{N}).$$

Thus we can view $\frac{\pi^2}{6}$ as the leading coefficient of the asymptotic expansion of the partial sums of $\sum_{n=1}^{\infty} 1/n^2$.

One can then try to inspect the partial sums of the expressions in (3.4), (3.5), (3.6), but the coefficients bear no obvious relationship to the right-hand sides:

$$\sum_{n=1}^{N} 1 = N$$

$$\sum_{n=1}^{N} n = \frac{1}{2}N^2 + \frac{1}{2}N$$

$$\sum_{n=1}^{N} n^2 = \frac{1}{3}N^3 + \frac{1}{2}N^2 + \frac{1}{6}N.$$

For (3.7), the classical *Faulhaber formula* (or *Bernoulli formula*) gives

$$\begin{aligned} \sum_{n=1}^{N} n^s &= \frac{1}{s+1} \sum_{j=0}^{s} \binom{s+1}{j} B_j N^{s+1-j} \\ &= \frac{1}{s+1} N^{s+1} + \frac{1}{2} N^s + \frac{s}{12} N^{s-1} + \ldots + B_s N \end{aligned} \tag{3.10}$$

for $s \geq 2$, which has a vague resemblance to (3.7), but again the connection is not particularly clear.

The problem here is the discrete nature of the partial sum

$$\sum_{n=1}^{N} n^s = \sum_{n \leq N} n^s,$$

which (if $N$ is viewed as a real number) has jump discontinuities at each positive integer value of $N$. These discontinuities yield various artefacts when trying to approximate this sum by a polynomial in $N$. (These artefacts also occur in (3.2), but happen in that case to be obscured in the error term $O(1/N)$; but for the divergent sums (3.4), (3.5), (3.6), (3.7), they are large enough to cause real trouble.)

However, these issues can be resolved by replacing the abruptly truncated partial sums $\sum_{n=1}^{N} n^s$ with *smoothed sums* $\sum_{n=1}^{\infty} \eta(\frac{n}{N}) n^s$, where $\eta : \mathbf{R}^+ \to \mathbf{R}$ is a *cutoff function*, or more precisely a compactly supported bounded function that equals 1 at 0. The case when $\eta$ is the indicator function $1_{[0,1]}$ then corresponds to the traditional partial sums, with all the attendant discretisation artefacts; but if one chooses a smoother cutoff, then these artefacts begin to disappear (or at least become lower order), and the true asymptotic expansion becomes more manifest.

Note that smoothing does not affect the asymptotic value of sums that were already absolutely convergent, thanks to the *dominated convergence theorem*. For instance, we have

$$\sum_{n=1}^{\infty} \eta(\frac{n}{N}) \frac{1}{n^2} = \frac{\pi^2}{6} + o(1)$$

whenever $\eta$ is a cutoff function (since $\eta(\frac{n}{N}) \to 1$ pointwise as $N \to \infty$ and is uniformly bounded). If $\eta$ is equal to 1 on a neighbourhood of the origin, then the integral test argument then recovers the $O(1/N)$ decay rate:

$$\sum_{n=1}^{\infty} \eta(\frac{n}{N}) \frac{1}{n^2} = \frac{\pi^2}{6} + O(\frac{1}{N}).$$

However, smoothing can greatly improve the convergence properties of a divergent sum. The simplest example is *Grandi's series*

$$\sum_{n=1}^{\infty} (-1)^{n-1} = 1 - 1 + 1 - \ldots.$$

The partial sums

$$\sum_{n=1}^{N} (-1)^{n-1} = \frac{1}{2} + \frac{1}{2}(-1)^{N-1}$$

oscillate between 1 and 0, and so this series is not conditionally convergent (and certainly not absolutely convergent). However, if one performs analytic continuation on the series

$$\sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n^s} = 1 - \frac{1}{2^s} + \frac{1}{3^s} - \dots$$

and sets $s = 0$, one obtains a formal value of $1/2$ for this series. This value can also be obtained by smooth summation. Indeed, for any cutoff function $\eta$, we can regroup

$$\sum_{n=1}^{\infty} (-1)^{n-1} \eta(\frac{n}{N}) =$$

$$\eta(1/N) + \sum_{m=1}^{\infty} \frac{\eta((2m-1)/N) - 2\eta(2m/N) + \eta((2m+1)/N)}{2}.$$

If $\eta$ is twice continuously differentiable (i.e. $\eta \in C^2$), then from Taylor expansion we see that the summand has size $O(1/N^2)$, and also (from the compact support of $\eta$) is only non-zero when $m = O(N)$. This leads to the asymptotic

$$\sum_{n=1}^{\infty} (-1)^{n-1} \eta(\frac{n}{N}) = \frac{1}{2} + O(\frac{1}{N})$$

and so we recover the value of $1/2$ as the leading term of the asymptotic expansion.

**Exercise 3.7.1.** Show that if $\eta$ is merely once continuously differentiable (i.e. $\eta \in C^1$), then we have a similar asymptotic, but with an error term of $o(1)$ instead of $O(1/N)$. This is an instance of a more general principle that smoother cutoffs lead to better error terms, though the improvement sometimes stops after some degree of regularity.

**Remark 3.7.1.** The most famous instance of smoothed summation is *Cesáro summation*, which corresponds to the cutoff function $\eta(x) := (1-x)_+$. Unsurprisingly, when Cesáro summation is applied to Grandi's series, one again recovers the value of $1/2$.

If we now revisit the divergent series (3.4), (3.5), (3.6), (3.7) with smooth summation in mind, we finally begin to see the origin of the right-hand sides.

Indeed, for any fixed smooth cutoff function $\eta$, we will shortly show that

$$\sum_{n=1}^{\infty} \eta(\frac{n}{N}) = -\frac{1}{2} + C_{\eta,0} N + O(\frac{1}{N}) \tag{3.11}$$

$$\sum_{n=1}^{\infty} n\eta(\frac{n}{N}) = -\frac{1}{12} + C_{\eta,1} N^2 + O(\frac{1}{N}) \tag{3.12}$$

$$\sum_{n=1}^{\infty} n^2\eta(\frac{n}{N}) = C_{\eta,2} N^3 + O(\frac{1}{N}) \tag{3.13}$$

(3.14)

and more generally

$$\sum_{n=1}^{\infty} n^s\eta(\frac{n}{N}) = -\frac{B_{s+1}}{s+1} + C_{\eta,s} N^{s+1} + O(\frac{1}{N}) \tag{3.15}$$

for any fixed $s = 1, 2, 3, \ldots$ where $C_{\eta,s}$ is the *Archimedean factor*

$$C_{\eta,s} := \int_0^\infty x^s \eta(x)\, dx \tag{3.16}$$

(which is also essentially the *Mellin transform* of $\eta$). Thus we see that the values (3.4), (3.5), (3.6), (3.7) obtained by analytic continuation are nothing more than the constant terms of the asymptotic expansion of the *smoothed* partial sums. This is not a coincidence; we will explain the equivalence of these two interpretations of such sums (in the model case when the analytic continuation has only finitely many poles and does not grow too fast at infinity) later in this section.

This interpretation clears up the apparent inconsistencies alluded to earlier. For instance, the sum $\sum_{n=1}^\infty n = 1 + 2 + 3 + \ldots$ consists only of non-negative terms, as does its smoothed partial sums $\sum_{n=1}^\infty n\eta(\frac{n}{N})$ (if $\eta$ is non-negative). Comparing this with (3.13), we see that this forces the highest-order term $C_{\eta,1}N^2$ to be non-negative (as indeed it is), but does not prohibit the *lower-order* constant term $-\frac{1}{12}$ from being negative (which of course it is).

Similarly, if we add together (3.12) and (3.11) we obtain

$$\sum_{n=1}^{\infty} (n+1)\eta(\frac{n}{N}) = -\frac{7}{12} + C_{\eta,1} N^2 + C_{\eta,0} N + O(\frac{1}{N}) \tag{3.17}$$

while if we subtract 1 from (3.12) we obtain

$$\sum_{n=2}^{\infty} n\eta(\frac{n}{N}) = -\frac{13}{12} + C_{\eta,1} N^2 + O(\frac{1}{N}). \tag{3.18}$$

These two asymptotics are not inconsistent with each other; indeed, if we shift the index of summation in (3.18), we can write

$$\sum_{n=2}^{\infty} n\eta(\frac{n}{N}) = \sum_{n=1}^{\infty}(n+1)\eta(\frac{n+1}{N}) \tag{3.19}$$

and so we now see that the discrepancy between the two sums in (3.8), (3.9) come from the shifting of the cutoff $\eta(\frac{n}{N})$, which is invisible in the formal expressions in (3.8), (3.9) but become manifestly present in the smoothed sum formulation.

**Exercise 3.7.2.** By Taylor expanding $\eta(n+1/N)$ and using (3.11), (3.19) show that (3.17) and (3.18) are indeed consistent with each other, and in particular one can deduce the latter from the former.

**3.7.1. Smoothed asymptotics.** We now prove (3.11), (3.12), (3.13), (3.15). We will prove the first few asymptotics by *ad hoc* methods, but then switch to the systematic method of the *Euler-Maclaurin formula* to establish the general case.

For sake of argument we shall assume that the smooth cutoff $\eta : \mathbf{R}^+ \to \mathbf{R}$ is supported in the interval $[0,1]$ (the general case is similar, and can also be deduced from this case by redefining the $N$ parameter). Thus the sum $\sum_{n=1}^{\infty} \eta(\frac{n}{N})x^s$ is now only non-trivial in the range $n \leq N$.

To establish (3.11), we shall exploit the *trapezoidal rule*. For any smooth function $f : \mathbf{R} \to \mathbf{R}$, and on an interval $[n, n+1]$, we see from Taylor expansion that

$$f(n+\theta) = f(n) + \theta f'(n) + O(\|f\|_{\dot{C}^2})$$

for any $0 \leq \theta \leq 1$, $\|f\|_{\dot{C}^2} := \sup_{x\in\mathbf{R}} |f''(x)|$. In particular we have

$$f(n+1) = f(n) + f'(n) + O(\|f\|_{\dot{C}^2})$$

and

$$\int_n^{n+1} f(x)\ dx = f(n) + \frac{1}{2}f'(n) + O(\|f\|_{\dot{C}^2});$$

eliminating $f'(n)$, we conclude that

$$\int_n^{n+1} f(x)\ dx = \frac{1}{2}f(n) + \frac{1}{2}f(n+1) + O(\|f\|_{\dot{C}^2}).$$

Summing in $n$, we conclude the trapezoidal rule

$$\int_0^N f(x)\ dx = \frac{1}{2}f(0) + f(1) + \ldots + f(N-1) + \frac{1}{2}f(N) + O(N\|f\|_{\dot{C}^2}).$$

We apply this with $f(x) := \eta(\frac{x}{N})$, which has a $\dot{C}^2$ norm of $O(1/N^2)$ from the chain rule, and conclude that

$$\int_0^N \eta(\frac{x}{N})\ dx = \frac{1}{2} + \sum_{n=1}^{\infty} \eta(\frac{n}{N}) + O(1/N).$$

But from (3.16) and a change of variables, the left-hand side is just $Nc_{\eta,0}$. This gives (3.11).

The same argument does not quite work with (3.12); one would like to now set $f(x) := x\eta(\frac{x}{N})$, but the $\dot{C}^2$ norm is now too large ($O(1/N)$ instead of $O(1/N^2)$). To get around this we have to refine the trapezoidal rule by performing the more precise Taylor expansion

$$f(n+\theta) = f(n) + \theta f'(n) + \frac{1}{2}\theta^2 f''(n) + O(\|f\|_{\dot{C}^3})$$

where $\|f\|_{\dot{C}^3} := \sup_{x\in\mathbf{R}} |f'''(x)|$. Now we have

$$f(n+1) = f(n) + f'(n) + \frac{1}{2}f''(n) + O(\|f\|_{\dot{C}^3})$$

and

$$\int_n^{n+1} f(x)\ dx = f(n) + \frac{1}{2}f'(n) + \frac{1}{6}f''(n) + O(\|f\|_{\dot{C}^3}).$$

We cannot simultaneously eliminate both $f'(n)$ and $f''(n)$. However, using the additional Taylor expansion

$$f'(n+1) = f'(n) + f''(n) + O(\|f\|_{\dot{C}^3})$$

one obtains

$$\int_n^{n+1} f(x)\ dx = \frac{1}{2}f(n) + \frac{1}{2}f(n+1) + \frac{1}{12}(f'(n) - f'(n+1)) + O(\|f\|_{\dot{C}^3})$$

and thus on summing in $n$, and assuming that $f$ vanishes to second order at $N$, one has (by *telescoping series*)

$$\int_0^N f(x)\ dx = \frac{1}{2}f(0) + \frac{1}{12}f'(0) + \sum_{n=1}^{N} f(n) + O(N\|f\|_{\dot{C}^3}).$$

We apply this with $f(x) := x\eta(\frac{x}{N})$. After a few applications of the chain rule and product rule, we see that $\|f\|_{\dot{C}^3} = O(1/N^2)$; also, $f(0) = 0$, $f'(0) = 1$, and $\int_0^N f(x)\ dx = N^2 c_{\eta,1}$. This gives (3.12).

The proof of (3.13) is similar. With a fourth order Taylor expansion, the above arguments give

$$f(n+1) = f(n) + f'(n) + \frac{1}{2}f''(n) + \frac{1}{6}f'''(x) + O(\|f\|_{\dot{C}^4}),$$

$$\int_n^{n+1} f(x)\ dx = f(n) + \frac{1}{2}f'(n) + \frac{1}{6}f''(n) + \frac{1}{24}f'''(n) + O(\|f\|_{\dot{C}^4})$$

and

$$f'(n+1) = f'(n) + f''(n) + \frac{1}{2}f'''(n) + O(\|f\|_{\dot{C}^4}).$$

Here we have a minor miracle (equivalent to the vanishing of the third Bernoulli number $B_3$) that the $f'''$ term is automatically eliminated when we eliminate the $f''$ term, yielding

$$\int_n^{n+1} f(x)\ dx = \frac{1}{2}f(n) + \frac{1}{2}f(n+1) + \frac{1}{12}(f'(n) - f'(n+1)) + O(\|f\|_{\dot{C}^4})$$

and thus

$$\int_0^N f(x)\ dx = \frac{1}{2}f(0) + \frac{1}{12}f'(0) + \sum_{n=1}^{N} f(n) + O(N\|f\|_{\dot{C}^4}).$$

With $f(x) := x^2\eta(\frac{x}{N})$, the left-hand side is $N^3 c_{\eta,2}$, the first two terms on the right-hand side vanish, and the $\dot{C}^4$ norm is $O(1/N^2)$, giving (3.13).

Now we do the general case (3.15). We define the *Bernoulli numbers* $B_0, B_1, \ldots$ recursively by the formula

$$\sum_{k=0}^{s-1} \binom{s}{k} B_k = s \tag{3.20}$$

for all $s = 1, 2, \ldots$, or equivalently

$$B_{s-1} := 1 - \frac{s-1}{2}B_{s-2} - \frac{(s-1)(s-2)}{3!}B_{s-3} - \ldots - \frac{1}{s}B_0.$$

The first few values of $B_s$ can then be computed:

$$B_0 = 1; B_1 = 1/2; B_2 = 1/6; B_3 = 0; B_4 = -1/30; \ldots.$$

From (3.20) we see that

$$\sum_{k=0}^{\infty} \frac{B_k}{k!}[P^{(k)}(1) - P^{(k)}(0)] = P'(1) \tag{3.21}$$

for any polynomial $P$ (with $P^{(k)}$ being the $k$-fold derivative of $P$); indeed, (3.20) is precisely this identity with $P(x) := x^s$, and the general case then follows by linearity.

As (3.21) holds for all polynomials, it also holds for all formal power series (if we ignore convergence issues). If we then replace $P$ by the formal power series

$$P(x) = e^{tx} = \sum_{k=0}^{\infty} t^k \frac{x^k}{k!}$$

we conclude the formal power series (in $t$) identity

$$\sum_{k=0}^{\infty} \frac{B_k}{k!} t^k (e^t - 1) = te^t$$

leading to the familiar generating function

$$\sum_{k=0}^{\infty} \frac{B_k}{k!} t^k = \frac{te^t}{e^t - 1} \tag{3.22}$$

for the Bernoulli numbers.

If we apply (3.21) with $P$ equal to the antiderivative of another polynomial $Q$, we conclude that

$$\int_0^1 Q(x)\ dx + \frac{1}{2}(Q(1) - Q(0)) + \sum_{k=2}^{\infty} \frac{B_k}{k!}[Q^{(k-1)}(1) - Q^{(k-1)}(0)] = Q(1)$$

which we rearrange as the identity

$$\int_0^1 Q(x)\ dx = \frac{1}{2}(Q(0) + Q(1)) - \sum_{k=2}^{\infty} \frac{B_k}{k!}[Q^{(k-1)}(1) - Q^{(k-1)}(0)]$$

which can be viewed as a precise version of the trapezoidal rule in the polynomial case. Note that if $Q$ has degree $d$, the only the summands with $2 \leq k \leq d$ can be non-vanishing.

Now let $f$ be a smooth function. We have a Taylor expansion

$$f(x) = Q(x) + O(\|f\|_{\dot{C}^{s+2}})$$

for $0 \leq x \leq 1$ and some polynomial $Q$ of degree at most $s+1$; also

$$f^{(k-1)}(x) = Q^{(k-1)}(x) + O(\|f\|_{\dot{C}^{s+2}})$$

for $0 \leq x \leq 1$ and $k \leq s+2$. We conclude that

$$\begin{aligned}\int_0^1 f(x)\ dx = \frac{1}{2}(f(0) + f(1)) \\ - \sum_{k=2}^{s+1} \frac{B_k}{k!}[f^{(k-1)}(1) - f^{(k-1)}(0)] \\ + O(\|f\|_{\dot{C}^{s+2}}).\end{aligned}$$

Translating this by an arbitrary integer $n$ (which does not affect the $\dot{C}^{s+2}$ norm), we obtain

$$\begin{aligned}\int_n^{n+1} f(x)\ dx = \frac{1}{2}(f(n) + f(n+1)) - \sum_{k=2}^{s+1} \frac{B_k}{k!}[f^{(k-1)}(n+1) - f^{(k-1)}(n)] \\ + O(\|f\|_{\dot{C}^{s+2}}).\end{aligned} \tag{3.23}$$

Summing the telescoping series, and assuming that $f$ vanishes to a sufficiently high order at $N$, we conclude the *Euler-Maclaurin formula*

$$\int_0^N f(x)\ dx = \frac{1}{2}f(0) + \sum_{n=1}^{N} f(n) + \sum_{k=2}^{s+1} \frac{B_k}{k!} f^{(k-1)}(0) + O(N\|f\|_{\dot{C}^{s+2}}). \tag{3.24}$$

We apply this with $f(x) := x^s \eta(\frac{x}{N})$. The left-hand side is $c_{\eta,s} N^s$. All the terms in the sum vanish except for the $k = s+1$ term, which is $\frac{B_{s+1}}{s+1}$. Finally, from many applications of the product rule and chain rule (or by viewing $f(x) = N^s g(x/N)$ where $g$ is the smooth function $g(x) := x^s \eta(x)$) we see that $\|f\|_{\dot{C}^{s+2}} = O(1/N^2)$, and the claim (3.15) follows.

**Remark 3.7.2.** By using a higher regularity norm than the $\dot{C}^{s+2}$ norm, we see that the error term $O(1/N)$ can in fact be improved to $O(1/N^B)$ for any fixed $B > 0$, if $\eta$ is sufficiently smooth.

**Remark 3.7.3.** One can *formally* derive the (untruncated) Euler-Maclaurin formula

$$\int_0^N f(x)\ dx = \frac{1}{2}(f(0) - f(N)) + \sum_{n=1}^{N} f(n) + \sum_{k=2}^{\infty} \frac{B_k}{k!} (f^{(k-1)}(0) - f^{(k-1)}(N)) \tag{3.25}$$

(which, after truncating and assuming that $f$ vanishes to high order at $N$, formally gives the main terms of (3.24)) as follows. If we let $D := \frac{d}{dx}$ be the differentiation operator, then the Taylor expansion formula

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2!} f''(x) + \ldots$$

becomes

$$\begin{aligned} f(x+h) &= (1 + hD + \frac{h^2}{2!} D^2 + \ldots) f(x) \\ &= e^{hD} f(x) \end{aligned}$$

(thus translation is the exponentiation of differentiation). In particular, $f(x) = e^{xD} f(0)$, and so (3.25) is formally

$$\int_0^N e^{xD}\ dx = \frac{1}{2}(1 - e^{ND}) + \sum_{n=1}^{N} e^{nD} + \sum_{k=2}^{\infty} \frac{B_k}{k!} D^{k-1}(1 - e^{ND}),$$

where we use the vanishing of $f$ at $N$ to justify the $O(e^{ND})$ error term. Formally, the integral on the left-hand side is $\frac{e^{ND}-1}{D}$, while the geometric series $\sum_{n=1}^{N} e^{nD}$ is $\frac{e^{(N+1)D}-e^D}{e^D-1}$. Meanwhile, from (3.22) one formally has

$$\frac{1}{D} + \frac{1}{2} + \sum_{k=2}^{\infty} \frac{B_k}{k!} D^{k-1} = \frac{e^D}{e^D - 1},$$

and the claim then follows (at a formal level) by some routine algebraic manipulation.

**Exercise 3.7.3.** Use (3.23) to derive Faulhaber's formula (3.10). Note how the presence of boundary terms at $N$ cause the right-hand side of (3.10) to be quite different from the right-hand side of (3.15); thus we see how non-smooth partial summation creates artefacts that can completely obscure the smoothed asymptotics.

**3.7.2. Connection with analytic continuation.** Now we connect the interpretation of divergent series as the constant term of smoothed partial sum asymptotics, with the more traditional interpretation via analytic continuation. For sake of concreteness we shall just discuss the situation with the Riemann zeta function series $\sum_{n=1}^\infty \frac{1}{n^s}$, though the connection extends to far more general series than just this one.

In the previous section, we have computed asymptotics for the partial sums

$$\sum_{n=1}^\infty \frac{1}{n^s}\eta(\frac{n}{N})$$

when $s$ is a negative integer. A key point (which was somewhat glossed over in the above analysis) was that the function $x^{-s}\eta(x)$ was smooth, even at the origin; this was implicitly used to bound various $\dot{C}^k$ norms in the error terms.

Now suppose that $s$ is a complex number with $\mathrm{Re}(s) < 1$, which is not necessarily a negative integer. Then $x^{-s}\eta(x)$ becomes singular at the origin, and the above asymptotic analysis is not directly applicable. However, if one instead considers the telescoped partial sum

$$\sum_{n=1}^\infty \frac{1}{n^s}\eta(\frac{n}{N}) - \sum_{n=1}^\infty \frac{1}{n^s}\eta(2n/N),$$

with $\eta$ equal to 1 near the origin, then by applying (3.24) to the function $f(x) := x^{-s}\eta(\frac{x}{N}) - x^{-s}\eta(2x/N)$ (which vanishes near the origin, and is now smooth everywhere), we soon obtain the asymptotic

$$\sum_{n=1}^\infty \frac{1}{n^s}\eta(\frac{n}{N}) - \sum_{n=1}^\infty \frac{1}{n^s}\eta(2n/N) = c_{\eta,-s}(N^{1-s} - (N/2)^{1-s}) + O(1/N). \tag{3.26}$$

Applying this with $N$ equal to a power of two and summing the telescoping series, one concludes that

$$\sum_{n=1}^\infty \frac{1}{n^s}\eta(\frac{n}{N}) = \zeta(s) + c_{\eta,-s}N^{1-s} + O(1/N) \tag{3.27}$$

for some complex number $\zeta(s)$ which is basically the sum of the various $O(1/N)$ terms appearing in (3.26). By modifying the above arguments, it is not difficult to extend this asymptotic to other numbers than powers of two, and to show that $\zeta(s)$ is independent of the choice of cutoff $\eta$.

From (3.27) we have

$$\zeta(s) = \lim_{N\to\infty} \sum_{n=1}^{\infty} \frac{1}{n^s} \eta(\frac{n}{N}) - c_{\eta,-s} N^{1-s},$$

which can be viewed as a definition of $\zeta$ in the region $\mathrm{Re}(s) < 1$. For instance, from (3.15), we have now proven (3.3) with this definition of $\zeta(s)$. However it is difficult to compute $\zeta(s)$ exactly for most other values of $s$.

For each fixed $N$, it is not hard to see that the expression $\frac{1}{n^s}\eta(\frac{n}{N}) - c_{\eta,-s}N^{1-s}$ is complex analytic in $s$. Also, by a closer inspection of the error terms in the Euler-Maclaurin formula analysis, it is not difficult to show that for $s$ in any compact region of $\{s \in \mathbf{C} : \mathrm{Re}(s) < 1\}$, these expressions converge uniformly as $N \to \infty$. Applying *Morera's theorem*, we conclude that our definition of $\zeta(s)$ is complex analytic in the region $\{s \in \mathbf{C} : \mathrm{Re}\, s < 1\}$.

We still have to connect this definition with the traditional definition (3.1) of the zeta function on the other half of the complex plane. To do this, we observe that

$$c_{\eta,-s} N^{1-s} = \int_0^N x^{-s} \eta(\frac{x}{N})\, dx = \int_1^N x^{-s} \eta(\frac{x}{N})\, dx - \frac{1}{s-1}$$

for $N$ large enough. Thus we have

$$\zeta(s) = \frac{1}{s-1} + \lim_{N\to\infty} \sum_{n=1}^{\infty} \frac{1}{n^s} \eta(\frac{n}{N}) - \int_1^N x^{-s} \eta(\frac{x}{N})\, dx$$

for $\mathrm{Re}\, s < 1$. The point of doing this is that this definition also makes sense in the region $\mathrm{Re}(s) > 1$ (due to the absolute convergence of the sum $\sum_{n=1}^{\infty} \frac{1}{n^s}$ and integral $\int_1^\infty x^{-s} dx$. By using the trapezoidal rule, one also sees that this definition makes sense in the region $\mathrm{Re}(s) > 0$, with locally uniform convergence there also. So we in fact have a globally complex analytic definition of $\zeta(s) - \frac{1}{s-1}$, and thus a *meromorphic* definition of $\zeta(s)$ on the complex plane. Note also that this definition gives the asymptotic

$$\zeta(s) = \frac{1}{s-1} + \gamma + O(|s-1|) \tag{3.28}$$

near $s = 1$, where $\gamma = 0.577\ldots$ is *Euler's constant*.

We have thus seen that asymptotics on smoothed partial sums of $\frac{1}{n^s}$ gives rise to the familiar meromorphic properties of the Riemann zeta function $\zeta(s)$. It turns out that by combining the tools of Fourier analysis and

complex analysis, one can reverse this procedure and deduce the asymptotics of $\frac{1}{n^s}$ from the meromorphic properties of the zeta function.

Let's see how. Fix a complex number $s$ with $\mathrm{Re}(s) < 1$, and a smooth cutoff function $\eta : \mathbf{R}^+ \to \mathbf{R}$ which equals one near the origin, and consider the expression

$$\sum_{n=1}^{\infty} \frac{1}{n^s} \eta(\frac{n}{N}) \tag{3.29}$$

where $N$ is a large number. We let $A > 0$ be a large number, and rewrite this as

$$N^A \sum_{n=1}^{\infty} \frac{1}{n^{s+A}} f_A(\log(n/N))$$

where

$$f_A(x) := e^{Ax} \eta(e^x).$$

The function $f_A$ is in the Schwartz class. By the Fourier inversion formula, it has a Fourier representation

$$f_A(x) = \int_{\mathbf{R}} \hat{f}_A(t) e^{-ixt}\ dt$$

where

$$\hat{f}_A(x) := \frac{1}{2\pi} \int_{\mathbf{R}} f_A(x) e^{ixt}\ dx$$

and so (3.29) can be rewritten as

$$N^A \sum_{n=1}^{\infty} \frac{1}{n^{s+A}} \int_{\mathbf{R}} \hat{f}_A(t) (n/N)^{-it}\ dt.$$

The function $\hat{f}_A$ is also Schwartz. If $A$ is large enough, we may then interchange the integral and sum and use (3.1) to rewrite (3.29) as

$$\int_{\mathbf{R}} \hat{f}_A(t) N^{A+it} \zeta(s + A + it)\ dt.$$

Now we have

$$\hat{f}_A(t) = \frac{1}{2\pi} \int_{\mathbf{R}} e^{(A+it)x} \eta(e^x)\ dx;$$

integrating by parts (which is justified when $A$ is large enough) we have

$$\hat{f}_A(t) = -\frac{1}{2\pi(A+it)} F(A+it)$$

where

$$F(A+it) = \int_{\mathbf{R}} e^{(A+it+1)x} \eta'(e^x)\ dx = \int_0^{\infty} y^{A+it} \eta'(y)\ dy.$$

We can thus write (3.29) as a contour integral

$$\frac{-1}{2\pi i}\int_{s+A-i\infty}^{s+A+i\infty} \zeta(z)\frac{N^{z-s}F(z-s)}{z-s}\,dz.$$

Note that $\eta'$ is compactly supported away from zero, which makes $F(A+it)$ an entire function of $A+it$, which is uniformly bounded whenever $A$ is bounded. Furthermore, from repeated integration by parts we see that $F(A+it)$ is rapidly decreasing as $t\to\infty$, uniformly for $A$ in a compact set. Meanwhile, standard estimates show that $\zeta(\sigma+it)$ is of polynomial growth in $t$ for $\sigma$ in a compact set. Finally, the meromorphic function $\zeta(z)\frac{N^{z-s}F(z-s)}{z-s}$ has a simple pole at $z=1$ (with residue $\frac{N^{1-s}F(1-s)}{1-s}$) and at $z-s$ (with residue $\zeta(s)F(0)$). Applying the residue theorem, we can write (3.29) as

$$\frac{-1}{2\pi i}\int_{s-B-i\infty}^{s-B+i\infty} \zeta(z)\frac{N^{z-s}F(z-s)}{z-s}\,dz - \frac{N^{1-s}F(1-s)}{1-s} - \zeta(s)F(0)$$

for any $B>0$. Using the various bounds on $\zeta$ and $F$, we see that the integral is $O(N^{-B})$. From integration by parts we have $F(0)=-1$ and

$$F(1-s) = -(1-s)\int_0^\infty y^{-s}\eta(y)\,dy = -(1-s)c_{\eta,-s}$$

and thus we have

$$\sum_{n=1}^\infty \frac{1}{n^s}\eta(\frac{n}{N}) = \zeta(s) + c_{\eta,-s}N^{1-s} + O(N^{-B})$$

for any $B>0$, which is (3.15) (with the refined error term indicated in Remark 3.7.2).

The above argument reveals that the simple pole of $\zeta(s)$ at $s=1$ is directly connected to the $c_{\eta,-s}N^{1-s}$ term in the asymptotics of the smoothed partial sums. More generally, if a Dirichlet series

$$D(s) = \sum_{n=1}^\infty \frac{a_n}{n^s}$$

has a meromorphic continuation to the entire complex plane, and does not grow too fast at infinity, then one (heuristically at least) has the asymptotic

$$\sum_{n=1}^\infty \frac{a_n}{n^s}\eta(\frac{n}{N}) = D(s) + \sum_\rho c_{\eta,\rho-s-1}r_\rho N^{\rho-s} + \dots$$

where $\rho$ ranges over the poles of $D$, and $r_\rho$ are the residues at those poles. For instance, one has the famous *explicit formula*

$$\sum_{n=1}^\infty \Lambda(n)\eta(\frac{n}{N}) = c_{\eta,0}N - \sum_\rho c_{\eta,\rho-1}N^\rho + \dots$$

where $\Lambda$ is the *von Mangoldt function*, $\rho$ are the non-trivial zeroes of the Riemann zeta function (counting multiplicity, if any), and ... is an error term (basically arising from the trivial zeroes of zeta); this ultimately reflects the fact that the Dirichlet series

$$\sum_{n=1}^{\infty} \frac{\Lambda(n)}{n^s} = -\frac{\zeta'(s)}{\zeta(s)}$$

has a simple pole at $s = 1$ (with residue $+1$) and simple poles at every zero of the zeta function with residue $-1$ (weighted again by multiplicity, though it is not believed that multiple zeroes actually exist).

The link between poles of the zeta function (and its relatives) and asymptotics of (smoothed) partial sums of arithmetical functions can be used to compare elementary methods in analytic number theory with complex methods. Roughly speaking, elementary methods are based on leading term asymptotics of partial sums of arithmetical functions, and are mostly based on exploiting the simple pole of $\zeta$ at $s = 1$ (and the lack of a simple zero of Dirichlet $L$-functions at $s = 1$); in contrast, complex methods also take full advantage of the zeroes of $\zeta$ and Dirichlet $L$-functions (or the lack thereof) in the entire complex plane, as well as the functional equation (which, in terms of smoothed partial sums, manifests itself through the *Poisson summation formula*). Indeed, using the above correspondences it is not hard to see that the prime number theorem (for instance) is equivalent to the lack of zeroes of the Riemann zeta function on the line $\mathrm{Re}(s) = 1$.

With this dictionary between elementary methods and complex methods, the Dirichlet hyperbola method in elementary analytic number theory corresponds to analysing the behaviour of poles and residues when multiplying together two Dirichlet series. For instance, by using the formula (3.11) and the hyperbola method, together with the asymptotic

$$\sum_{n=1}^{\infty} \frac{1}{n} \eta(\frac{n}{N}) = \int_1^{\infty} \eta(\frac{x}{N}) \, \frac{dx}{x} + \gamma + O(1/N)$$

which can be obtained from the trapezoidal rule and the definition of $\gamma$, one can obtain the asymptotic

$$\sum_{n=1}^{\infty} \tau(n) \eta(\frac{n}{N}) = \int_1^{\infty} \log x \eta(\frac{x}{N}) \, dx + 2\gamma c_{\eta,0} N + O(\sqrt{N})$$

where $\tau(n) := \sum_{d|n} 1$ is the divisor function (and in fact one can improve the $O(\sqrt{N})$ bound substantially by being more careful); this corresponds to the fact that the Dirichlet series

$$\sum_{n=1}^{\infty} \frac{\tau(n)}{n^s} = \zeta(s)^2$$

has a double pole at $s = 1$ with expansion

$$\zeta(s)^2 = \frac{1}{(s-1)^2} + 2\gamma\frac{1}{s-1} + O(1)$$

and no other poles, which of course follows by multiplying (3.28) with itself.

**Remark 3.7.4.** In the literature, elementary methods in analytic number theorem often use sharply truncated sums rather than smoothed sums. However, as indicated earlier, the error terms tend to be slightly better when working with smoothed sums (although not much gain is obtained in this manner when dealing with sums of functions that are sensitive to the primes, such as $\Lambda$, as the terms arising from the zeroes of the zeta function tend to dominate any saving in this regard).

## 3.8. Finitary consequences of the invariant subspace problem

One of the most notorious open problems in functional analysis is the *invariant subspace problem* for Hilbert spaces, which I will state here as a conjecture:

**Conjecture 3.8.1** (Invariant Subspace Problem, ISP0)**.** *Let $H$ be an infinite dimensional complex Hilbert space, and let $T : H \to H$ be a bounded linear operator. Then $H$ contains a proper closed invariant subspace $V$ (thus $TV \subset V$).*

As stated this conjecture is quite infinitary in nature. Just for fun, I set myself the task of trying to find an equivalent reformulation of this conjecture that only involved finite-dimensional spaces and operators. This turned out to be somewhat difficult, but not entirely impossible, if one adopts a sufficiently generous version of "finitary" (cf. [**Ta2008**, §1.3]). Unfortunately, the finitary formulation that I arrived at ended up being rather complicated (in particular, involving the concept of a "barrier"), and did not obviously suggest a path to resolving the conjecture; but it did at least provide some simpler finitary consequences of the conjecture which might be worth focusing on as subproblems.

I should point out that the arguments here are quite "soft" in nature and are not really addressing the heart of the invariant subspace problem; but I think it is still of interest to observe that this problem is not purely an infinitary problem, and does have some non-trivial finitary consequences.

**3.8.1. Initial reductions.** The first reduction is to get rid of the closed invariant subspace $V$, as this will be the most difficult object to finitise. We rephrase ISP0 as

**Conjecture 3.8.2** (Invariant Subspace Problem, ISP1)**.** *Let $H$ be an infinite dimensional complex Hilbert space, and let $T : H \to H$ be a bounded linear operator. Then there exist unit vectors $v, w \in H$ such that $\langle T^n v, w\rangle = 0$ for all natural numbers $n \in \mathbf{N}$.*

Indeed, to see that ISP1 implies ISP0, we simply take $V$ to be the closed invariant subspace generated by the orbit $v, Tv, T^2v, \ldots$, which is proper since it is orthogonal to $w$. To see that ISP0 implies ISP1, we let $v$ be an arbitrary unit vector in the invariant subspace $V$, and $w$ be an arbitrary unit vector in the orthogonal complement $V^\perp$.

The claim is obvious if $H$ is not separable (just let $v$ be arbitrary, and $w$ to be a normal vector to the separable space spanned by $v, Tv, T^2v, \ldots$), so we may normalise $H$ to be $\ell^2(\mathbf{N})$. We may also normalise $T$ to be a contraction (thus $\|T\|_{op} \leq 1$), and let $(a_{ij})_{i,j\geq 1}$ be the coefficients of $T$.

The next step is to restrict $T$ to a compact space of operators. Define a *growth function* to be a monotone increasing function $F : \mathbf{N} \to \mathbf{N}$. Given any growth function $F$, we say that a linear contraction $T : \ell^2(\mathbf{N}) \to \ell^2(\mathbf{N})$ with coefficients $(a_{ij})_{i,j\geq 1}$ is *$F$-tight* if one has the bound

$$\sup_{1\leq i\leq N} \sum_{j\geq F(N)} |a_{ij}|^2 \leq \frac{1}{N} \tag{3.30}$$

and

$$\sup_{1\leq j\leq N} \sum_{i\geq F(N)} |a_{ij}|^2 \leq \frac{1}{N}. \tag{3.31}$$

For instance, if the matrix $(a_{ij})_{i,j\geq 1}$ is band-limited to the region $|j - i| \leq 10$, it is $F$-tight with $F(N) := N + 11$. If it is limited to the region $i/2 \leq j \leq 2i$, then it is $F$-tight with $F(N) := 2N + 1$. So one can view $F$-tightness as a weak version of the band-limited property.

The significance of this concept lies in the following lemma:

**Lemma 3.8.3** (Sequential compactness)**.** (i) *Every contraction $T : \ell^2(\mathbf{N}) \to \ell^2(\mathbf{N})$ is $F$-tight with respect to at least one growth function $F$.*

(ii) *If $F$ is a growth function and $T_1, T_2, \ldots$ is a sequence of $F$-tight contractions, then there exists a subsequence $T_{n_k}$ which converges in the strong operator topology to an $F$-tight contraction $T$. Furthermore, the adjoints $T^*_{n_k}$ converge in the strong operator topology to $T^*$.*

**Proof.** To prove (i), observe that if $T$ is a contraction and $N \geq 1$, then

$$\sup_{1\leq i\leq N} \sum_{j=1}^{\infty} |a_{ij}|^2 \leq 1$$

and

$$\sup_{1\leq j\leq N}\sum_{i=1}^{\infty}|a_{ij}|^2\leq 1$$

and hence by the monotone convergence theorem we can find $F(N)$ such that (3.30), (3.31). By increasing $F(N)$ as necessary one can make $F$ monotone.

To prove (ii), we apply the usual Arzelá-Ascoli diagonalisation argument to extract a subsequence $T_{n_k}=(a_{i,j,n_k})_{i,j\geq 1}$ that converges componentwise (i.e. in the weak operator topology) to a limit $T=(a_{i,j})_{i,j\geq 1}$. From Fatou's lemma we see that $T$ is an $F$-tight contraction. From the tightness one can upgrade the weak operator topology convergence to strong operator topology convergence (i.e.

$$\lim_{k\to\infty}\sum_{j=1}^{\infty}|a_{i,j,n_k}-a_{i,j}|^2=0$$

for all $i$) by standard arguments, and similarly for the adjoints. □

We will similarly need a way to compactify the unit vectors $v,w$. If $F$ is a growth function and $0<N_1<N_2<N_3<\ldots$ are natural numbers, we say that a unit vector $v=(v_i)_{i=1}^{\infty}$ is $F,N_1,N_2,\ldots$-*tight* if one has

$$\sum_{i\geq N_k}|v_i|^2\leq\frac{1}{F(N_{k-1})} \tag{3.32}$$

for all $k\geq 1$, with the convention that $N_0=0$. Similarly, we say that $v$ is $F,N_1,\ldots,N_K$-*tight* if one has (3.32) for all $1\leq k\leq K$. One has the following variant of Lemma 3.8.3:

**Lemma 3.8.4** (Sequential compactness)**.** *Let $F$ be a growth function.*

(i) *Every unit vector $v$ is $F,N_1,N_2,\ldots$-tight with respect to at least one increasing sequence $0<N_1<N_2<\ldots$. In fact any finite number of unit vectors $v_1,\ldots,v_m$ can be made $F,N_1,N_2,\ldots$-tight with the same increasing sequence $0<N_1<N_2<\ldots$.*

(ii) *If $0<N_1<N_2<\ldots$, and for each $k\geq 1$, $v_k$ is a $F,N_1,\ldots,N_k$-tight unit vector, then there exists a subsequence $v_{n_l}$ of $v_k$ that converges strongly to an $F,N_1,N_2,\ldots$-tight unit vector $v$.*

The proof of this lemma is routine and is omitted.

In view of these two lemmas, ISP0 or ISP1 is equivalent to

**Conjecture 3.8.5** (Invariant Subspace Problem, ISP2)**.** *Let $F$ be a growth function, and let $T=(a_{ij})_{i,j\geq 1}$ be an $F$-tight contraction. Then there exist a sequence $0<N_1<N_2<\ldots$ and a pair of $F,N_1,N_2,\ldots$-tight unit vectors $v,w\in\ell^2(\mathbf{N})$ such that $\langle T^n v,w\rangle=0$ for all natural numbers $n\in\mathbf{N}$.*

The compactness given by the $F$-tightness and $F, N_1, N_2, \ldots$ will be useful for finitising later.

**3.8.2. Finitising.** Now we need a more complicated object.

**Definition 3.8.6.** A *barrier* is a family $\mathcal{T}$ of finite tuples $(N_1, \ldots, N_m)$ of increasing natural numbers $0 < N_1 < \ldots < N_m$, such that

(i) Every infinite sequence $N_1 < N_2 < \ldots$ of natural numbers has at least one initial segment $(N_1, \ldots, N_m)$ in $\mathcal{T}$; and

(ii) If $(N_1, \ldots, N_m)$ is a sequence in $\mathcal{T}$, then no initial segment $(N_1, \ldots, N_{m'})$ with $m' < m$ lies in $\mathcal{T}$.

Examples of barriers include

(1) The family of all tuples $(N_1, \ldots, N_m)$ of increasing natural numbers with $m = 10$;

(2) The family of all tuples $(N_1, \ldots, N_m)$ of increasing natural numbers with $m = N_1 + 1$;

(3) The family of all tuples $(N_1, \ldots, N_m)$ of increasing natural numbers with $m = N_{N_1+1} + 1$.

We now claim that ISP2 is equivalent to the following finitary statement. Let $\ell^2(N) \equiv \mathbf{C}^N$ denote the $\ell^2$ space on $\{1, \ldots, N\}$.

**Conjecture 3.8.7** (Finitary invariant subspace problem, FISP0)**.** *Let $F$ be a growth function, and let $\mathcal{T}$ be a barrier. Then there exists a natural number $N_*$ such that for every $F$-tight contraction $T : \ell^2(F(N_*)) \to \ell^2(F(N_*))$, there exists a tuple $(N_1, \ldots, N_m)$ in $\mathcal{T}$ with $0 < N_1 < \ldots < N_m < N_*$, and $F, N_1, \ldots, N_m$-tight unit vectors $v, w \in \ell^2(F(N_*))$, such that $|\langle T^n v, w\rangle| \leq \frac{1}{F(N_m)}$ for all $0 \leq n \leq F(N_m)$.*

We now show that ISP2 and FISP0 are equivalent.

**Proof of ISP2 assuming FISP0.** Let $F$ be a growth function, and let $T$ be an $F$-tight contraction. Let $\mathcal{T}'$ denote the set of all tuples $0 < N_1 < \ldots < N_m$ with $m > 1$ such that there does not exist $F, N_1, \ldots, N_m$-tight unit vectors $v, w \in \ell^2(\mathbf{N})$ such that $|\langle T^n v, w\rangle| \leq \frac{2}{m}$ holds for all $0 \leq n \leq m$. Let $\mathcal{T}$ be those elements of $\mathcal{T}'$ that contain no proper initial segment in $\mathcal{T}'$.

Suppose first that $\mathcal{T}$ is not a barrier. Then there exists an infinite sequence $0 < N_1 < N_2 < \ldots$ such that $(N_1, \ldots, N_m) \notin \mathcal{T}$ for all $m$, and thus $(N_1, \ldots, N_m) \notin \mathcal{T}'$ for all $m$. In other words, for each $m$ there exists $F, N_1, \ldots, N_m$-tight unit vectors $v_m, w_m \in \ell^2(\mathbf{N})$ such that $|\langle T^n v_m, w_m\rangle| \leq \frac{2}{m}$ for all $0 \leq n \leq m$. By Lemma 3.8.4, we can find a subsequence $v_{m_j}, w_{m_j}$ that converge strongly to $F, N_1, N_2, \ldots$-tight unit vectors $v, w$. We conclude that $\langle T^n v, w\rangle = 0$ for all $n \geq 0$, and ISP2 follows.

Now suppose instead that $\mathcal{T}$ is a barrier. Let $F'$ be a growth function larger than $F$ to be chosen later. Then the $F$-tight contraction $T$ is also $F'$-tight, as is the restriction $T|_N : \ell^2(N) \to \ell^2(N)$ of $T$ to any finite subspace. Using FISP0, we can thus find $0 < N_1 < \ldots < N_m < N_*$ with $(N_1, \ldots, N_m) \in \mathcal{T}$ and $F', N_1, \ldots, N_m$-tight unit vectors $v, w \in \ell^2(F'(N_*))$ such that

$$|\langle (T|_{F'(N_m)})^n v, w\rangle| \leq \frac{1}{F'(N_m)}$$

for all $0 \leq n \leq F'(N_m)$, and in particular for all $0 \leq n \leq m$. Note that $v, w$ are almost in $\ell^2(N_m)$, up to an error of $1/F'(N_{m-1})$. From this and the $F$-tightness of the contraction $T$, we see (if $F'$ is sufficiently rapid) that $(T|_{F'(N_m)})^n v$ and $T^n v$ differ by at most $1/m$ for $0 \leq n \leq m$. We conclude that

$$|\langle T^n v, w\rangle| \leq \frac{2}{m},$$

and so $(N_1, \ldots, N_m) \notin \mathcal{T}$, a contradiction. This yields the proof of ISP2 assuming FISP0.

**Proof of FISP0 assuming ISP2.** Suppose that FISP0 fails. Then there exists a growth function $F$ and a barrier $\mathcal{T}$ such that, for every $N_*$, there exists an $F$-tight contraction $T_{N_*} : \ell^2(F(N_*)) \to \ell^2(F(N_*))$ such that there does not exist any tuples $(N_1, \ldots, N_m)$ in $\mathcal{T}$ with $0 < N_1 < \ldots < N_m < N_*$, and $F, N_1, \ldots, N_m$-tight unit vectors $v, w \in \ell^2(F(N_*))$, such that $|\langle T^n v, w\rangle| \leq \frac{1}{F(N_m)}$ for all $0 \leq n \leq F(N_m)$.

We extend each $T_{N_*}$ by zero to an operator on $\ell^2(\mathbf{N})$, which is still a $F$-tight contraction. Using Lemma 3.8.3, one can find a sequence $N_{*,k}$ going to infinity such that $T_{N_{*,k}}$ converges in the strong (and dual strong) operator topologies to an $F$-tight contraction $T$. Let $F'$ be a growth function larger than $F$ to be chosen later. Applying ISP2, there exists an infinite sequence $0 < N_1 < N_2 < \ldots$ and $F', N_1, N_2, \ldots$-tight unit vectors $v, w \in \ell^2(\mathbf{N})$ such that $\langle T^n v, w\rangle = 0$ for all $n \geq 0$.

As $\mathcal{T}$ is a barrier, there exists a finite initial segment $(N_1, \ldots, N_m)$ of the above sequence that lies in $\mathcal{T}$. For $k$ sufficiently large, we have $N_{*,k} \geq N_m$, and also we see from the strong operator norm convergence of $T_{N_{*,k}}$ to $T$ (and thus $T^n_{N_{*,k}}$ to $T^n$ for any $n$, as all operators are uniformly bounded) that

$$|\langle T^n_{N_{*,k}} v, w\rangle| \leq \frac{1}{F'(N_m)}$$

for all $0 \leq n \leq F(N_m)$.

Now we restrict $v, w$ to $\ell^2(F(N_{*,k}))$, and then renormalise to create unit vectors $v', w' \in \ell^2(F(N_{*,k}))$. For $k$ large enough, we have

$$\|v - v'\|, \|w - w'\| \leq 1/F'(N_m)$$

and we deduce (for $F'$ large enough) that $v', w'$ are $F, N_1, N_2, \ldots, N_m$-tight and $|\langle T^n v, w\rangle| \leq \frac{1}{F(N_m)}$ for all $0 \leq n \leq F(N_m)$. But this contradicts the construction of the $T_{N_*}$, and the claim follows.

**3.8.3. A special case.** The simplest example of a barrier is the family of 1-tuples $(N)$, and one of the simplest examples of an $F$-tight contraction is a contraction that is 1-band-limited, i.e. the coefficients $a_{ij}$ vanish unless $|i-j| \leq 1$. We thus obtain

**Conjecture 3.8.8** (Finitary invariant subspace problem, special case, FISP1)**.** *Let $F$ be a growth function and $\varepsilon > 0$. Then there exists a natural number $N_*$ such that for every 1-band-limited contraction $T : \ell^2(F(N_*)) \to \ell^2(F(N_*))$, there exists $0 < N < N_*$ and unit vectors $v, w \in \ell^2(F(N_*))$ with*

$$\sum_{j \geq N} |v_j|^2, \sum_{j \geq N} |w_j|^2 \leq \varepsilon^2$$

*(i.e. $v, w$ are $\varepsilon$-close to $\ell^2(N)$) such that $|\langle T^n v, w\rangle| \leq \frac{1}{F(N)}$ for all $0 \leq n \leq F(N)$.*

This is perhaps the simplest case of ISP that I do not see how to resolve[7]. Here is a slightly weaker version that I still cannot resolve:

**Conjecture 3.8.9** (Finitary invariant subspace problem, special case, FISP2)**.** *Let $F$ be a growth function, let $\varepsilon > 0$, and let $T : \ell^2(\mathbf{N}) \to \ell^2(\mathbf{N})$ be a 1-band-limited contraction. Then there exists $N > 0$ and unit vectors $v, w \in \ell^2(\mathbf{N})$ such that*

$$\sum_{j \geq N} |v_j|^2, \sum_{j \geq N} |w_j|^2 \leq \varepsilon^2$$

*(i.e. $v, w$ are $\varepsilon$-close to $\ell^2(N)$) such that $|\langle T^n v, w\rangle| \leq \frac{1}{F(N)}$ for all $0 \leq n \leq F(N)$.*

This claim is implied by ISP but is significantly weaker than it. Informally, it is saying that one can find two reasonably localised vectors $v, w$, such that the orbit of $v$ is highly orthogonal to $w$ for a very long period of time, much longer than the degree to which $v, w$ are localised.

**3.8.4. Notes.** I am indebted to Henry Towsner for many discussions on this topic, and to the MathOverflow community for describing the concept of a barrier.

[7] Note that the finite-dimensional operator $T : \ell^2(F(N_*)) \to \ell^2(F(N_*))$ will have plenty of (generalised) eigenvectors, but there is no particular reason why any of them are "tight" in the sense that they are $\varepsilon$-close to $\ell^2(N)$.

## 3.9. The Guth-Katz result on the Erdős distance problem

*Combinatorial incidence geometry* is the study of the possible combinatorial configurations between geometric objects such as lines and circles. One of the basic open problems in the subject has been the *Erdős distance problem* [**Er1946**]:

**Problem 3.9.1** (Erdős distance problem)**.** Let $N$ be a large natural number. What is the least number $\#\{|x_i - x_j| : 1 \leq i < j \leq N\}$ of distances that are determined by $N$ points $x_1, \ldots, x_N$ in the plane?

Erdos called this least number $g(N)$. For instance, one can check that $g(3) = 1$ and $g(4) = 2$, although the precise computation of $g$ rapidly becomes more difficult after this. By consiging $N$ points in arithmetic progression, we see that $g(N) \leq N - 1$. By considering the slightly more sophisticated example of a $\sqrt{N} \times \sqrt{N}$ lattice grid (assuming that $N$ is a square number for simplicity), and using some analytic number theory, one can obtain the slightly better asymptotic bound $g(N) = O(N/\sqrt{\log N})$.

On the other hand, lower bounds are more difficult to obtain. As observed by Erdos, an easy argument, ultimately based on the incidence geometry fact that any two circles intersect in at most two points, gives the lower bound $g(N) \gg N^{1/2}$. The exponent $1/2$ has been slowly increasing over the years by a series of increasingly intricate arguments combining incidence geometry facts with other known results in combinatorial incidence geometry (most notably the *Szemerédi-Trotter theorem* [**SzTr1873**]) and also some tools from additive combinatorics; however, these methods seemed to fall quite short of getting to the optimal exponent of 1. Indeed, until last year, the best lower bound known was approximately $N^{0.8641}$, due to Katz and Tardos [**KaTa2004**].

Very recently, though, Guth and Katz [**GuKa2010b**] have obtained a near-optimal result:

**Theorem 3.9.2.** *One has $g(N) \gg N/\log N$.*

The proof neatly combines together several powerful and modern tools in a new way: a recent geometric reformulation of the problem due to Elekes and Sharir [**ElSh2010**]; the polynomial method as used recently by Dvir [**Dv2009**], Guth [**Gu2010**], and Guth-Katz [**GuKa2010**] on related incidence geometry problems (discussed in [**Ta2009b**, §1.1, 1.7]); and the somewhat older method of cell decomposition (discussed in [**Ta2010b**, §1.4]). A key new insight is that the polynomial method (and more specifically, the *polynomial Ham Sandwich theorem*) can be used to efficiently create cells.

In this post, I thought I would sketch some of the key ideas used in the proof, though I will not give the full argument here (the paper [**GuKa2010b**]

itself is largely self-contained, well motivated, and of only moderate length). In particular I will not go through all the various cases of configuration types that one has to deal with in the full argument, but only some illustrative special cases.

To simplify the exposition, I will repeatedly rely on "pigeonholing cheats". A typical such cheat: if I have $n$ objects (e.g. $n$ points or $n$ lines), each of which could be of one of two types, I will assume[8] that either all $n$ of the objects are of the first type, or all $n$ of the objects are of the second type. A related such cheat[9]: if one has $n$ objects $A_1, \ldots, A_n$ (again, think of $n$ points or $n$ circles), and to each object $A_i$ one can associate some natural number $k_i$ (e.g. some sort of "multiplicity" for $A_i$) that is of "polynomial size" (of size $O(N^{O(1)})$), then I will assume in fact that all the $k_i$ are in a fixed dyadic range $[k, 2k]$ for some $k$. Using the notation $X \sim Y$ to denote the assertion that $C^{-1}Y \leq X \leq CY$ for an absolute constant $C$, we thus have $k_i \sim k$ for all $i$, thus $k_i$ is morally constant.

I will also use asymptotic notation rather loosely, to avoid cluttering the exposition with a certain amount of routine but tedious bookkeeping of constants. In particular, I will use the informal notation $X \lll Y$ or $Y \ggg X$ to denote the statement that $X$ is "much less than" $Y$ or $Y$ is "much larger than" $X$, by some large constant factor.

**3.9.1. Reduction to a linear problem.** Traditionally, the Erdos distance problem has been attacked by first casting it as a question about incidences between circles, by starting with the trivial observation that if two points $x_i, x_j$ are equidistant from a third point $x_k$, then $x_i, x_j$ lie on a circle centred at $x_k$.

The incidence geometry of circles, however, is not quite as well understood as the incidence geometry of lines, and so one often then converted the circle incidence problem to a line incidence problem, for instance by using the elementary Euclidean geometry fact that if two circles intersected at a pair of points, then the centres of these circles would lie on the perpendicular bisector of that pair of points. Indeed, by combining this elementary observation with the celebrated *Szemerédi-Trotter theorem* [**SzTr1873**] that gives a sharp incidence bound for a collection of lines and points, Chung, Szemerédi, and Trotter [**ChSzTr1992**] were already able to obtain the respectable lower bound of $g(N) \gg N^{4/5-o(1)}$.

[8] In truth, I can only assume that at least $n/2$ of the objects are of the first type, or at least $n/2$ of the objects are of the second type; but in practice, having $n/2$ instead of $n$ only ends up costing an unimportant multiplicative constant in the type of estimates used here.

[9] In practice, the dyadic pigeonhole principle can only achieve this after throwing away all but about $n/\log N$ of the original $n$ objects; it is this type of logarithmic loss that eventually leads to the logarithmic factor in the main theorem.

The first innovation of Guth and Katz (which builds upon earlier work in this direction in [**ElSh2010**]) is to use elementary Euclidean geometry to recast the distance problem as a linear problem in a slightly different fashion, namely as an incidence problem of lines in *three dimensions* $\mathbf{R}^3$ rather than two.

Let's see how this works. To prove the main theorem, we will argue by contradiction, assuming that $g(N) \ll N/\log N$ for some large $N$. Thus, we can find $N$ points $x_1, \ldots, x_N$ in the plane which only subtend $g(N) \ll N/\log N$ distances. We think of these distances as the lengths of line segments $\overline{x_i x_j}$ connecting two of the points $x_i, x_j$. There are $\binom{N}{2} \sim N^2$ different line segments $\overline{x_i x_j}$ that have these lengths. A standard application of the Cauchy-Schwarz inequality then shows that there must be many pairs of distinct but congruent line segments $\overline{x_i x_j}$, $\overline{x_k x_l}$ (i.e. line segments of equal length $|x_i - x_j| = |x_k - x_l|$); in fact, their number must be

$$\frac{\binom{N}{2}^2}{g(N)} - \binom{N}{2} \gg N^3 \log N.$$

So, we have a lot ($\gg N^3 \log N$) of pairs $\overline{x_i x_j}, \overline{x_k x_l}$ of congruent line segments. Now we use a trivial but fundamentally important observation of Elekes and Sharir to recast the problem in terms of rigid motions:

**Proposition 3.9.3.** *Two line segments $\overline{AB}, \overline{CD}$ in the plane are congruent if and only if there is an orientation-preserving* rigid motion *that maps $A$ to $C$ and $B$ to $D$. Furthermore, this rigid motion is unique.*

**Proof.** Translate $A$ to $C$ and then rotate appropriately. □

**Remark 3.9.4.** The above observation exploits the fact that Euclidean geometry is a *Klein geometry* - a geometry determined by a Lie group of isometries, which in this case is the *two-dimensional special Euclidean group* $SE(2) \equiv SO(2) \ltimes \mathbf{R}^2$ of orientation-preserving rigid motions. It is plausible that the arguments here can extend to other Klein geometries. However, it is much less clear as to what one can salvage from this argument when working with a much less symmetric geometry, such as one coming from a more general metric space. This is in contrast to much of the previous work on this problem, which exploited somewhat different features of Euclidean geometry, such as incidence properties or arithmetic structure.

From the above proposition we thus see that we can find $\gg N^3 \log N$ quintuples $(x_i, x_j, x_k, x_l, R)$, where $R \in SE(2)$ is such that $R(x_i) = R(x_k)$ and $R(x_j) = R(x_l)$.

We now dualise the problem; rather than think about rigid motions acting on pairs of points, we think about pairs of points describing a set

of rigid motions. For any pair of points $x, y$, let $\ell_{x\to y} \subset SE(2)$ be the set of all rigid motions that map $x$ to $y$. This is a one-dimensional subset of $SE(2)$; indeed, $\ell_{x\to y}$ consists of the translation from $x$ to $y$, together with rotations around an origin $p$ that lies on the perpendicular bisector of $\overline{xy}$, with a rotation angle $\theta$ obeying the relation

$$\left|\cot\frac{\theta}{2}\right| = \frac{|\overline{pm}|}{|\overline{mx}|}$$

where $m := \frac{x+y}{2}$ is the midpoint of $x$ and $y$. If we discard the translation (which can easily be dealt with by a separate argument) as well as the point reflection (corresponding to the case $p = m, \theta = \pi$), and focus only on the rotations by angles less than $\pi$, we in fact see that the origin $p$ of the rotation depends in a linear fashion on the quantity $\cot\frac{\theta}{2}$. Thus, if we parameterise (most of) $SE(2)$ by the coordinates $(p, \cot\frac{\theta}{2}) \in \mathbf{R}^3$, we thus see that we can view each $\ell_{x\to y}$ as a line in $\mathbf{R}^3$, and we will adopt this perspective for the rest of this post.

**Remark 3.9.5.** It may seem slightly miraculous that the set $\ell_{x\to y}$ of rigid motions from $x$ to $y$ has the geometry of a line; we will give an explanation of this phenomenon elsewhere.

Let $\mathcal{L}$ be the collection of all the lines $\ell_{x_i\to x_k} \subset \mathbf{R}^3$ generated by pairs of points $x_i, x_k$ from our collection, thus there are $\sim N^2$ such lines[10]. The $\ggg N^3 \log N$ quintuples $(x_i, x_j, x_k, x_l, R)$ described earlier give rise to $\ggg N^3 \log N$ intersecting pairs $\ell, \ell' \in \mathcal{L}$ of lines in $\mathcal{L}$. So the question now comes down to a simple question about incidences of lines: is it possible for $\sim N^2$ lines in three-dimensional space $\mathbf{R}^3$ to generate $\ggg N^3 \log N$ pairs $\ell, \ell'$ of intersecting lines? If the answer to this question is "no", then we are done.

Unfortunately, the answer to the question is "yes". One quickly comes up with two basic counterexamples to that question:

(1) (Concurrency) If one has $\sim N^2$ lines that all go through the same point $p$, then we will have $\sim N^4$ pairs of intersecting lines.

(2) (Coplanarity) If one has $\sim N^2$ lines that all lie in the same plane $\pi$, with no two lines being parallel, then we will have $\sim N^4$ pairs of intersecting lines.

Slightly less obviously, there is a third counterexample that comes from a *regulus* (or *hyperbolic paraboloid*) - a *doubly ruled surface* in $\mathbf{R}^3$. A typical example of a regulus is the set $\{(x, y, z) \in \mathbf{R}^3 : z = xy\}$. On the one hand, this surface is ruled by the lines $\{(x, y, xy) : y \in \mathbf{R}\}$ for $x \in \mathbf{R}$; on the other hand, it is also ruled by the lines $\{(x, y, xy) : x \in \mathbf{R}\}$ for $y \in \mathbf{R}$. If we then

[10] It is easy to see that different pairs of points $(x_i, x_k)$ lead to different lines.

pick $\sim N^2$ lines from the first family of lines and $\sim N^2$ from the second family, then we obtain another family of $\sim N^2$ lines that lead to $\sim N^4$ pairs of intersecting lines.

However, all is not lost here, because we are not dealing with an *arbitrary* family $\mathcal{L}$ of $\sim N^2$ lines in $\mathbf{R}^3$, but rather with a *special* family that was generated ultimately by $N$ points $x_1, \ldots, x_N$ in $\mathbf{R}^2$. As such, the special structure of this family can be used to rule out the concurrency, coplanarity, and regulus counterexmaples. Indeed, observe that if a rigid motion $R$ maps $x_i$ to $x_j$, then it cannot also map $x_i$ to $x_k$ for some $k \neq j$. This implies that $\ell_{x_i \to x_j}$ and $\ell_{x_i \to x_k}$ cannot intersect. This limits the amount of concurrency in $\mathcal{L}$; letting $i$ run from 1 to $N$, we indeed conclude that at most $N$ lines in $\mathcal{L}$ can meet at a point, which eliminates the concurrency counterexample[11].

In a similar spirit, observe that for a given plane $\pi$, the requirement that a line in $\mathbf{R}^3$ lie in that plane is a codimension two constraint on that line (the space of lines in a plane is two-dimensional, while the space of lines in $\mathbf{R}^3$ is four-dimensional). Thus, for fixed $x_i$, the requirement that $\ell_{x_i \to x_j}$ lie in $\pi$ is a codimension two constraint on $x_j$, and thus by *Bezout's theorem*, it should[12] only be satisfiable by $O(1)$ values of $x_j$. Letting $i$ run from 1 to $N$, we conclude that any plane $\pi$ can contain at most $O(N)$ lines from $\mathcal{L}$, thus neutralising the coplanarity counterexample. The same argument also shows that any regulus also can contain at most $O(N)$ lines from $\mathcal{L}$ (because a regulus is an algebraic surface of degree 2, and so the Bezout argument still applies).

Now, it turns out that the elimination of the concurrency, coplanarity, and regulus obstructions allows us to now give the right answer to the previous question. Indeed, Guth and Katz show

**Theorem 3.9.6.** *Let $\mathcal{L}$ be a collection of $\sim N^2$ lines in $\mathbf{R}^3$, such that at most $N$ lines in $\mathcal{L}$ meet at a point, and such that any plane or regulus contains at most $O(N)$ lines in $\mathcal{L}$. Then there are at most $O(N^3 \log N)$ pairs $\ell, \ell' \in \mathcal{L}$ of intersecting lines in $\mathcal{L}$.*

The above discussion has shown (modulo a few details) that Theorem 3.9.6 implies Theorem 3.9.2, so it suffices now to show Theorem 3.9.6. This is a significant simplification, because it is an assertion about incidences of lines, rather than congruence of line segments or incidences of circles, and we know a lot more about how configurations of lines intersect than we do about configurations of circles or of congruences of line segments.

[11] More precisely, the best one now can do is get $\sim N$ groups of $N$ concurrent lines, but this only yields $\sim N^3$ pairs of intersecting lines, which is acceptable.

[12] One has to check that the constraints are non-degenerate, but this is routine; actually, in this particular case one can also argue directly using elementary Euclidean geometry instead of Bezout's theorem.

Of course, it remains to prove Theorem 3.9.6. It is convenient to pigeonhole based on the concurrency of the lines in $\mathcal{L}$. Let $k \geq 2$, and suppose that has a set of points $S$, such that for each point $p$ in $S$ there are at least $k$ lines in $\mathcal{L}$ passing through $p$. From the concurrency bound we know that $k \leq N$. Then each point in $p$ contributes $\sim k^2$ pairs of intersecting lines, so we need to show that $P$ does not get much larger than $N^3/k^2$. And indeed this is the case:

**Theorem 3.9.7.** *Let $\mathcal{L}$ be a collection of $\sim N^2$ lines in $\mathbf{R}^3$, such that any plane or regulus contains at most $O(N)$ lines in $\mathcal{L}$. Let $2 \leq k \leq N$, and let $S$ be a set of points, each of which has at least $k$ lines in $\mathcal{L}$ passing through it. Then $|S| \ll N^3/k^2$.*

A simple dyadic summation argument allows one to deduce Theorem 3.9.6 from[13] Theorem 3.9.7. The hypothesis that each plane or regulus contains at most $O(N)$ lines, incidentally, is very reminiscent of the *Wolff axiom* in the work on the Kakeya conjecture; see [**Wo1995**].

It is worth noting that the bound in Theorem 3.9.7 is completely sharp. Indeed, consider two parallel grids

$$\{(i,j,0) \in \mathbf{Z}^2 \times \{0\} : 1 \leq i,j \leq \sqrt{N}\}$$

and

$$\{(i,j,1) \in \mathbf{Z}^2 \times \{1\} : 1 \leq i,j \leq \sqrt{N}\}$$

and let $\mathcal{L}$ be the set of lines connecting a point in the first grid to a point in the second grid. Then one can verify that $\mathcal{L}$ obeys the required hypotheses for Theorem 3.9.7, and a number-theoretic calculation[14] shows that for any $2 \leq k \ll N$, the number of points in $\mathbf{R}^3$ that are incident to at least $k$ lines in $\mathcal{L}$ is $\sim N^3/k^2$.

It is also worth noting that if one replaces the real line $\mathbf{R}$ by a finite field $\mathbf{F}_q$, then the claim fails for large $k$. Indeed, if one sets $N = q^2$ and $\mathcal{L}$ to be the set of *all* lines in $\mathbf{F}_q^3$, then the hypotheses of Theorem 3.9.7 hold, but each of the $q^3$ points in $\mathbf{F}_q^3$ is incident to $\sim q^2$ lines in $\mathcal{L}$, which soon leads to a significant violation of the conclusion of that theorem. Thus, any proof of Theorem 3.9.7 in the large $k$ case cannot be purely algebraic in nature must somehow use a property of the real line that is not shared by finite fields. This property will be the *ordered* nature of the real line, which manifests itself in the Szemerédi-Trotter theorem and in the ham sandwich theorem, both of which are used in the proof. However, this example does not prevent the small $k$ case from being purely algebraic.

[13] Note that the case $k = 1$ is not relevant, as the points associated to this case do not contribute any pairs of intersecting lines.

[14] To see this, first look at the cases $k = 2$ and $k \sim N^2$, and then interpolate the arguments; details are given in [**GuKa2010b**].

So, the only thing left to do is to prove Theorem 3.9.7. It turns out that the multiplicity 2 case $k = 2$ and the higher multiplicity case $k > 2$ have to be treated separately, basically because $k$ concurrent lines will be trapped in a plane when $k = 2$, but will usually not be so trapped when $k > 2$. Also, note that the regulus obstruction only generates intersections of multiplicity 2; when $k > 2$, the hypothesis that a regulus contains $O(N)$ points is not used and so can be dropped.

**3.9.2. The $k = 2$ case.** We first verify Theorem 3.9.7 in the $k = 2$ case, i.e. we show that the lines $\mathcal{L}$ in that theorem can intersect in at most $O(N^3)$ points. As hinted in the previous discussion, the argument here is purely algebraic in nature, using just the polynomial method (similar to, say, how Dvir [**Dv2009**] proved the finite fields Kakeya conjecture, as discussed in [**Ta2009b**, §1.1]). As such, the $k = 2$ result in fact holds over an arbitrary field, but we will stick to $\mathbf{R}$ for sake of notational consistency.

The polynomial method rests on two basic facts:

(1) Given a set $S$ of points in $\mathbf{R}^3$, there is a non-trivial polynomial $P : \mathbf{R}^3 \to \mathbf{R}$ of degree $O(|S|^{1/3})$ that vanishes at all of these points simultaneously.

(2) If a polynomial $P : \mathbf{R}^3 \to \mathbf{R}$ of degree at most $d$ vanishes on more than $d$ points of a line $\ell$, then it must in fact vanish on all of $\ell$.

Both facts are easy to prove (see e.g. [**Ta2009b**, §1.1]). But, as we will see, they have to be applied carefully in order to obtain a good estimate.

Suppose for sake of contradiction that we can find a set $S$ of points in $\mathbf{R}^3$ of cardinality $|S| \gg N^3$ such that each point in $S$ is incident to at least two lines from $\mathcal{L}$. The strategy is then to use Fact 1 find a low-degree polynomial $P$ that vanishes on $S$, so that $S$ is contained in the low-degree surface $\Sigma := \{x \in \mathbf{R}^3 : P(x) = 0\}$. This surface $\Sigma$ is not necessarily irreducible; it may be the union of several irreducible subsurfaces. By the nature of $S$, there will be lots of lines in $\mathcal{L}$ that intersect this surface $\Sigma$ quite frequently; by Fact 2, this should force the lines to be trapped inside $\Sigma$. Thus we have a low-degree surface $\Sigma$ that contains a lot of lines. Hopefully, this forces many of the irreducible components of $\Sigma$ to be singly ruled surfaces, doubly ruled surfaces, or planes. The latter two cases cannot contain too many lines in $\mathcal{L}$ by hypothesis, so we expect the dominant family of lines in $\mathcal{L}$ to be those coming from the singly ruled surfaces. But one can use Bezout-type theorems to control the number of times lines from one singly ruled surface of controlled degree can intersect lines from another such surface, and hopefully this gives the desired $O(N^3)$ bound for $P$.

Let's now execute this strategy. Fact 1 gives a non-trivial polynomial $P$ of degree $O(|S|^{1/3})$ that vanishes at all the points in $S$. This turns out to be

too large of a degree bound to close the argument. The problem is that the set of points $S$ that we want to apply this fact to is not completely arbitrary; in fact, by its nature, many points in $S$ are going to be collinear (because each point in $S$ is incident to at least two lines from $\mathcal{L}$), and so heuristically many of these points in $S$ are in fact "redundant" for the purposes of finding a polynomial that vanishes on all of $S$.

We can handle this by the "random refinement" trick. Let us write $|S| = QN^3$ for some $Q \ggg 1$. Each of these $QN^3$ points is incident to two lines in $\mathcal{L}$; as there are $\sim N^2$ lines in $\mathcal{L}$, we thus expect[15] each line in $\mathcal{L}$ to be incident to $\sim QN$ points in $S$.

Now, let us destroy some of this collinearity by taking a random subset $S'$ of $S$ of density about $1/Q$. Then $S'$ will have size about $N^3$, and each line in $\mathcal{L}$ is now expected to be incident to $O(N)$ elements of $S'$. We apply Fact 1 to get a non-trivial polynomial $P$ of degree $O(N)$ that vanishes at every point in $S'$. Applying Fact 2 (and assuming that certain constants were chosen properly), this implies that $P$ also vanishes at every line in $\mathcal{L}$, and thus also at every point in $S$. This should be contrasted with what would have happened if we applied Fact 1 to $S$ directly, giving a degree bound of $O(Q^{1/3}N)$ rather than $O(N)$.

So now we have a surface $\Sigma := \{x \in \mathbf{R}^3 : P(x) = 0\}$ of degree $O(N)$ that contains all the $\sim N^2$ lines in $\mathcal{L}$. We split this surface into irreducible components. These components may be planes, reguli, singly ruled surfaces, or not ruled at all. To simplify things let us just consider four extreme cases[16]:

(1) $\Sigma$ consists of the union of planes.

(2) $\Sigma$ consists of the union of reguli.

(3) $\Sigma$ consists of the union of singly ruled surfaces.

(4) $\Sigma$ consists of the union of non-ruled surfaces.

In the first case, degree considerations tell us that there are at most $O(N)$ planes, and by hypothesis each of them contains $O(N)$ lines in $\mathcal{L}$. Within each plane, there are at most $O(N^2)$ points of intersection, and between any two planes $\pi_1, \pi_2$, all points of intersection must lie in the common line $\pi_1 \cap \pi_2$, which is incident to the $O(N)$ lines of $\mathcal{L}$ in $\pi_1, \pi_2$ in at most $O(N)$

[15] Actually, what could happen more generally is that only a fraction of the lines in $\mathcal{L}$, say $\alpha N^2$ lines for some $0 < \alpha \leq 1$, are contributing the bulk of the incidences, with each line being incident to about $\sim QN/\alpha$ points in $S$; but let us consider the $\alpha = 1$ case here for simplicity, as it is the worst case, and the other cases are similar but require a bit more bookkeeping.

[16] For the full proof, one also has to consider hybrid cases in which more than one of the above four types appear, and so there are some cross-terms to deal with, but these are relatively easy to control and will not be discussed here.

points. Adding this all together we get at most $O(N^3)$ points of intersection, which is acceptable.

A similar argument (which we omit) handles the second case, so we move on to the third case. Let's assume for simplicity that each singly ruled surface has the same degree $d$; since the total degree of $\Sigma$ is $O(N)$, we must then have at most $O(N/d)$ such surfaces. In a singly ruled surface $\Gamma$, all but $O(1)$ of the lines in $\Gamma$ will come from the single ruling of $\Gamma$ (i.e. all but $O(1)$ of the lines will be generators); the contribution of these exceptional lines is negligible (as there at most $O(N)$ such lines in all, leading to $O(N^3)$ points of intersection) and we shall simply ignore them. There are two remaining types of intersection to consider; the intersections between two lines from the same ruling of a single ruled surface, and the intersections from lines that do not lie in the same ruled surface.

For the first type of intersection, one can show that in a ruled surface of degree $d$, any line in the ruling can intersect the other lines in the ruling in at most $O(d)$ points. So the total number of intersections arising in this way is $O(N^2 d)$; since $d = O(N)$, this is acceptable.

To handle the second type of intersection, observe that any line not incident to a ruled surface $\Gamma$ can intersect $\Gamma$ at most $O(d)$ times; summing over the $O(N/d)$ surfaces $\Gamma$ and the $O(N^2)$ lines in $\mathcal{L}$ we obtain at most $O(N^3)$ incidences, as desired.

Finally, we consider the fourth case, where there are no ruled surfaces anywhere in $\Sigma$. Here, one uses an algebraic observation of Cayley [**Sa1915**] that classifies ruled surfaces by a bounded number of algebraic conditions:

**Proposition 3.9.8.** *A surface $\Gamma$ is ruled if and only if, for every point $p$ in $\Gamma$, there exists a line $\ell_p$ through $p$ that is tangent to $\Gamma$ to order three (thus, if $P$ is a defining function for $\Gamma$, there exists a non-zero vector $v$ such that $P(p) = D_v P(p) = D_v^2 P(p) = D_v^3 P(p) = 0$).*

The "only if" direction is obvious, but what we need here is the more difficult "if" operation. The basic idea is to show (e.g. by using the *Picard uniqueness theorem* for ODE) that the foliation induced by the tangent lines $\ell_p$ consist entirely of straight lines, thus giving the desired ruling of $\Sigma$. The precise order of vanishing is not required for this argument; any bounded number instead of three would suffice here.

The condition that there exists a non-zero vector $v \in \mathbf{R}^3$ such that $D_v P(p) = D_v^2 P(p) = D_v^3 P(p) = 0$ can be combined by elementary elimination theory (or high school algebra, for that matter) into a single algebraic condition $FL(P)(p) = 0$ on $p$, where $FL(P)$ is a polynomial of degree $O(\deg(P))$ known as the *flecnode polynomial* of $P$. If $\Sigma := \{x : P(x) = 0\}$

contains no ruled surfaces, then the above proposition tells us that the flecnode polynomial $FL(P)$ shares no common factors with $P$.

On the other hand, if $\ell$ is a line in $\mathcal{L}$ oriented in the direction $v$, and $p$ is a point in $\mathcal{L}$, then $\ell$ lies in $\Sigma$, and thus $P(p) = D_v P(p) = D_v^2 P(p) = D_v^3 P(p) = 0$. Thus we see that each line $\ell$ in $\mathcal{L}$ is contained in the zero locus of both $P$ and its flecnode polynomial $FL(P)$. However, by applying Bezout's theorem to a generic two-dimensional slice of $\mathbf{R}^3$, we see that the zero locus of $P$ and of $FL(P)$ can only intersect in $O(\deg(P) \times \deg(FL(P))) = O(\deg(P)^2)$ lines at most. But if we chose the constants correctly, this number will be less than the number of lines in $\mathcal{L} \sim N^2$, leading to the required contradiction.

**3.9.3. The $k > 2$ case.** Now we turn to the $k > 2$ case. Our starting point here will be the *Szemerédi-Trotter theorem* [**SzTr1873**], which asserts that given a finite set $P$ of points and a finite set $L$ of lines, the number of incidences $|I(P,L)| := |\{(p,\ell) \in P \times L : p \in \ell\}|$ is bounded by

$$|I(P,L)| \ll |P|^{2/3}|L|^{2/3} + |P| + |L|.$$

The theorem is usually stated in the plane $\mathbf{R}^2$, but it automatically extends to higher dimensions, and in particular to $\mathbf{R}^3$, by a random projection argument. This theorem can be proven by crossing number methods (as discussed in [**Ta2008**, §2.10]) or by cell decomposition (as discussed in [**Ta2010b**, §1.6]). In both cases, the order structure of $\mathbf{R}$ (and in particular, the fact that a line divides a plane into two half-planes) is crucial. But we will not need to know the proof of this theorem, instead using it as a black box.

An easy corollary of this theorem is that if $L$ is a family of lines, $P$ is a family of points, and $2 \leq k \ll |L|^{1/2}$ is such that each point in $P$ is incident to at least $k$ points in $L$, then

$$|P| \ll |L|^2/k^3. \tag{3.33}$$

If one applies this bound directly to our situation, we obtain the bound $|P| \ll N^4/k^3$, which is inferior to the desired bound $|P| \ll N^3/k^2$. Actually, it is not surprising that we get an inferior bound, because we have not yet exploited the crucial non-coplanarity hypothesis.

However, it is possible to amplify the Szemerédi-Trotter bound by the method of *cell decomposition*. As discussed in [**Ta2010b**, §1.6], the idea is to carefully carve up the ambient space (which, in this case, is $\mathbf{R}^3$) into smaller regions or "cells", each of which only contains a small fraction of the points $P$ and which encounters only a small fraction of the lines $L$. One then applies an existing bound (in this case, (3.33)) to each cell, and sums up over all cells to get what should presumably be a better bound (where

the key point being that the cells should be constructed in such a way that each line only encounters a small fraction of the cells).

There is however a catch to this method: if one creates too many cells, then one starts running into the problem that too many points in $P$ and too many lines in $L$ will now lie on the *boundary* of the cell, rather than in the *interior*. So one has to carefully optimise the complexity of the cell decomposition.

In previous literature, the most popular way to create cells was by a random construction, for instance using randomly chosen points and lines from $P$ and $L$ to create planes, which one then uses to slice up space into polytope cells, possibly with an additional non-random "cleanup" stage to remove some degeneracies or other potential difficulties in the cell structure. One of the key innovations in the Guth-Katz paper is to instead create cells via the polynomial method, and specifically by the polynomial Ham Sandwich theorem. This allows for a very efficient and even cell decomposition; the walls of the cell will no longer be flat planes, but will still be algebraic sets of controlled degree, and this turns out to be good enough for the application at hand. This is inspired by previous applications of the polynomial ham sandwich theorem to incidence geometry problems, as discussed in [**Ta2009b**, §1.7].

Let us first recall the polynomial ham sandwich theorem (for three dimensions), which one can think of as a continuous version of Fact 1 from the previous section:

**Theorem 3.9.9** (Polynomial ham sandwich theorem)**.** *Let $X_1, \ldots, X_m$ be $m$ bounded open sets in $\mathbf{R}^3$. Then there exists a non-trivial polynomial $P : \mathbf{R}^3 \to \mathbf{R}$ of degree $O(m^{1/3})$ such that the algebraic set $\{x \in \mathbf{R}^3 : P(x) = 0\}$ bisects each of the $X_i$, thus $\{x \in X_i : P(x) < 0\}$ and $\{x \in X_i : P(x) > 0\}$ have volume equal to half that of $X_i$.*

See for instance [**Ta2009b**, §1.7] for a proof of this fact (based on the *Borsuk-Ulam theorem*).

By taking the $X_1, \ldots, X_m$ to be the $\varepsilon$-neighbourhood of finite sets $S_1, \ldots, S_m$, and sending $\varepsilon$ to zero (using some basic compactness properties of the (projective) space of polynomials to extract a limit), one can conclude a useful combinatorial corollary:

**Corollary 3.9.10** (Discretised polynomial ham sandwich theorem)**.** *Let $S_1, \ldots, S_m$ be finite sets of points in $\mathbf{R}^3$. Then there exists a non-trivial polynomial $P : \mathbf{R}^3 \to \mathbf{R}$ of degree $O(m^{1/3})$ such that the algebraic set $\{x \in \mathbf{R}^3 : P(x) = 0\}$ bisects each of the $S_i$, in the sense that $\{x \in S_i : P(x) < 0\}$ and $\{x \in S_i : P(x) > 0\}$ each have cardinality at most $|S_i|/2$.*

Note that the algebraic set $\{x \in \mathbf{R}^3 : P(x) = 0\}$ may capture some of the points of $S_i$; indeed, this is necessarily the case if $|S_i|$ is odd. This possibility will need to be addressed later in the argument.

We can iterate this corollary to obtain a nice cell decomposition:

**Corollary 3.9.11** (Cell decomposition)**.** *Let $S$ be a finite set of points in $\mathbf{R}^3$, and let $m \geq 1$. Then there exists a non-trivial polynomial $P$ of degree $O(m^{1/3})$ and a decomposition of $\{x \in \mathbf{R}^3 : P(x) \neq 0\}$ into at most $O(m)$ cells $C$, each of which is an open set with boundary in $\{x \in \mathbf{R}^3 : P(x) = 0\}$, and each of which contains at most $O(|S|/m)$ points of $S$. (We allow $C$ to be disconnected.)*

**Proof.** Without loss of generality we may take $m = 2^j$ to be a power of two. The proposition is trivial for $m = 1$, and by using Corollary 3.9.10 it is easy to see (with the right choice of implied constants) that the claim for $m$ implies the claim for $2m$ (with the same implied constants), by bisecting each of the cells obtained for $m$ by an additional bisecting polynomial that one then multiplies with the existing polynomial. □

Note that Fact 1 from the previous section can be viewed as the case $m = |S|$ of the above decomposition, in which the cell walls have now absorbed all the points in $S$. But for us, the optimal value of $m$ will be significantly less than $|S|$, to balance the contribution of the points on the cell walls with the points in the interior.

We now apply this situation to the situation in Theorem 3.9.7. Fix $3 \leq k \leq N$, and suppose for contradiction that we can find a set $S$ of points with $|S| \ggg N^3/k^2$, with each point in $S$ incident to at least $k$ lines in $\mathcal{L}$.

We will apply the cell decomposition for a certain parameter $m$; it turns out that the optimal value here is $m \sim (N/k)^3$. Thus we obtain a non-trivial polynomial $P$ of degree $O(N/k)$, and a collection of $O((N/k)^3)$ cells, each of which contains $O((k/N)^3|S|)$ points of $S$ in the interior.

By throwing away at most half of the points in $S$, we can end up in one of two extreme cases:

(1) (Cellular case) All points in $S$ are in the interior of a cell.

(2) (Algebraic case) All points in $S$ are in the boundary of a cell.

The cellular case becomes easier for $m$ large, while the algebraic case becomes easier for $m$ small; the choice $m \sim (N/k)^3$ is used to balance the two cases.

Let us first consider the cellular case. Then we see that $\gg (N/k)^3$ cells $C$ will contain $\gg (k/N)^3|S|$ points of $S$. Each such point in such a cell $C$ is

incident to at least $k$ lines from $\mathcal{L}$. Applying the Szemerédi-Trotter bound (3.33) inside this cell $C$, we conclude that

$$(k/N)^3|S| \ll |\mathcal{L}_C|^2/k^3,$$

where $\mathcal{L}_C$ is the set of lines in $\mathcal{L}$ that intersect $C$. Since $|S| \gg N^3/k^2$, we conclude that

$$|\mathcal{L}_C| \gg k^2.$$

On the other hand, as $P$ has degree $O(N/k)$, we see from Bezout's theorem that each line in $\mathcal{L}$ intersects $O(N/k)$ cells, and so

$$\sum_C |\mathcal{L}_C| \ll (N/k)|\mathcal{L}| \ll N^3/k.$$

Since the number of cells here is $\gg (N/k)^3$, we obtain

$$(N/k)^3 k^2 \lll N^3/k$$

which is a contradiction.

Now consider the algebraic case. We have $S$ points, each incident to $k$ lines from $\mathcal{L}$, which has cardinality $\sim N^2$. We thus expect each line to be incident to $\sim k|S|/N^2 \gg N/k$ points in $S$. For simplicity we will assume that every line is of this type (in reality, we would have to throw out some lines that do not intersect enough points in $S$).

By construction, all the points in $S$ lie in the algebraic set $\Sigma := \{x \in \mathbf{R}^3 : P(x) = 0\}$, which has degree $O(N/k)$. Applying Fact 2, we conclude that every line in $\mathcal{L}$ is in fact trapped inside $\Sigma$.

This is now a situation fairly similar to that of the previous section. However, we now use the hypothesis that $k \geq 3$ to obtain a better bound. Every point $p$ in $S$ has at least three lines of $\mathcal{L}$ passing through it, each of which lies in $\Sigma$. This leads to two possibilities for $p$:

(1) (Singular case) $p$ is a singular point of $\Sigma$ (i.e. $\nabla P(p) = 0$).

(2) (Flat case) $p$ is a flat point of $\Sigma$ (i.e. the $p$ is non-singular, but the second fundamental form of $\Sigma$ vanishes at $p$).

Indeed, if $p$ was non-singular, then $\Sigma$ has a unique tangent plane at $p$ along which at least three lines of $\mathcal{L}$ are tangent; as they all lie in $\Sigma$, this forces the second fundamental form of $\Sigma$ to vanish.

By throwing out at most half of the points in $S$, we may now reduce to two subcases:

(1) (Singular subcase) All points of $S$ are singular points of $\Sigma$.

(2) (Flat subcase) All points of $S$ are flat points of $\Sigma$.

Let us consider first the singular subcase. The points $S$ are now contained in the zero locus of $P$ and of $\nabla P$; the latter is an intersection of

three algebraic surfaces of degree $O(N/k)$; by reducing $P$ to be square-free, we can assume that $P$ and $\nabla P$ have no common factor. We already saw from Fact 2 that all the lines in $\mathcal{L}$ were trapped in the zero locus of $P$; the same argument shows that they are also trapped in the zero locus of $\nabla P$. But by Bezout's theorem applied to a generic two-dimensional slice of $\mathbf{R}^3$ (as was done in the previous section, using $FL(P)$ instead of $\nabla P$), we see that the zero locus of $P$ and $\nabla P$ can intersect in at most $O(N/k) \times O(N/k)$ lines, which will contradict the bound $|\mathcal{L}| \sim N^2$ if the constants are chosen properly.

Now we turn to the flat subcase. Just as the three polynomial components $\partial_{e_1} P, \partial_{e_2} P, \partial_{e_3} P$ of $\nabla P$ detects singular points of $\Sigma$, there are nine polynomials that detect flat points of $\Sigma$, namely the components $Q_{i,j}$ of the three-dimensional vector fields

$$Q_i := (D_{e_i \times \nabla P} \nabla P) \times \nabla P$$

for $i, j = 1, 2, 3$. All nine polynomials $Q_{i,j}$ vanish at a flat point. (This observation was also used in the preceding problem [**GuKa2010**].)

The argument for the singular subcase almost carries over without difficulty to the flat subcase, but there is one problem: it is not necessarily the case that $P$ does not share a common factor with the $Q_{i,j}$, so that we cannot quite apply the Bezout argument immediately. Indeed, $P$ could contain a plane, which of course consists entirely of flat points. However, if $P$ has no planes, then the set of flat points has positive codimension, and the argument proceeds as before. At the other extreme, if $P$ consists entirely of planes, then by degree considerations there are at most $O(N/k)$ planes present. But by hypothesis, each plane contains at most $O(N)$ lines from $\mathcal{L}$. Since $|\mathcal{L}| \sim N^2$, this leads to a contradiction (if the implied constants are chosen correctly). The general case (in which $P$ has some planes, but does not consist entirely of planes) can be established by combining the previous two arguments properly.

## 3.10. The Bourgain-Guth method for proving restriction theorems

One of my favourite unsolved problems in harmonic analysis is the *restriction problem*. This problem, first posed explicitly by Elias Stein (see e.g. [**St1979**]), can take many equivalent forms, but one of them is this: one starts with a smooth compact hypersurface $S$ (possibly with boundary) in $\mathbf{R}^d$, such as the unit sphere $S = S^2$ in $\mathbf{R}^3$, and equips it with surface measure $d\sigma$. One then takes a bounded measurable function $f \in L^\infty(S, d\sigma)$ on

this surface, and then computes the (inverse) Fourier transform

$$\widehat{fd\sigma}(x) = \int_S e^{2\pi i x\cdot\omega} f(\omega) d\sigma(\omega)$$

of the measure $fd\sigma$. As $f$ is bounded and $d\sigma$ is a finite measure, this is a bounded function on $\mathbf{R}^d$; from the dominated convergence theorem, it is also continuous. The restriction problem asks whether this Fourier transform also decays in space, and specifically whether $\widehat{fd\sigma}$ lies in[17] $L^q(\mathbf{R}^d)$ for some $q < \infty$. By the *closed graph theorem*, this is the case if and only if there is an estimate of the form

$$\|\widehat{fd\sigma}\|_{L^q(\mathbf{R}^d)} \leq C_{q,d,S} \|f\|_{L^\infty(S,d\sigma)} \tag{3.34}$$

for some constant $C_{q,d,S}$ that can depend on $q, d, S$ but not on $f$. By a limiting argument, to provide such an estimate, it suffices to prove such an estimate under the additional assumption that $f$ is smooth.

**Remark 3.10.1.** Strictly speaking, the above problem should be called the *extension problem*, but it is dual to the original formulation of the *restriction problem*, which asks to find those exponents $1 \leq q' \leq \infty$ for which the Fourier transform of an $L^{q'}(\mathbf{R}^d)$ function $g$ can be meaningfully restricted to a hypersurface $S$, in the sense that the map $g \mapsto \hat{g}|_S$ can be continuously defined from $L^{q'}(\mathbf{R}^d)$ to, say, $L^1(S, d\sigma)$. A duality argument shows that the exponents $q'$ for which the restriction property holds are the dual exponents to the exponents $q$ for which the extension problem holds.

There are several motivations for studying the restriction problem[18]. The problem is connected to the classical question of determining the nature of the convergence of various Fourier summation methods (and specifically, Bochner-Riesz summation); very roughly speaking, if one wishes to perform a partial Fourier transform by restricting the frequencies (possibly using a well-chosen weight) to some region $B$ (such as a ball), then one expects this operation to well behaved if the boundary $\partial B$ of this region has good restriction (or extension) properties. More generally, the restriction problem for a surface $S$ is connected to the behaviour of Fourier multipliers whose symbols are singular at $S$. The problem is also connected to the analysis of various linear PDE such as the Helmholtz equation, Schrödinger equation, wave equation, and the (linearised) Korteweg-de Vries equation, because solutions to such equations can be expressed via the Fourier transform in the form $fd\sigma$ for various surfaces $S$ (the sphere, paraboloid, light cone, and cubic for the Helmholtz, Schrödinger, wave, and linearised Korteweg de Vries equation respectively). A particular family of restriction-type theorems for

[17] This is a natural space to control decay because it is translation invariant, which is compatible on the frequency space side with the modulation invariance of $L^\infty(S, d\sigma)$.

[18] For a further discussion of these topics, see [**Ta2003b**].

such surfaces, known as *Strichartz estimates*, play a foundational role in the nonlinear perturbations of these linear equations (e.g. the nonlinear Schrödinger equation, the nonlinear wave equation, and the Korteweg-de Vries equation). Last, but not least, there is a a fundamental connection between the restriction problem and the *Kakeya problem*, which roughly speaking concerns how tubes that point in different directions can overlap. Indeed, by superimposing special functions of the type $\widehat{f d\sigma}$, known as *wave packets*, and which are concentrated on tubes in various directions, one can "encode" the Kakeya problem inside the restriction problem; in particular, the conjectured solution to the restriction problem implies the conjectured solution to the Kakeya problem. Finally, the restriction problem serves as a simplified toy model for studying discrete exponential sums whose coefficients do not have a well controlled phase; this perspective was, for instance, used in [**Gr2005**] to establish Roth's theorem in the primes by Fourier-analytic methods, which was in turn one of the main inspirations for our later work establishing arbitrarily long progressions in the primes, although we ended up using ergodic-theoretic arguments instead of Fourier-analytic ones and so did not directly use restriction theory in that paper.

The estimate (3.34) is trivial for $q = \infty$ and becomes harder for smaller $q$. The geometry, and more precisely the *curvature*, of the surface $S$, plays a key role: if $S$ contains a portion which is completely flat, then it is not difficult to concoct an $f$ for which $\widehat{f d\sigma}$ fails to decay in the normal direction to this flat portion, and so there are no restriction estimates for any finite $q$. Conversely, if $S$ is not infinitely flat at any point, then from the method of stationary phase, the Fourier transform $\widehat{d\sigma}$ can be shown to decay at a power rate at infinity, and this together with a standard method known as the $TT^*$ *argument* can be used to give non-trivial restriction estimates for finite $q$. However, these arguments fall somewhat short of obtaining the best possible exponents $q$. For instance, in the case of the sphere $S = S^{d-1} \subset \mathbf{R}^d$, the Fourier transform $\widehat{d\sigma}(x)$ is known to decay at the rate $O(|x|^{-(d-1)/2})$ and no better as $d \to \infty$, which shows that the condition $q > \frac{2d}{d-1}$ is necessary in order for (3.34) to hold for this surface. The *restriction conjecture for* $S^{d-1}$ asserts that this necessary condition is also sufficient. However, the $TT^*$-based argument gives only the *Tomas-Stein theorem* [**To1975**], which in this context gives (3.34) in the weaker range[19] $q \geq \frac{2(d+1)}{d-1}$.

Over the last two decades, there was a fair amount of work in pushing past the Tomas-Stein barrier. For sake of concreteness let us work just with

[19] On the other hand, by the nature of the $TT^*$ method, the Tomas-Stein theorem does allow the $L^\infty(S, d\sigma)$ norm on the right-hand side to be relaxed to $L^2(S, d\sigma)$, at which point the Tomas-Stein exponent $\frac{2(d+1)}{d-1}$ becomes best possible. The fact that the Tomas-Stein theorem has an $L^2$ norm on the right-hand side is particularly valuable for applications to PDE, leading in particular to the Strichartz estimates mentioned earlier.

the restriction problem for the unit sphere $S^2$ in $\mathbf{R}^3$. Here, the restriction conjecture asserts that (3.34) holds for all $q > 3$, while the Tomas-Stein theorem gives only $q \geq 4$. By combining a multiscale analysis approach with some new progress on the Kakeya conjecture, Bourgain [**Bo1991**] was able to obtain the first improvement on this range, establishing the restriction conjecture for $q > 4 - \frac{2}{15}$. The methods were steadily refined over the years; until recently, the best result [**Ta2003**] was that the conjecture held for all $q > 3\frac{1}{3}$, which proceeded by analysing a "bilinear $L^2$" variant of the problem studied previously by Bourgain [**Bo1995**] and by Wolff [**Wo2001**]. This is essentially the limit of that method; the relevant bilinear $L^2$ estimate fails for $q < 3 + \frac{1}{3}$.

On the other hand, the full range $q > 3$ of exponents in (3.34) was obtained by Bennett, Carbery, and myself [**BeCaTa2006**] (with an alternate proof later given by Guth [**Gu2010**]), but only under the additional assumption of *non-coplanar interactions*. In three dimensions, this assumption was enforced by replacing (3.34) with the weaker trilinear (and localised) variant

(3.35)
$$\|\widehat{f_1 d\sigma_1}\widehat{f_2 d\sigma_2}\widehat{f_3 d\sigma_3}\|_{L^{q/3}(B(0,R))} \leq C_{q,d,S_1,S_2,S_3,\varepsilon} R^{\varepsilon} \|f_1\|_{L^\infty(S_1,d\sigma_1)} \|f_2\|_{L^\infty(S_2,d\sigma_2)} \|f_3\|_{L^\infty(S_3,d\sigma_3)}$$

where $\varepsilon > 0$ and $R \geq 1$ are arbitrary, $B(0,R)$ is the ball of radius $R$ in $\mathbf{R}^3$, and $S_1, S_2, S_3$ are compact portions of $S$ whose unit normals $n_1(), n_2(), n_3()$ are never coplanar, thus there is a uniform lower bound

$$|n_1(\omega_1) \wedge n_2(\omega_2) \wedge n_3(\omega_3)| \geq c$$

for some $c > 0$ and all $\omega_1 \in S_1, \omega_2 \in S_2, \omega_3 \in S_3$. If it were not for this non-coplanarity restriction, (3.35) would be equivalent to (3.34) (by setting $S_1 = S_2 = S_3$ and $f_1 = f_2 = f_3$, with the converse implication coming from Hölder's inequality; the $R^\varepsilon$ loss can be removed by a lemma from [**Ta1999**]). At the time we wrote this paper, we tried fairly hard to try to remove this non-coplanarity restriction in order to recover progress on the original restriction conjecture, but without much success.

Very recently, though, Bourgain and Guth [**BoGu2010**] found a new way to use multiscale analysis to "interpolate" between the result of Bennett, Carbery and myself (that has optimal exponents, but requires non-coplanar interactions), with a more classical square function estimate of Córdoba that handles the coplanar case. A direct application of this interpolation method already ties with the previous best known result in three dimensions (i.e. that (3.34) holds for $q > 3\frac{1}{3}$). But it also allows for the insertion of additional input, such as the best Kakeya estimate currently known in three dimensions, due to Wolff [**Wo1995**]. This enlarges the range slightly to $q > 3.3$. The method also can extend to variable-coefficient settings, and in

some of these cases (where there is so much "compression" going on that no additional Kakeya estimates are available) the estimates are best possible.

As is often the case in this field, there is a lot of technical book-keeping and juggling of parameters in the formal arguments of Bourgain and Guth, but the main ideas and numerology[20] can be expressed fairly readily. In this section, I would like to record this numerology for the simplest of the Bourgain-Guth arguments, namely a reproof of (3.34) for $p > 3\frac{1}{3}$.

In order to focus on the ideas in the paper (rather than on the technical details), I will adopt an informal, heuristic approach, for instance by interpreting the uncertainty principle and the pigeonhole principle rather liberally, and by focusing on main terms in a decomposition and ignoring secondary terms. I will also be somewhat vague with regard to asymptotic notation such as $\ll$. Making the arguments rigorous requires a certain amount of standard but tedious effort (and is one of the main reasons why the Bourgain-Guth paper is as long as it is), which I will not focus on here.

**3.10.1. The Córdoba square function estimate.** In two dimensions, the restriction theory is well understood, due to the work of Córdoba, Fefferman, and others (see [**Co1985**] for a survey). The situation is particularly simple when one looks at bilinear expressions such as

$$\|F_1 F_2\|_{L^2(\mathbf{R}^2)}$$

where $F_1 := \widehat{f_1 d\sigma_1}$, $F_2 := \widehat{f_2 d\sigma_2}$, and $d\sigma_1, d\sigma_2$ are surface measures on two smooth compact curves $S_1, S_2$ that are *transverse* in the sense that the unit normals of $S_1$ are never[21] oriented in the same direction as the unit normals of $S_2$. In this case, we can use Plancherel's theorem to rewrite the above expression as a convolution

$$\|f_1 d\sigma_1 * f_2 d\sigma_2\|_{L^2(\mathbf{R}^2)}.$$

The transversality of $S_1$ and $S_2$, combined with the inverse function theorem, shows that $f_1 d\sigma_1 * f_2 d\sigma_2$ is a non-degenerate pushforward of the tensor product $f_1 \otimes f_2$, and so one obtains the basic bilinear restriction estimate

$$\|F_1 F_2\|_{L^2(\mathbf{R}^2)} \ll \|f_1\|_{L^2(S_1, d\sigma_1)} \|f_2\|_{L^2(S_2, d\sigma_2)}.$$

This estimate (and higher-dimensional analogues thereof) lead to the *bilinear* $X^{s,b}$ *estimates* which are of fundamental importance in nonlinear dispersive equations (particularly those in which the nonlinearity contains derivatives).

[20] In mathematics, *numerology* refers to the empirically observed relationships between various key exponents and other numerical parameters; in many cases, one can use shortcuts such as dimensional analysis or informal heuristic, to compute these exponents long before the formal argument is completely in place.

[21] A model case to consider here are two arcs of the unit circle, one near $(1, 0)$ and one near $(0, 1)$.

This bilinear estimate can be localised. Suppose one splits $S_1$ into arcs $S_{1,\alpha}$ of diameter $\sim 1/r$ for some $r \gg 1$, which induces a decomposition $F_1 = \sum_\alpha F_{1,\alpha}$ of $F_1$ into components $F_{1,\alpha} := f_1 \widehat{1_{S_{1,\alpha}} d\sigma_1}$. Similarly decompose $F_2 = \sum_\beta F_{2,\beta}$. Then we have

$$F_1 F_2 = \sum_\alpha \sum_\beta F_{1,\alpha} F_{2,\beta}.$$

The Fourier transform of $F_{1,\alpha} F_{2,\beta}$ is supported in the Minkowski sum $S_{1,\alpha} + S_{2,\beta}$. The transversality of $S_1, S_2$ ensures that these sums are basically disjoint as $\alpha, \beta$ varies, so by almost orthogonality one has

$$\|F_1 F_2\|_{L^2(\mathbf{R}^2)} \ll (\sum_\alpha \sum_\beta \|F_{1,\alpha} F_{2,\beta}\|^2_{L^2(\mathbf{R}^2)})^{1/2}$$

or equivalently

$$\|F_1 F_2\|_{L^2(\mathbf{R}^2)} \ll \|(\sum_\alpha |F_{1,\alpha}|^2)^{1/2} (\sum_\beta |F_{2,\beta}|^2)^{1/2}\|_{L^2(\mathbf{R}^2)}.$$

Actually, this estimate is morally localisable to balls $B(x,r)$ of radius $r$; heuristically, we have

$$\|F_1 F_2\|_{L^2(B(x,r))} \ll \|(\sum_\alpha |F_{1,\alpha}|^2)^{1/2}) (\sum_\beta |F_{2,\beta}|^2)^{1/2}\|_{L^2(B(x,r))}. \tag{3.36}$$

Informally[22], this is due to the uncertainty principle: localising in space to scale $r$ wouuld cause the arcs $S_{1,\alpha}, S_{2,\beta}$ in Fourier space to blur out at the scale $1/r$, but this will not significantly affect the almost disjointness of the Minkowski sums $S_{1,\alpha} + S_{2,\beta}$.

Furthermore, the uncertainty principle suggests to us that $F_{1,\alpha}$ and $F_{2,\beta}$ are essentially constant on balls $B(x,r)$ of radius $r$. As such, the expression inside the norm on the right-hand side of (3.36) is morally constant on such balls, which allows us to apply Hölder's inequality and conclude that

$$\|F_1 F_2\|_{L^q(B(x,r))} \ll \|(\sum_\alpha |F_{1,\alpha}|^2)^{1/2}) (\sum_\beta |F_{2,\beta}|^2)^{1/2}\|_{L^q(B(x,r))} \tag{3.37}$$

for any $q \leq 2$.

This is a bilinear estimate, but for heuristic purposes it is morally equivalent to the linear estimate

$$\|F\|_{L^q(B(x,r))} \ll \|(\sum_\alpha |F_\alpha|^2)^{1/2})\|_{L^q(B(x,r))} \tag{3.38}$$

[22] To make this rigorous, one would use a smoother cutoff than $1_{B(x,r)}$, and in particular it is convenient to use a cutoff which is compactly supported in Fourier space rather than physical space; we will not discuss these technicalities further here.

for $q \leq 4$, where $F = \widehat{f d\sigma}$ and $d\sigma$ is the surface measure on a curve $S$ which "exhibits curvature" and such that $F$ is "dominated by transverse interactions", $F_\alpha = \widehat{f 1_{S_\alpha} d\sigma}$, and $S$ is partitioned into arcs $S_\alpha$ of diameter $\sim 1/r$. For the purposes of numerology, we will pretend that (3.38) is true as stated, though in practice one has to actually work with the bilinearisation (3.37) instead.

**Remark 3.10.2.** Córdoba [**Co1982**] used (a rigorous form of) (3.38) to establish the restriction conjecture (3.34) for curves in the plane (such as the unit circle) in the optimal range $q > 4$.

The estimate (3.38) is a two-dimensional one, but it can be stepped up to a three-dimensional estimate

$$\|F\|_{L^q(B(x,r))} \ll \|(\sum_\alpha |F_\alpha|^2)^{1/2})\|_{L^q(B(x,r))} \tag{3.39}$$

for $q \leq 4$, where $F = \widehat{f d\sigma}$, $d\sigma$ is now surface measure on the sphere $S^2 \subset \mathbf{R}^3$, which one decomposes into caps $S_\alpha$ of diameter $O(1/r)$, $f$ is supported on the $O(1/r)$-neighbourhood of a great circle in $S^2$ with $F_\alpha := \widehat{f 1_{S_\alpha} d\sigma}$, and $F$ is "dominated by transverse interactions" in a sense that we will not quantify precisely here. This gives efficient control on $F$ in terms of square functions, but only in the "transverse coplanar case" in which the frequencies that dominate $F$ are both coplanar (in the sense that they all lie roughly on the same great circle) and transverse.

**3.10.2. The Bourgain-Guth argument.** Now we sketch how the Bourgain-Guth argument works to establish (3.34) for $q > \frac{10}{3}$. Fix $q$; we may assume $q < 4$. For each radius $R \geq 1$, let $Q_R$ be the best constant in the local restriction estimate

$$\|F\|_{L^q(B(x,R))} \leq Q_R \|f\|_{L^\infty(S^2)}$$

where $F := \widehat{f d\sigma}$. To show (3.34), one has to show that $Q_R$ is bounded uniformly in $R$. Actually, thanks to an "epsilon removal lemma" that was proven in [**Ta1999**] using a variant of the Tomas-Stein argument, it suffices to show that the logarithmic growth estimate $Q_R \ll R^\varepsilon$ for any $\varepsilon > 0$.

An effective technique for achieving this is an *induction on scales* argument, bounding $Q_R$ efficiently in terms of $Q_R$ for various scales $R'$ between 1 and $R$. This technique was introduced in [**Bo1991**], using the intermediate scale $R' := \sqrt{R}$ (which is a natural scale for the purposes of approximating spherical caps by disks while still respecting the uncertainty principle). The subsequent paper [**Wo2001**] adapted this argument by also relying on scales $R' = R^{1-\varepsilon}$ that were much closer to $R$. The Bourgain-Guth argument is closer in spirit to this latter approach.

Specifically, one sets $K := R^\varepsilon$ to be a small power of $R$, and divides the sphere $S^2$ into largish caps $S_\alpha$ of radius $\sim 1/K$, thus splitting $F = \sum_\alpha F_\alpha$. At the same time, we cover $B(x,R)$ by smallish balls $B(y,K)$ of radius $K$. On each such ball $B(y,K)$, the functions $F_\alpha$ are morally constant, as per the uncertainty principle. Of course, the amplitude of the $F_\alpha$ on $B(y,K)$ depend on $\alpha$; for each small ball $B(y,K)$, only a fraction of the $F_\alpha$ will "dominate" the sum $F$. Roughly speaking, we can then sort the balls $B(y,K)$ into three classes:

(1) (Non-coplanar case) There exist three dominant caps $S_\alpha$ which do not lie within $O(1/K)$ of a great circle.

(2) (Non-transverse case) All the dominant caps $S_\alpha$ lie in a cap of size $o(1)$.

(3) (Transverse coplanar case) All the dominant caps lie within $O(1/K)$ of a great circle, but at least two of them are at distance $\sim 1$ from each other.

In the first case, one can control $\|F\|_{L^q(B(y,K))}$ by $O(K^{O(1)})$ non-coplanar interactions of the form $\|F_1F_2F_3\|_{L^{q/3}(B(y,K))}$, where $F_1, F_2, F_3$ are portions of $F$ on non-coplanar portions of the sphere $S^2$. In this case, one can use (3.35) and obtain a contribution of $O(K^{O(1)}) = O(R^{O(\varepsilon)})$ in this case.

It has been known for some time [**TaVaVe1998**] that the non-transverse case can always be eliminated. Basically, if we group the caps $S_\alpha$ into larger caps $\tilde{S}_\beta$ of radius $1/K' = o(1)$, and decompose $F = \sum_\beta \tilde{F}_\beta$ accordingly, then in the non-transverse case we can morally bound

$$|F| \ll (\sum_\beta |\tilde{F}_\beta|^q)^{1/q}$$

and so

$$\|F\|_{L^q(B(x,R))} \ll (\sum_\beta \|\tilde{F}_\beta\|^q_{L^q(B(x,R))})^{1/q}.$$

However, a standard parabolic rescaling argument (which, strictly speaking, requires one to generalise the sphere to a larger family of similarly curved surfaces, but let us ignore this technical detail) shows that

$$\|\tilde{F}_\beta\|_{L^q(B(x,R))} \ll Q_{R/K'}(K')^{4/q-2}$$

and so (since there are $\sim (K')^2$ large caps $\tilde{S}_\beta$)

$$\|F\|_{L^q(B(x,R))} \ll (K')^{6/q-2} Q_{R/K'}.$$

Since $q > 3$, the exponent of $K'$ here is positive, and so this is a good term for the recurrence.

Finally, we deal with the transverse, coplanar case. Here, the main tool is the Córdoba-type square function estimate (3.39). Being coplanar, there

are only about $O(K)$ caps $S_\alpha$ that contribute here, so we can pay a factor of $O(K^{1/2-1/q})$ and convert the square function to a $\ell^q$-function:

$$\|F\|_{L^q(B(y,K))} \ll K^{1/2-1/q} \|(\sum_\alpha |F_\alpha|^q)^{1/q})\|_{L^q(B(y,K))}.$$

Summing over all such balls, we obtain

$$\|F\|_{L^q(B(x,R))} \ll K^{1/2-1/q} \sum_\alpha \|F_\alpha\|^q_{L^q(B(x,R))})^{1/q}).$$

Again, a parabolic rescaling gives

$$\|F_\alpha\|_{L^q(B(x,R))} \ll K^{4/q-2} Q_{R/K}$$

so the net contribution to $\|F\|_{L^q(B(x,R))}$ here is $O(K^{1/2-1/q} K^{6/q-2} Q_{R/K})$. This leads to the recursion

$$Q_R \ll R^{O(\varepsilon)} + (K')^{6/q-2} Q_{R/K'} + K^{1/2-1/q} K^{6/q-2} Q_{R/K}.$$

For $q > 10/3$, the exponents of $K'$ and $K$ are negative, and this allows one to induct on scales and get the required bound $Q_R \ll R^{O(\varepsilon)}$.

The argument given above is not optimal; the main inefficiency here is the factor of $O(K^{1/2-1/q})$ that one pays to convert the square function to the $\ell^q$ function. This factor is only truly present if almost every cap $S_\alpha$ along a great circle is genuinely contributing to $F$. However, one can use Kakeya estimates to prevent this event from happening too often. Indeed, thanks to the nature of parabolic scaling, the functions $F_\alpha$ are not merely essentially constant on balls of radius $K$, but are in fact essentially constant on $K \times K^2$ tubes oriented in the normal direction of $S_\alpha$. One can use a Kakeya estimate (such as the one in [**Wo1995**]) to then prevent these tubes from congregating too often with too high of a multiplicity; quantifying this, Bourgain and Guth were able to relax the constraint $q > 10/3$ to $q > 3.3$. Unfortunately, there are still some residual inefficiencies, and even with the full Kakeya conjecture, the argument given in that paper only gets down to $3\frac{3}{11}$.

*Chapter 4*

# Nonstandard analysis

## 4.1. Real numbers, nonstandard real numbers, and finite precision arithmetic

The *real number system* can be thought of as the idealised infinite precision limit of finite precision arithmetic. In finite precision arithmetic, every quantity that cannot be expressed exactly (e.g. as a terminating decimal) will instead come with error bars; for instance, a computation of $\pi$ in finite precision arithmetic may yield the estimate $\pi = 3.14 \pm 0.01$, and then a later refinement of that computation may yield the improvement $\pi = 3.141 \pm 0.001$. Thus a transcendental quantity such as $\pi$ would not be expressed as any single finite precision number; instead, $\pi$ would be viewed as the idealisation (or limit) of a class of mutually consistent finite precision measurements ($3.14 \pm 0.01$, $3.141 \pm 0.001$, etc.).

With this perspective, the assertion that two real numbers $x$, $y$ are equal means that they are equal to any given finite precision. Thus, for instance, the real numbers $x = 0.999\ldots$ and $y = 1.000\ldots$ are equal, because they are both equal to $1 \pm 0.1$, both equal to $1 \pm 0.01$, and so forth. At a more formal level, this is essentially just the standard construction of the real numbers as the space of Cauchy sequences of rationals, quotiented out by equivalence.

In particular, a real number $x$ is zero if and only if it is indistinguishable from zero to any finite amount of precision, i.e. one has $|x| < 0.1$, $|x| < 0.01$, and so forth. This is essentially the *Archimedean principle*. More generally, if one wishes to prove that $x = y$, it suffices to show that $x = y + O(\varepsilon)$ for every $\varepsilon > 0$. This trick of giving oneself an *epsilon of room* is a fundamental technique in analysis; see [**Ta2010**].

Among other things, this explains why, in probability theory, an event can have zero probability without being impossible. For instance, let $x$ be a real number chosen uniformly at random from the interval $[0,1]$, and consider the event $E$ that $x$ equals 0.5. This is contained in the event that $x$ lies in the subinterval $[0.45, 0.55]$, which occurs with probability 0.1; thus $E$ has probability at most 0.1. Similarly, as 0.5 is contained in $[0.495, 0.505]$, we see that $E$ occurs with probability at most 0.01. More generally, the probability of $E$ cannot be distinguished from zero to any finite precision, and so $E$ occurs with probability zero. Nevertheless, the event $E$ is of course not totally empty; it is conceivable (but arbitrarily unlikely) that the randomly chosen $x$ will indeed equal 0.5 exactly.

There are more precise number systems than the reals, most notably the *hyperreals* (or *nonstandard reals* ${}^*\mathbf{R}$), which do not necessarily identify two numbers if they agree to any finite accuracy (i.e. they no longer obey the Archimedean principle[1]); in particular such number systems admit non-trivial *infinitesimals*, whose magnitude is less than any standard positive real number such as 0.1 or 0.01, but which is not zero. For instance, whereas in the standard reals, the sequences $1/n$ and $1/n^2$ both go to zero as $n \to \infty$, in nonstandard analysis, if one takes an *ultralimit* as $n$ goes to an unbounded nonstandard natural number $N$, $1/n$ and $1/n^2$ will now converge to two different limits $1/N$, $1/N^2$, which are both infinitesimal, but distinct from each other and from zero (in particular, $1/N^2$ will be infinitesimal compared to $1/N$). So we see that ultralimits not only capture the classical limiting value of a sequence, but also retain information about the rate and nature of convergence. But this comes at the cost of losing the (standard) Archimedean principle (at least if one uses the standard definition of a natural number).

The relationship between standard numbers and nonstandard numbers in nonstandard analysis is somewhat analogous to that between *single-precision* and *double-precision* numbers in computer programming. In modern computer languages such as *C*, integers often come in two[2] forms: a "short" integer, often using just two bytes of storage (and thus only able to represent numbers up to $2^{16} - 1 = 65535$ in size), and a "long" integer, often using four bytes of storage (and thus able to represent numbers as large as $2^{32} - 1 = 4,294,967,295$). ) These are analogous to standard integers and nonstandard integers; for instance, the analogue of an unbounded nonstandard integer is a long integer which is too large to be represented correctly as a short integer. The analogy is not quite perfect, because short integers

[1] More precisely, they do not obey the *standard* Archimedean principle - there exist positive nonstandard reals less than $1/n$ for every standard natural number $n$; but they still obey the *nonstandard* Archimedean principle, in which $n$ now ranges over the *nonstandard* natural numbers rather than the standard ones.

[2] Let us ignore for now the distinction between signed and unsigned integers.

are not closed under basic operations such as addition; it is possible for the sum (say) of two short integers to cause an overflow and give an incorrect result (or an error message). In contrast, the standard integers are closed under all standard operations (e.g. addition, multiplication, exponentiation, the Ackermann function, etc.); one cannot ever reach a truly unbounded nonstandard integer just by using standard operations on standard integers.

Similarly, in computer languages, real numbers are modeled by single-precision floating point numbers, which have a certain maximum size and a certain number of digits (or bits) of precision, and then double-precision numbers, which can be significantly larger and have many more bits of precision. These are analogous to standard reals and nonstandard reals respectively. A bounded nonstandard real is then like a floating point number with additional digits of precision; the operation of taking the standard part of such a number is analogous to rounding off a double-precision number to a single-precision one. Again, the analogy is not perfect: in computer programming, such roundoff can create roundoff errors, but in nonstandard analysis the map $x \to \mathrm{st}(x)$ is a homomorphism and produces no such errors.

In a similar spirit, an infinitesimal is analogous to a double-precision number which would round to 0 if represented in single-precision. Note that in double precision one can create numbers such as $0.9999\ldots$ which would round to 1 in single-precision, which can be viewed as one way of interpreting the standard real identity $0.9999\ldots = 1$.

Various continuous operations in standard analysis can be interpreted as discrete operations in nonstandard analysis. For instance, suppose one wanted to integrate a standard continuous function $f : [0,1] \to \mathbf{R}$ from 0 to 1. One way to do it is to extend $f$ to the nonstandard numbers in the usual fashion, form the Riemann sum $\frac{1}{N}\sum_{n=1}^{N} f(n/N)$ for some unbounded natural number $N$, and take the standard part; the Riemann integrability of $f$ ensures that this will give the standard integral $\int_0^1 f(x)\ dx$. One can think of the computational analogue of this, namely numerical integration of a continuous function using an extremely small step size, and rounding off all errors below the single-precision level. In numerical analysis, the accumulated numerical error in such procedures will sometimes be visible at the single-precision level; but this does not happen in nonstandard analysis (ultimately due to the properties of ultrafilters used in the construction of the nonstandard reals).

## 4.2. Nonstandard analysis as algebraic analysis

One of the features of *nonstandard analysis*, as opposed to its standard counterpart, is that it efficiently conceals almost all of the epsilons and deltas that are so prevalent in standard analysis (cf. [**Ta2008**, §1.5]). As a

consequence, analysis acquires a much more *algebraic* flavour when viewed through the nonstandard lens.

Consider for instance the concept of continuity for a function $f : [0,1] \to \mathbf{R}$ on the unit interval. The standard definition of continuity uses a bunch of epsilons and deltas:

- $f$ is continuous iff for every $\varepsilon > 0$ and $x \in [0,1]$ there exists $\delta > 0$ such that for all $y$ in $[0,1]$ with $|y-x| < \delta$ one has $|f(y)-f(x)| < \varepsilon$.

The nonstandard definition, which is logically equivalent, is as follows:

- $f$ is continuous iff for every $x \in [0,1]$ and $y \in {}^*[0,1]$ with $y = x + o(1)$, one has ${}^*f(y) = {}^*f(x) + o(1)$.

Here ${}^*f : {}^*[0,1] \to {}^*\mathbf{R}$ is the ultralimit extension[3] of the original function $f : [0,1] \to \mathbf{R}$ to the ultrapowers ${}^*[0,1]$ and ${}^*\mathbf{R}$, and $o(1)$ denotes an infinitesimal, i.e. a nonstandard real whose magnitude is smaller than any standard $\varepsilon > 0$. It is a good exercise to test one's understanding of nonstandard analysis to verify that these two definitions are indeed equivalent.

There is an analogous nonstandard characterisation of uniform continuity:

- $f$ is uniformly continuous iff for every $x \in {}^*[0,1]$ and $y \in {}^*[0,1]$ with $y = x + o(1)$, one has ${}^*f(y) = {}^*f(x) + o(1)$.

One can now quickly give a nonstandard proof[4] of the classical fact that continuous functions on a compact interval such as $[0,1]$ are automatically uniformly continuous. Indeed, if $x, y \in {}^*[0,1]$ are such that $y = x + o(1)$, then $x$ and $y$ have the same *standard part* $z = \mathrm{st}(x) = \mathrm{st}(y)$, which lies in $[0,1]$. If $f$ is continuous, then $f(y) = f(z) + o(1)$ and $f(x) = f(z) + o(1)$, hence $f(y) - f(x) = o(1)$.

One can also use nonstandard analysis to phrase continuity (on compact domains, at least) in a very succinct algebraic fashion:

- $f$ is continuous if and only if ${}^*f$ commutes with the standard part function $\mathrm{st} : x \to \mathrm{st}(x)$.

Note how the number of quantifiers required to define continuity has decreased[5] all the way to zero. It is this elimination of quantifiers that allows the theory to be algebrised; as a zeroth approximation, one can view algebra as the mathematics of quantifier-free statements.

---

[3] See [**Ta2008**, §1.5] or Section 4.4 for definitions of these terms and more discussion.

[4] It is instructive to see how the nonstandard proof and the standard proof are ultimately just reformulations of each other, and in particular how both rely ultimately on the Bolzano-Weierstrass theorem. In the nonstandard world, Bolzano-Weierstrass is needed to demonstrate existence of the standard part.

[5] Of course, they have all been hidden in the definition of the standard part function.

## 4.3. Compactness and contradiction: the correspondence principle in ergodic theory

The *correspondence principle* between finite dynamical systems[6] and infinite dynamical systems, that allows one to convert certain statements about the former to logically equivalent statements about the latter. The most well-known instance of this principle is the *Furstenberg correspondence principle* that connects combinatorial statements about large subsets of integers with ergodic theoretic statements about large subsets of measure-preserving systems, but the principle is more general than this, as we hope to demonstrate in this section.

Informally, the correspondence principle equates four types of results:

(1) Local quantitative results for concrete finite systems.

(2) Local qualitative results for concrete infinite systems.

(3) Continuous quantitative results for abstract finite systems.

(4) Continuous qualitative results for abstract infinite systems.

The meaning of these terms should become clearer once we give some specific examples.

There are many contexts in which this principle shows up (e.g. in Ramsey theory, recurrence theory, graph theory, group theory, etc.) but the basic ingredients are always the same. Namely, the correspondence between Type 1 and Type 2 (or Type 3 and Type 4) arises from a weak sequential compactness property, which, roughly speaking asserts that given any sequence of (increasingly large) finite systems, there exists a subsequence of such systems which converges (in a suitably "weak" sense) to an infinite system. (We will define these terms more precisely in concrete situations later.) More informally, any "sufficiently large" finite system can be "approximated" in some weak sense by an infinite system[7]

By combining compactness with a proof by contradiction argument, one obtains a *compactness and contradiction argument* that yields a correspondence principle: any qualitative statement about infinite systems (e.g. that a certain quantity is always strictly positive) is equivalent to a quantitative statement about sufficiently large finite systems (e.g. a certain quantity is always uniformly bounded from below). This principle forms a crucial bridge between finitary (or quantitative) mathematics and infinitary (or qualitative) mathematics; in particular, by taking advantage of this principle, tools

[6] I will be vague here about what "dynamical system" means; very broadly, just about anything with a group action could qualify here.

[7] One can make this informal statement more rigorous using nonstandard analysis and/or ultrafilters, but we will not take such an approach here; see Section 4.4.

from one type of mathematics can be used to prove results in the other (cf. [**Ta2008**, §1.3]).

In addition to the use of compactness, the other key pillar of the correspondence principle is that of *abstraction* - the ability to generalise from a concrete system (on a very explicit space, e.g. the infinite cube $\{0,1\}^{\mathbf{Z}}$) to an more general abstract setting (e.g. an abstract dynamical system, measure-preserving system, group, etc.) One of the reasons for doing this is that there are various maneuvres one can do in the abstract setting (e.g. passing from a system to a subsystem, a factor, or an extension, or by reasoning by analogy from other special systems that are different from the original concrete system) which can be quite difficult to execute or motivate if one stays within the confines of a single concrete setting. (See also Section 1.6 for further discussion.)

We now turn to several specific examples of this principle in various contexts. We begin with the more "combinatorial" or "non-ergodic theoretical" instances of this principle, in which there is no underlying probability measure involved; these situations are simpler than the ergodic-theoretic ones, but already illustrate many of the key features of this principle in action.

**4.3.1. The correspondence principle in Ramsey theory.** We begin with the classical correspondence principle that connects Ramsey results about finite colourings, to Ramsey results about infinite colourings, or (equivalently) about the topological dynamics of covers of open sets. We illustrate this by demonstrating the equivalence of three statements. The first two are as follows:

**Theorem 4.3.1** (van der Waerden theorem, Type 2 formulation)**.** *Suppose the integers are coloured by finitely many colours. Then there exist arbitrarily long monochromatic arithmetic progressions.*

**Theorem 4.3.2** (van der Waerden theorem, Type 1 formulation)**.** *For every $c$ and $k$ there exists $N$ such that whenever $\{1,\dots,N\}$ is coloured by $c$ colours, there exists a monochromatic arithmetic progression of length $k$.*

It is easy to see that Theorem 4.3.2 implies Theorem 4.3.1. Conversely, to deduce Theorem 4.3.2 from Theorem 4.3.1, we use the compactness and contradiction argument as follows. Assume for contradiction that Theorem 4.3.1 was true, but Theorem 4.3.2 was false. Untangling the quantifiers, this means that there exist positive integers $k$, $c$, a sequence $N_n$ going to infinity, and colourings $\{1,\dots,N_n\} = A_{n,1} \cup \dots \cup A_{n,c}$ of $\{1,\dots,N_n\}$ into $c$ colours, none of which contain any monochromatic arithmetic progressions of length $k$.

By shifting the sets $\{1,\dots,N_n\}$ and redefining $N_n$ a little, we can replace $\{1,\dots,N_n\}$ by $\{-N_n,\dots,N_n\}$. This sequence of colourings on the

increasingly large sets $\{-N_n, \ldots, N_n\}$. One can now extract a subsequence of such colourings on finite sets of integers that converge pointwise or weakly to a colouring on the whole set of integers by the usual "Arzelá-Ascoli diagonalisation trick". Indeed, by passing to an initial subsequence (and using the infinite pigeonhole principle), one can ensure that all of these colourings eventually become a constant colour at 0; refining to another subsequence, we can ensure it is a constant colour at 1; then at $-1, 2, -2$, and so forth. Taking a diagonal subsequence of these sequences, we obtain a final subsequence of finite colourings that converges pointwise to an infinite limit colouring. By Theorem 4.3.1, this limit colouring contains an monochromatic arithmetic progression of length $k$. Now note that the property of being monochromatic at this progression is a local one: one only needs to inspect the colour of finitely many of the integers in order to verify this property. Because of this, this property of the infinite limit colouring will also be shared by the finite colourings that are sufficiently far along the converging sequence. But we assumed at the very beginning that none of these finite colourings have a monochromatic arithmetic progression of length $k$, a contradiction, and the claim follows.

The above argument, while simple, has all the basic ingredients of the correspondence principle in action: a proof by contradiction, use of weak compactness to extract an infinite limiting object, application of the infinitary result to that object, and checking that the conclusion of that result is sufficiently[8] "finitary", "local", or "continuous" that it extends back to some of the finitary sequence, leading to the desired contradiction.

A key disadvantage of the use of the compactness and contradiction argument, though, is that it is quite difficult to extract specific quantitative bounds from any argument that uses this argument; for instance, while one can eventually "proof mine" the above argument (combined with some standard proof of Theorem 4.3.1) to eventually get a bound on $N$ in terms of $k$ and $d$, such a bound is extremely poor (of Ackermann function type).

Theorem 4.3.1 can be reformulated in a more abstract form:

**Theorem 4.3.3** (van der Waerden theorem, Type 4 version)**.** *Let $X$ be a compact space, let $T : X \to X$ be a homeomorphism, and let $(V_\alpha)_{\alpha \in A}$ be an open cover of $X$. Then for any $k$ there exists a positive integer $n$ and an open set $V_\alpha$ in the cover such that $V_\alpha \cap T^{-n} V_\alpha \cap \ldots T^{-(k-1)n} V_\alpha$ is non-empty.*

[8] It is essential that one manages to reduce to purely local properties before passing from the converging sequence to the limit, or vice versa, since non-local properties are usually not preserved by the limit. For instance, consider the colouring of $\{-N, \ldots, N\}$ which colours every integer between -N/2 and N/2 blue, and all the rest red. Then this converges weakly to the all-blue colouring, and clearly the (non-local) property of containing at least one red element is not preserved by the limit.

The deduction of Theorem 4.3.3 from Theorem 4.3.1 is easy, after using the compactness to refine the open cover to a finite subcover, picking a point $x_0$ in $X$, and then colouring each integer $n$ by the index $\alpha$ of the first open set $V_\alpha$ that contains $T^n x_0$. The converse deduction of Theorem 4.3.1 from Theorem 4.3.3 is the one which shows the "dynamical" aspect of this theorem: we can encode a colouring $\mathbf{Z} = A_1 \cup \ldots \cup A_c$ of the integers as a point $x_0 := (c_n)_{n\in\mathbf{Z}}$ in the infinite product space $\{1,\ldots,c\}^{\mathbf{Z}}$, where $c_n$ is the unique class such that $n \in A_{c_n}$ (indeed, one can think of this product space as the space of all c-colourings of the integers). The infinite product space is compact with the product (or weak) topology used earlier, thus a sequence of colourings converge to a limit iff they converge locally (or pointwise). This space also comes with the standard shift $T : (x_n)_{n\in\mathbf{Z}} \to (x_{n-1})_{n\in\mathbf{Z}}$ (corresponding to right shift on the space of colourings). If we let $X$ be the closure of the orbit $\{T^n x_0 : n \in \mathbf{Z}\}$, and let $V_1,\ldots,V_c$ be the open cover $V_i := \{(x_n)_{n\in\mathbf{Z}} : x_0 = i\}$, it is straightforward to show that Theorem 4.3.3 implies Theorem 4.3.1.

**4.3.2. The correspondence principle for finitely generated groups.** The combinatorial correspondence used above for colourings can also be applied to other situations, such as that of finitely generated groups. Recall that if $G$ is a group generated by a finite set $S$, we say that $G$ has *polynomial growth* if there exists constants $K$, $d$ such that for every $r \geq 1$, the ball $B_r$ of radius $r$ (i.e. the set of words in $S$ of length at most $r$) has cardinality at most $Kr^d$. Such groups were classified by a well-known theorem of Gromov [**Gr1981**]:

**Theorem 4.3.4** (Gromov's theorem on polynomial growth, Type 4 version)**.** *Let $G$ be a finitely generated group of polynomial growth. Then $G$ is virtually nilpotent (i.e. it has a finite index subgroup that is nilpotent).*

This theorem is discussed further in Section 2.5.

As observed in [**Gr1981**], Theorem 4.3.4 is equivalent to a finitary version:

**Theorem 4.3.5** (Gromov's theorem on polynomial growth, Type 3)**.** *For every integers $s$, $K$, $d$, there exists $s$ such that any finitely generated group with $s$ generators, such that $B_r$ has cardinality at most $Kr^d$ for all $1 \leq r \leq R$, is virtually nilpotent.*

It is clear that Theorem 4.3.5 implies Theorem 4.3.4. For the converse implication, we use the compactness and contradiction argument. We sketch the details as follows. First, we make things more concrete (i.e. move from

Type 4.3.3 and Type 4.3.4 to Type 4.3.1 and Type 4.3.2 respectively) by observing that every group $G$ on $s$ generators can be viewed as a quotient $\mathbf{F}_s/\Gamma$ of the (nonabelian) free group on $s$ generators by some normal subgroup $\Gamma$.

Suppose for contradiction that Theorem 4.3.5 failed in this concrete setting; then there exists $s$, $K$, $d$, a sequence $R_n$ going to infinity, and a sequence $G_n = \mathbf{F}_s/\Gamma_n$ of groups such that each $G_n$ obeys the volume condition $|B_r| \leq Kr^d$ for all $1 \leq r \leq R_n$.

The next step, as before, is to exploit weak sequential compactness and extract a subsequence of groups $G_n = \mathbf{F}_s/\Gamma_n$ that "converge" to some limit $G = \mathbf{F}_s/\Gamma$, in the "weak" or "pointwise" sense that $\Gamma_n$ converges pointwise (or locally) to $\Gamma$ (much as with the convergence of colourings in the previous setting). The Arzelá-Ascoli argument as before shows that we can find a subsequence of $G_n = \mathbf{F}_s/\Gamma_n$ which do converge pointwise to some limit object $G = \mathbf{F}_s/\Gamma$; one can check that the property of being a normal subgroup is sufficiently "local" that it is preserved[9] under limits, thus $\Gamma$ is a normal subgroup of $\mathbf{F}_s$ and so $G$ is well-defined as a group.

As volume growth is a local condition (involving only words of bounded length for any fixed $r$), we then easily conclude that $G$ is of polynomial growth, and thus by Theorem 4.3.4 is virtually nilpotent. Some nilpotent algebra reveals that every virtually nilpotent group is finitely presented, so in particular there are a finite list of relations among the generators which guarantee this virtual nilpotency property. Such properties are local enough that they must then persist to groups $G_n$ sufficiently far along the subsequence, contradicting Theorem 4.3.5.

A slight modification of the above argument also reveals that the step and index of the nilpotent subgroup of $G$ can be bounded by some constant depending only on $K$, $d$, $s$; this gives Theorem 4.3.5 meaningful content even when $G$ is finite (in contrast to Theorem 4.3.4, which is trivial for finite groups). On the other hand, an explicit bound for this constant (or for $R$) in terms of $s$, $K$, $d$ was only obtained quite recently, in [**ShTa2010**].

**4.3.3. The correspondence principle for dense sets of integers.** Now we turn to the more "ergodic" variants of the correspondence principle, starting with the fundamental Furstenberg correspondence principle connecting combinatorial number theory with ergodic theory. We will illustrate this with the classic example of Szemerédi's theorem [**Sz1975**].

There are many finitary versions of Szemerédi's theorem. Here is one:

**Theorem 4.3.6** (Szemerédi's theorem, Type 1 version)**.** *Let $k \geq 2$ and $0 < \delta \leq 1$. Then there exists a positive integer $N = N(\delta, k)$ such that*

[9] One way to view this convergence is that algebraic identity obeyed by the generators of $G$, is eventually obeyed by the groups sufficiently far along the convergent subsequence, and conversely.

*every subset $A$ of $\{1,\dots,N\}$ with $|A| \geq \delta N$ contains at least one $k$-term arithmetic progression.*

The standard "Type 2" formulation of this theorem is the assertion that any subset of the integers of positive upper density has arbitrarily long arithmetic progressions. While this statement is indeed easily shown to be equivalent to Theorem 4.3.6, the Furstenberg correspondence principle instead connects this formulation to a rather different one, in which the deterministic infinite set is replaced by a *random* one. Recall that a random subset of integers $\mathbf{Z}$ is a random variable $A$ taking values in the power set $2^{\mathbf{Z}}$ of the integers (or more formally, with a distribution that is a Borel probability measure on $2^{\mathbf{Z}}$ with the product topology), and so in particular the probabilities of any cylinder events such as

$$\mathbf{P}(3,5 \in A; 7, 11 \notin A)$$

that involve only finitely many of the elements of $A$, are well-defined as numbers between 0 and 1. The Carathéodory extension theorem[10] (combined with some topological properties of $2^{\mathbf{Z}}$) shows, conversely, that any assignment of numbers between 0 and 1 to each cylinder set, which obeys various compatibility conditions such as

$$\mathbf{P}(3 \in A) = \mathbf{P}(3,5 \in A) + \mathbf{P}(3 \in A; 5 \notin A)$$

can be shown to give rise to a well-defined random set $A$.

We say that a random set $A$ of integers is *stationary* if for every integer $h$, the shifted set $A+h$ has the same probability distribution as $A$. In terms of cylinder events, this is equivalent to a collection of assertions such as

$$\mathbf{P}(3,5 \in A; 7, 11 \notin A) = \mathbf{P}(3-h,5-h \in A; 7-h, 11-h \notin A)$$

and so forth. One can then equate Theorem 4.3.6 with

**Theorem 4.3.7** (Szemerédi's theorem, Type 2 version)**.** *Let $A$ be a stationary random infinite set of integers such that $\mathbf{P}(0 \in A) > 0$ (which, by stationarity, implies that $\mathbf{P}(n \in A) > 0$ for all $n$), and let $k \geq 2$. Then, with positive probability, $A$ contains a $k$-term arithmetic progression for each $k$.*

It is not difficult to show that Theorem 4.3.6 implies Theorem 4.3.7. We briefly sketch the converse implication, which (unsurprisingly) goes via the usual compactness-and-contradiction argument. Suppose for contradiction that Theorem 4.3.7 is true, but Theorem 4.3.6 fails. Then we can find $k$ and $\delta$, a sequence of $N_n$ going to infinity, and sets $A_n \subset \{1,\dots,N_n\}$ with $|A_n| \geq \delta N_n$ with no $k$-term arithmetic progressions.

[10] See e.g. [**Ta2011**, §1.7] for a discussion of this theorem.

We now need to extract a stationary random infinite set $A$ of integers as a limit of the deterministic finite sets $A_n$. The way one does that is by randomly translating each of the $A_n$. More precisely, let $B_n$ denote the random finite set $A_n - h$, where $h$ is chosen from $\{1, \ldots, N_n\}$ uniformly at random. The probability distribution $\mu_n$ of $B_n$ is a discrete probability measure on $2^{\mathbf{Z}}$ which is "almost stationary" in the sense that $B_n + 1$ (say) has a distribution very close to $B_n$; for instance probabilities such as $\mathbf{P}(3, 5 \in B_n)$ and $\mathbf{P}(3, 5 \in B_n + 1)$ can easily be seen to differ only by $O(1/N_n)$. Also, the fact that $|A_n| \geq \delta N_n$ equates to the assertion that $\mathbf{P}(0 \in B_n) \geq \delta$.

By using the same type of Arzel-Ascoli argument as before, we can show that some subsequence of the random variables $B_n$ converge weakly[11] to a limit $B$ in the sense that the cylinder probabilities of $B_n$ converge to those of B along this subsequence; thus for instance

$$\lim_{n\to\infty} \mathbf{P}(3, 5 \in B_n; 17 \notin B_n) = \mathbf{P}(3, 5 \in B; 17 \notin B).$$

Since the $B_n$ are approximately stationary, one can show that $B$ is exactly stationary. Since $\mathbf{P}(0 \in B_n)$ is bounded uniformly away from zero, one can show that $\mathbf{P}(0 \in B)$ is positive. Thus, we can apply Theorem 4.3.7 to show that $B$ contains a $k$-term arithmetic progression with positive probability. Since there are only countably many k-term arithmetic progressions, the countable pigeonhole principle then tells us that there exists some $k$-term arithmetic progression $a, a + r, \ldots, a + (k-1)r$ which lies in $B$ with positive probability. This is a "local" property on $B$. By weak convergence, this means that this same is true for $B_n$ for $n$ sufficiently far along this subsequence; in particular, the corresponding deterministic sets $A_n$ contain $k$-term arithmetic progressions, a contradiction. Thus Theorem 4.3.7 does imply Theorem 4.3.6.

Much as Theorem 4.3.1 is equivalent to Theorem 4.3.4, Theorem 4.3.7 can be reformulated in a more abstract manner, known as the *Furstenberg recurrence theorem*:

**Theorem 4.3.8** (Szemerédi's theorem, Type 4 version)**.** *Let $(X, \mu, T)$ be a probability space with a measure-preserving bimeasurable map $T : X \to X$ (thus $T$ is invertible, and $T, T^{-1}$ are measurable and measure-preserving), and let $A \subset X$ have positive measure $\mu(A) > 0$. Then there exists $r > 0$ such that $\mu(A \cap T^{-r}A \cap \ldots \cap T^{(k-1)r}A) > 0$.*

We leave the equivalence of Theorem 4.3.8 with Theorems 4.3.6, 4.3.7 as an exercise. (See also [**Ta2009**, §2.10] for further discussion.)

[11] To get from the cylinder probabilities back to a random variable, one uses the *Carathéodory extension theorem*. Weak convergence (or more precisely, weak-* convergence) of measures is also known as *vague convergence*.

**4.3.4. The correspondence principle for dense sets of integers, II.** The deduction of Theorem 4.3.6 from Theorem 4.3.7 gives that the set $A$ appearing in Theorem 4.3.6 has at least one $k$-term arithmetic progression, but if one inspects the argument more carefully, one sees that in fact one has a stronger statement that if $N$ is large enough. Namely, there exists some $1 \leq r \leq C(k,\delta)$ such that $A$ contains at least $c(k,\delta)N$ k-term arithmetic progressions $a, a+r, \ldots, a+(k-1)r$ of step r. We leave this derivation as an exercise.

It is possible however to find even more progressions in the set $A$:

**Theorem 4.3.9** (Szemerédi's theorem, Varnavides-type version)**.** *Let $k \geq 2$ and $0 < \delta \leq 1$. Then there exists a positive integer $N_0 = N_0(\delta, k)$ and $c = c(k,\delta) > 0$ such that every subset $A$ of $\{1, \ldots, N\}$ with $|A| \geq \delta N$ contains at least $cN^2$ $k$-term arithmetic progressions.*

This can be obtained from Theorem 4.3.7 by repeating the derivation of Theorem 4.3.6 with two additional twists. Firstly, it is not difficult to modify the $N_n$ appearing in this derivation to be prime (for instance, by appealing to *Bertrand's postulate*). This allows us to identify $\{1, \ldots, N_n\}$ with the finite field $\mathbf{Z}/N_n\mathbf{Z}$, giving us the ability to dilate within this set as well as translate. For technical reasons it is also convenient to restrict $A_n$ to lie in the bottom half $\{1, \ldots, \lfloor N_n/2 \rfloor\}$ of this set, which is also easy to arrange. We then argue as before, but with the randomly translated set $B_n := A_n - h$ replaced by the randomly translated and dilated set $B_n := (A_n - h) \cdot r$, where $h$ and $r$ are independently chosen at random from this finite field. If one does this, one finds that probabilities such as $\mathbf{P}(0, 1, \ldots, k-1 \in B_n)$ are essentially equal to the number of $k$-term arithmetic progressions in $A_n$, divided by $N_n^2$ (the restriction of $A_n$ to the bottom half of $\{1, \ldots, N_n\}$ is necessary to avoid certain "wraparound" issues). If one then repeats the previous arguments one can establish Theorem 4.3.9 from Theorem 4.3.7.

This type of argument was implicitly first introduced by Varnavides [**Va1959**]. Basically, this argument exploits the affine invariance (i.e. $\mathrm{Aff}(\mathbf{Z})$ invariance) of the space of arithmetic progressions, as opposed to mere translation invariance (i.e. $\mathbf{Z}$ invariance).

One can rephrase Theorem 4.3.8 in a quantitative ergodic formulation, essentially due to Bergelson, Host, McCutcheon, and Parreau [**BeHoMcPa2000**]:

**Theorem 4.3.10** (Szemerédi's theorem, Type 3 version)**.** *Let $k \geq 2$ and $0 < \delta \leq 1$. Then there exists $c(k,\delta) > 0$ such that for every measure-preserving system $(X, \mu, T)$ and any measurable set $A$ with $\mu(A) > \delta$, we have $\liminf_{N\to\infty} \mathbf{E}_{n \in \{1,\ldots,N\}} \mu(A \cap T^{-n}A \cap \ldots \cap T^{-(k-1)n}A) \geq c(k,\delta)$.*

**4.3.5. The correspondence principle for sparse sets of integers.** It is possible to squeeze even more finitary results out of Theorem 4.3.7 than

was done in the previous two sections. In particular, one has the following relative version of Szemerédi's theorem from [**GrTa2008**]:

**Theorem 4.3.11** (Relative Szemerdi theorem)**.** *Let $k \geq 2$ and $0 < \delta \leq 1$. Let $N$ be a sufficiently large integer, and let $R$ be a "sufficiently pseudorandom" subset of $\{1, \ldots, N\}$. Then every subset $A$ of $R$ with $|A| \geq \delta |R|$ contains one $k$-term arithmetic progression.*

I did not define above what "sufficiently pseudorandom"' meant as it is somewhat technical, but very roughly speaking it is a package of approximate independence conditions which include things like

$$\mathbf{E}_{n,h,k \in [N]} \nu(n)\nu(n+h)\nu(n+k)\nu(n+h+k) = 1 + o(1) \tag{4.1}$$

where $\nu(n) := \frac{N}{|R|} 1_R(n)$ is the normalised indicator function of $R$, and all arithmetic operations are taking place in the cyclic group $\mathbf{Z}/N\mathbf{Z}$.

The point of Theorem 4.3.11 is that it allows one to detect arithmetic progressions inside quite sparse sets of integers (typically,R will have size about $N/\log N$, so $A$ and $R$ would be logarithmically sparse). In particular, a relative Szemerédi theorem similar to this one (but somewhat stronger[12] was a key ingredient in the result [**GrTa2008**] that the primes contain arbitrarily long arithmetic progressions.

The derivation[13] of Theorem 4.3.11 from Theorem 4.3.7 is discussed in [**Ta2004**]. We sketch the main ideas here. Once again, we argue by contradiction. If Theorem 4.3.11 failed, then one can find $k$ and $\delta$, a sequence $N_n$ going to infinity, a sequence $R_n$ of "increasingly pseudorandom" subsets of $\{1, \ldots, N_n\}$, and sets $A_n \subset R_n$ with $|A_n| \geq \delta |R_n|$ such that none of the $A_n$ contain $k$-term arithmetic progressions.

As before, it is not difficult to ensure $N_n$ to be prime, and that $A_n$ lives in the bottom half $R_n \cap \{1, \ldots, \lfloor N_n/2 \rfloor\}$ of $R_n$. We then create the random translated and dilated set $B_n := (A_n - h) \cdot r$ as before; note that $B_n$ still has no $k$-term arithmetic progressions (except in the degenerate case $r = 0$, but this case is extremely rare and becomes irrelevant in the limit). We can also randomly translated and dilate $R_n$ by the same parameters to obtain a random set $S_n := (R_n - r) \cdot r$; thus $B_n$ is a relatively dense subset of the random (and potentially quite sparse) set $S_n$.

In the previous two sections, we looked at the (Borel) probability measure $\mu_n$ on the power set $2^{\mathbf{Z}}$ formed by the distribution of $B_n$, which can be

[12] Theorem 4.3.11 is unfortunately insufficient for this task, for the technical reason that the amount of pseudorandomness needed here depends on $\delta$; the relative Szemerédi's theorem developed in my paper with Ben only needs a number of pseudorandomness conditions that depend on $k$ but - crucially - not on $\delta$.

[13] There are other approaches known to obtain this implication, for instance via the Hahn-Banach theorem: see [**ReTrTuVa2008**], [**Go2010**], or [**Ta2011b**, §1.7].

viewed as a collection of cylinder statistics such as

$$\mu_n(C(3,5,\overline{17})) = \mathbf{P}(3,5 \in B_n; 17 \notin B_n)$$

where $C(3,5,\overline{17})$ is the cylinder set $\{A \subset \mathbf{Z} : 3,5 \in A; 17 \notin A\}$. In order for these statistics to actually arise from a Borel probability measure, these statistics have to be numbers lying between 0 and 1, and they have to obey compatibility conditions such as

$$\mu_n(C(3,\overline{17})) = \mu_n(C(3,5,\overline{17})) + \mu_n(C(3,\overline{5},\overline{17})),$$

and also

$$\mu_n(2^{\mathbf{Z}}) = \mu_n(C()) = 1.$$

Conversely, any set of non-negative numbers obeying these properties will give a Borel probability measure, thanks to the *Carathéodory extension theorem*.

In the sparse case, this approach does not work, because $\mu_n$ degenerates in the limit if $R_n$ is sparse. For instance, because $B_n$ is so sparse, the probability $\mathbf{P}(3,5 \in B_n; 17 \notin B_n)$ can go to zero; indeed, we expect this quantity to look like $O(|R_n|/N)^2$. To fix this we do two things. Firstly, we replace the absolute complement $\overline{B_n} = \{1,\ldots,N_n\}\backslash B_n$ that implicitly appears above by the relative complement $S_n$. Secondly, we introduce the normalising factor $N_n/|R_n|$, so for instance the cylinder set $C(3,5,\overline{17})$ will now be assigned the normalised weight

$$\mu_n(C(3,5,\overline{17})) = (\frac{N_n}{|R_n|})^3 \mathbf{P}(3,5 \in B_n; 17 \in S_n\backslash B_n)$$

and similarly for other cylinder sets. Perhaps more suggestively, we have

$$\mu_n(C(3,5,\overline{17})) = \mathbf{E}_{a,r \in \mathbf{Z}/N_n\mathbf{Z}} \Lambda_n(a+3r)\Lambda_n(a+5r)(\nu_n - \Lambda_n)(a+17r)$$

where $\Lambda_n := \frac{N_n}{|R_n|} 1_{A_n}$ and $\nu_n := \frac{N_n}{|R_n|} 1_{R_n}$.

This gives us a non-negative number assigned to every cylinder set, but unfortunately these numbers do not obey the compatibility conditions required to make these numbers arise from a probability measure. However, if one assumes enough pseudorandomness conditions such as (1), one can show that the compatibility conditions are satisfied approximately, thus for instance

$$\mu_n(C(3,\overline{17})) = \mu_n(C(3,5,\overline{17})) + \mu_n(C(3,\overline{5},\overline{17})) + o(1)$$

or equivalently

$$\mathbf{E}_{a,r \in \mathbf{Z}/N_n\mathbf{Z}} \Lambda(a+3r)(\nu-1)(a+5r)(\nu-\Lambda)(a+17r) = o(1).$$

These conditions can be checked using a large number of applications of the Cauchy-Schwarz inequality, which we omit here. Thus, $\mu_n$ is not a true probability measure, but is some sort of approximate "virtual probability

measure". It turns out that these virtual probability measures enjoy the same crucial weak compactness property as actual probability measures, and one can repeat all the previous arguments to deduce Theorem 4.3.11 from Theorem 4.3.7.

**4.3.6. The correspondence principle for graphs.** In the previous three sections we considered the correspondence principle in the integers $\mathbf{Z}$. It is not difficult to replace the integers with other amenable groups, such as $\mathbf{Z}^d$ for some fixed $d$. Now we discuss a somewhat different-looking instance of the correspondence principle, coming from graph theory, in which the underlying group is now the (small) permutation group[14] $\mathrm{Perm}_0(\mathbf{Z})$ consisting of those permutations $\sigma : \mathbf{Z} \to \mathbf{Z}$ which permute only finitely many elements. We illustrate the graph correspondence principle with the following old result of Ruzsa and Szemerédi [**RuSz1978**]:

**Theorem 4.3.12** (Triangle removal lemma, Type 3 version)**.** *For every $\epsilon > 0$ there exists $\delta > 0$ and $N_0 > 0$ such that for any $G = (V, E)$ be a graph on $N$ vertices for some $N \geq N_0$ which has at most $\delta N^3$ triangles, one can remove at most $\epsilon N^2$ edges from the graph to make the graph triangle free.*

We will not discuss why this theorem is of interest, other than to mention in passing that it actually implies the $k = 3$ case of Szemerédi's theorem; see e.g. [**TaVu2006**, §10.6]. It turns out that this result can be deduced from some infinitary versions of this lemma. Here is one instance:

**Theorem 4.3.13** (Triangle removal lemma, Type 2 version)**.** *Let $G$ be a random graph on the integers $\mathbf{Z}$ which is* exchangeable[15] *in the sense that any permutation of $G$ has the same distribution as $G$. Suppose that $G$ is almost surely triangle free, and let $\varepsilon > 0$. Then there exists a continuous function $F$ from the space $2^{\binom{\mathbf{Z}}{2}}$ of graphs on $\mathbf{Z}$ (with the product topology) to the space $2^{\binom{\mathbf{N}}{2}}$ of graphs on $\mathbf{N}$, which is equivariant with respect to permutations of $\mathbf{N}$, such that $G' := F(G)$ is surely (not just almost surely) a subgraph of $G$ which is triangle free, and such that $\mathbf{P}((1,2) \in G \backslash G') \leq \varepsilon$.*

**Theorem 4.3.14** (Triangle removal lemma, Type 3 version)**.** *Let $(X, \mu)$ be a probability space with a measure-preserving action of $\mathrm{Perm}_0(\mathbf{Z})$, and let $E_{12}, E_{23}, E_{31}$ be three measurable sets which are invariant under the stabiliser of $\{1,2\}$, $\{2,3\}$, and $\{3,1\}$ respectively. Suppose that $E_{12} \cap E_{23} \cap E_{31}$ has measure zero. Then one can find subsets $E'_{12}, E'_{23}, E'_{31}$ respectively of $E_{12}, E_{23}, E_{31}$, which remain invariant under the stabilisers of $\{1,2\}$, $\{2,3\}$, and $\{3,1\}$ respectively, such that $E'_{12} \cap E'_{23} \cap E'_{31}$ is empty (and not just measure zero).*

[14] It is convenient to work with this group, rather than the entire group $\mathrm{Perm}(\mathbf{Z})$ of permutations, as it is countable.

[15] This is the analogue of stationarity in this setting.

The proofs that Theorem 4.3.13 and Theorem 4.3.14 imply Theorem 4.3.12 are somewhat technical, but in the same spirit as the previous applications of the correspondence principle; see [**Ta2007b**], this paper respectively for details. In fact they prove slightly stronger statements than Theorem 4.3.12, in that they give a bit more information as to how the triangle-free graph $G'$ is obtained from the nearly triangle-free graph $G$. The same methods also apply to hypergraphs without much further difficulty, as is done in the above papers, but we will not discuss the details here.

**4.3.7. The correspondence principle over finite fields.** The correspondence principle can also be applied quite effectively to the finite field setting, which is a dyadic model for the integer setting. In [**TaZi2010**], for instance, the equivalence of the following two results was shown:

**Theorem 4.3.15** (Inverse Gowers conjecture for finite fields, Type 1 version)**.** *Let $\mathbf{F}$ be a finite field, let $k \geq 2$, let $\delta > 0$, and let $f : \mathbf{F}^n \to \mathbf{C}$ be a function on some vector space $\mathbf{F}^n$ bounded in magnitude by* 1*, and such that the* Gowers uniformity norm

$$\|f\|_{U^k(F^n)} := (\mathbf{E}_{h_1,\ldots,h_k,x\in\mathbf{F}^n} \Delta_{h_1} \ldots \Delta_{h_k} f(x))^{1/2^k}$$

*is larger than $\delta$, where $\Delta_h f(x) := f(x+h)\overline{f(x)}$. Then there exists a function $\phi : \mathbf{F}^n \to S^1$ which is a* phase polynomial *of degree at most $k$ in the sense that $\Delta_{h_1} \ldots \Delta_{h_k} \phi(x) = 1$ for all $h_1, \ldots, h_k, x \in \mathbf{F}^n$ (or equivalently, that $\|\phi\|_{U^k(F^n)} = 1$), such that we have the correlation $|\mathbf{E}_{x\in F^n} f(x)\overline{\phi(x)}| \geq c(F,k,\delta)$ for some $c(F,k,\delta) > 0$ independent of $n$.*

**Theorem 4.3.16** (Inverse Gowers conjecture for finite fields, Type 4 version)**.** *Let $(X,\mu)$ be a probability space, let $\mathbf{F}$ be a finite field, and let $g \mapsto T_g$ be a measure-preserving action of the infinite group $F^\omega := \lim_{\leftarrow} F^n$. Let $f \in L^\infty(X)$ be such that the Gowers-Host-Kra seminorm*

$$\|f\|_{U^k(X)} := \lim_{n\to\infty} (\mathbf{E}_{h_1,\ldots,h_k\in F^n} \int_X \Delta_{h_1} \ldots \Delta_{h_k} f\mu)^{1/2^k}$$

*is positive, where $\Delta_h f(x) := f(T^h x)\overline{f(x)}$. Then there exists $\phi : X \to S^1$ which is a phase polynomial in the sense that $\Delta_{h_1} \ldots \Delta_{h_k} \phi = 1$ a.e., and which correlates with $f$ in the sense that $\int_X f\overline{\phi}\, d\mu \neq 0$.*

In [**BeTaZi2010**], [**TaZi2011**], Theorem 4.3.16 was established, which by the correspondence principle alluded to above, implies Theorem 4.3.15. (See [**Ta2011b**, §1.5] for further discussion.)

Very roughly speaking, the reason why the correspondence principle is more effective here than on the integers is because the vector space $\mathbf{F}^n$ enjoys a massively transitive action of the general linear group $GL(\mathbf{F}^n)$ that mixes things around in a manner much stronger than even the affine action $\mathrm{Aff}(\mathbf{Z})$

mentioned earlier (which is basically 2-transitive but not $k$-transitive for any higher $k$).

**4.3.8. The correspondence principle for convergence of ergodic averages.** The final instance of the correspondence principle that we will discuss here goes in the opposite direction from previous instances. In the preceding seven instances, the interesting aspect of the principle was that one could use a qualitative result about infinite systems to deduce a quantitative result about finite systems. Here, we will do the reverse: we show how a result about infinite systems can be deduced from one from a finite system. We will illustrate this with a very simple result from ergodic theory, the *mean ergodic theorem*:

**Theorem 4.3.17** (Mean ergodic theorem, Type 4 version)**.** *Let $(X, \mu, T)$ be a measure-preserving system, and let $f \in L^2(X)$. Let $S_N$ be the averaging operators $S_N f := \mathbf{E}_{1 \leq n \leq N} T^n f$. Then $S_N f$ is a convergent sequence in $L^2(X)$ as $N \to \infty$.*

This is of course a well-known result with many short and elegant proofs; the proof method that we sketch here (essentially due to Avigad, Gerhardy, and Towsner [**AvGeTo2010**]) is lengthier and messier than the purely infinitary proofs, but can be extended to some situations in which it had been difficult to proceed in an infinitary manner (see e.g. [**Ta2008b**]).

The basic problem with finitising this theorem is that there is no uniform rate of convergence in the mean ergodic theorem: given any $\epsilon > 0$, we know that the averages $S_N f$ eventually lie within $\varepsilon$ of their limit for $N$ large enough, but it is known that the $N$ we need for this is not uniform in the choice of the system $(X, \mu, T)$ or the function f, and can indeed by arbitrarily large for given $\varepsilon$ even after fixing the size of $f$. So a "naive" finitisation does not work, much as a naive finitisation of the infinite convergence principle (every bounded monotone sequence converges), as discussed in [**Ta2008**, §1.3].

The resolution is in fact very similar to that discussed in [**Ta2008**, §1.3]. Observe that if $x_1, x_2, \ldots$ is any sequence in a complete metric space (e.g. the real line, or $L^2(X)$), the statement that "$x_n$" converges", or equivalently that "$x_n$ is Cauchy", is equivalent to

> For every $\varepsilon > 0$, there exists $n$ such that for all sufficiently large $m$, $d(x_n, x_m) \leq \varepsilon$,

which is in turn equivalent to the lengthier, but more finitistic, statement

> For every $\varepsilon > 0$ there exists $F_0 : \mathbf{N} \to \mathbf{N}$ such that for every $F : \mathbf{N} \to \mathbf{N}$ that grows faster than $F_0$ in the sense

that $F(n) > F_0(n)$ for all $n$, one has $d(x_n, x_{F(n)}) \leq \varepsilon$ for some $n$.

The point here is that once the function F is selected, one only has to verify the closeness of a single pair of elements in the sequence, rather than infinitely many. This makes it easier to finitise the convergence statement effectively. Indeed, Theorem 4.3.17 is easily seen (by another compactness and contradiction argument) to be equivalent to

**Theorem 4.3.18** (Mean ergodic theorem, Type 3 version)**.** *For every $\varepsilon > 0$ and every sufficiently rapid $F$ (i.e. $F$ grows faster than some $F_0$ depending on $\varepsilon$) there exists $N$ such that for every measure-preserving system $(X, \mu, T)$ and every $f \in L^2(X)$ with $\|f\|_{L^2(X)} \leq 1$, we have $\|S_n f - S_{F(n)} f\|_{L^2(X)} \leq \varepsilon$ for some $1 \leq n \leq N$.*

Note that this theorem is quantitative in the sense that $N$ depends only on $\varepsilon$ and $F$, and not on the underlying system; indeed one can give an explicit value for $N$, arising from iterating $F$ about $1/\varepsilon^2$ times. See [**Ta2009**, §2.8] for further discussion.

## 4.4. Nonstandard analysis as a completion of standard analysis

Many structures in mathematics are *incomplete* in one or more ways. For instance, the field of rationals $\mathbf{Q}$ or the reals $\mathbf{R}$ are *algebraically incomplete*, because there are some non-trivial algebraic equations (such as $x^2 = 2$ in the case of the rationals, or $x^2 = -1$ in the case of the reals) which could *potentially* have solutions (because they do not imply a necessarily false statement, such as $1 = 0$, just using the laws of algebra), but do not *actually* have solutions in the specified field.

Similarly, the rationals $\mathbf{Q}$, when viewed now as a metric space rather than as a field, are also *metrically incomplete*, beause there exist sequences in the rationals (e.g. the decimal approximations $3, 3.1, 3.14, 3.141, \ldots$ of the irrational number $\pi$) which could *potentially* converge to a limit (because they form a *Cauchy sequence*), but do not *actually* converge in the specified metric space.

A third type of incompleteness is that of *logical incompleteness*, which applies now to formal theories rather than to fields or metric spaces. For instance, *Zermelo-Frankel-Choice (ZFC) set theory* is logically incomplete, because there exist statements (such as the consistency of ZFC) which could *potentially* be provable by the theory (because it does not lead to a contradiction, or at least so we believe, just from the axioms and deductive rules of the theory), but is not *actually* provable in this theory.

A fourth type of incompleteness, which is slightly less well known than the above three, is what I will call *elementary incompleteness* (and which model theorists call the failure of the *countable saturation property*). It applies to any structure that is describable by a first-order language, such as a field, a metric space, or a universe of sets. For instance, in the language of ordered real fields, the real line $\mathbf{R}$ is elementarily incomplete, because there exists a sequence of statements (such as the statements $0 < x < 1/n$ for natural numbers $n = 1, 2, \ldots$) in this language which are *potentially* simultaneously satisfiable (in the sense that any finite number of these statements can be satisfied by some real number $x$) but are not *actually* simultaneously satisfiable in this theory.

In each of these cases, though, it is possible to start with an incomplete structure and *complete* it to a much larger structure to eliminate the incompleteness. For instance, starting with an arbitrary field $k$, one can take its algebraic completion (or *algebraic closure*) $\overline{k}$; for instance, $\mathbf{C} = \overline{\mathbf{R}}$ can be viewed as the algebraic completion of $\mathbf{R}$. This field is usually significantly larger than the original field $k$, but contains $k$ as a subfield, and every element of $\overline{k}$ can be described as the solution to some polynomial equation with coefficients in $k$. Furthermore, $\overline{k}$ is now *algebraically complete* (or *algebraically closed*): every polynomial equation in $\overline{k}$ which is potentially satisfiable (in the sense that it does not lead to a contradiction such as $1 = 0$ from the laws of algebra), is actually satisfiable in $\overline{k}$.

Similarly, starting with an arbitrary metric space $X$, one can take its *metric completion* $\overline{X}$; for instance, $\mathbf{R} = \overline{\mathbf{Q}}$ can be viewed as the metric completion of $\mathbf{Q}$. Again, the completion $\overline{X}$ is usually much larger than the original metric space $X$, but contains $X$ as a subspace, and every element of $\overline{X}$ can be described as the limit of some Cauchy sequence in $X$. Furthermore, $\overline{X}$ is now a complete metric space: every sequence in $\overline{X}$ which is potentially convergent (in the sense of being a Cauchy sequence), is now actually convegent in $\overline{X}$.

In a similar vein, we have the *Gödel completeness theorem*, which implies (among other things) that for any consistent first-order theory $T$ for a first-order language $L$, there exists at least one *completion* $\overline{T}$ of that theory $T$, which is a consistent theory in which every sentence in $L$ which is potentially true in $\overline{T}$ (because it does not lead to a contradiction in $\overline{T}$) is actually true in $\overline{T}$. Indeed, the completeness theorem provides at least one model (or *structure*) $\mathfrak{U}$ of the consistent theory $T$, and then the completion $\overline{T} = \mathrm{Th}(\mathfrak{U})$ can be formed by interpreting every sentence in $L$ using $\mathfrak{U}$ to determine its truth value. Note, in contrast to the previous two examples, that the completion is usually not unique in any way; a theory $T$ can have multiple inequivalent models $\mathfrak{U}$, giving rise to distinct completions of the same theory.

Finally, if one starts with an arbitrary structure $\mathfrak{U}$, one can form an *elementary completion* ${}^*\mathfrak{U}$ of it, which is a significantly larger structure which contains $\mathfrak{U}$ as a substructure, and such that every element of ${}^*\mathfrak{U}$ is an elementary limit of a sequence of elements in $\mathfrak{U}$ (I will define this term shortly). Furthermore, ${}^*\mathfrak{U}$ is elementarily complete; any sequence of statements that are potentially simultaneously satisfiable in ${}^*\mathfrak{U}$ (in the sense that any finite number of statements in this collection are simultaneously satisfiable), will actually be simultaneously satisfiable. As we shall see, one can form such an elementary completion by taking an *ultrapower* of the original structure $\mathfrak{U}$. If $\mathfrak{U}$ is the *standard universe* of all the *standard* objects one considers in mathematics, then its elementary completion ${}^*\mathfrak{U}$ is known as the *nonstandard universe*, and is the setting for *nonstandard analysis*.

As mentioned earlier, completion tends to make a space much larger and more complicated. If one algebraically completes a finite field, for instance, one necessarily obtains an infinite field as a consequence. If one metrically completes a countable metric space with no isolated points, such as $\mathbf{Q}$, then one necessarily obtains an uncountable metric space (thanks to the *Baire category theorem*). If one takes a logical completion of a consistent first-order theory that can model *true arithmetic*, then this completion is no longer describable by a recursively enumerable schema of axioms, thanks to *Gödel's incompleteness theorem*. And if one takes the elementary completion of a countable structure, such as the integers $\mathbf{Z}$, then the resulting completion ${}^*\mathbf{Z}$ will necessarily be uncountable.

However, there are substantial benefits to working in the completed structure which can make it well worth the massive increase in size. For instance, by working in the algebraic completion of a field, one gains access to the full power of *algebraic geometry*. By working in the metric completion of a metric space, one gains access to powerful tools of real analysis, such as the *Baire category theorem*, the *Heine-Borel theorem*, and (in the case of Euclidean completions) the *Bolzano-Weierstrass theorem*. By working in a logically and elementarily completed theory (aka a *saturated model*) of a first-order theory, one gains access to the branch of model theory known as *definability theory*, which allows one to analyse the structure of definable sets in much the same way that algebraic geometry allows one to analyse the structure of algebraic sets. Finally, when working in an elementary completion of a structure, one gains a *sequential compactness* property, analogous to the Bolzano-Weierstrass theorem, which can be interpreted as the foundation for much of nonstandard analysis, as well as providing a unifying framework to describe various correspondence principles between finitary and infinitary mathematics.

In this section, I wish to expand upon these above points with regard to elementary completion, and to present nonstandard analysis as a completion of standard analysis in much the same way as, say, complex algebra is a completion of real algebra, or real metric geometry is a completion of rational metric geometry.

**4.4.1. Elementary convergence.** In order to understand the concept of a metric completion of a metric space $X = (X, d)$, one needs to know about the distinction between a Cauchy sequence and a convergent sequence. Similarly, to talk about the elementary completion of a structure $\mathfrak{U}$, one needs the notion of an elementarily Cauchy sequence and an elementarily convergent sequence.

Let us set out some notation. We assume that we have some *first-order language* $L$, which allows one to form sentences involving the first-order logical symbols ($\forall$, $\exists$, $\vee$, $\wedge$, $\neg$, $\implies$, etc.), variables of one or more types, the equality symbol $=$, some constant symbols, and some operations and relations. For instance:

- $L$ could be the language of multiplicative groups, in which there is only one type of object (a group element), a constant symbol $e$, a binary operation $\cdot$ from pairs of group elements to group elements, and a unary operation $()^{-1}$ from group elements to group elements.
- $L$ could be the language of real ordered fields, in which there is one type of object (a field element), constant symbols $0, 1$, binary operations $+, \cdot$, and unary operations $-, ()^{-1}$ (with the latter only being defined for non-zero elements), and the order relation $<$.
- $L$ could be the language of (formal) metric spaces, in which there are two types of objects (points in the space, and real numbers), the constants, operations and relations of a real ordered field, and a metric operation $d$ from pairs of points in the space to real numbers.
- $L$ could be the language of sets, in which there is one type of object (a set) and one relation $\in$.
- etc., etc.

We assume that the language has at most countably many types, constants, operations, and relations. In particular, there are at most countably many sentences in $L$.

A *structure* $\mathfrak{U}$ for a language $L$ is a way of interpreting each of the object classes in $L$ as a set, and each of the constants, operations, and relations as an elements, functions, and relations on those sets respectively. For instance, a structure for the language of groups would be a set $G$, together with a constant symbol $e \in G$, a binary function $\cdot : G \times G \to G$, and a

unary operation $()^{-1} : G \to G$. In particular, groups are structures for the language of groups, but so are many non-groups. Each structure $\mathfrak{U}$ can be used to *interpret* any given sentence $S$ in $L$, giving it a truth value of true or false. We write $\mathfrak{U} \models S$ if $S$ is interpreted to be true by $\mathfrak{U}$. For instance, the axioms of a group can be expressed as a single sentence $A$, and a structure $\mathfrak{U}$ for the language of groups is a group if and only if $\mathfrak{U} \models A$.

Now we introduce the notion of elementary convergence.

**Definition 4.4.1** (Elementary convergence)**.** Let $\mathfrak{U}$ be a structure for a language $L$, and let $x_1, x_2, \ldots$ be a sequence of objects in $\mathfrak{U}$ (all of the same type). Let $x$ be another object in $\mathfrak{U}$ of the same type as the $x_n$.

- We say that the sequence $x_n$ is *elementarily Cauchy* if, for every predicate $P(x)$ that takes one variable of the same type as the $x_n$ as input, the truth value of $P(x_n)$ becomes eventually constant (i.e. either $P(x_n)$ is true for all sufficiently large $n$, or $P(x_n)$ is false for all sufficiently large $n$). We write this eventual truth value as $\lim_{n\to\infty} P(x_n)$.
- We say that the sequence $x_n$ is *elementarily convergent* to $x$ if we have $\lim_{n\to\infty} P(x_n) = P(x)$ for every predicate $P(x)$ that takes one variable of the same type as the $x_n$ or $x$ as input.

**Remark 4.4.2.** One can view the predicates $P$ (or more precisely, the sets $\{x \in \mathfrak{U} : P(x) \text{ true}\}$) as generating a topology on $\mathfrak{U}$ (or more precisely, on the domain of one of the object types of $L$ in $\mathfrak{U}$), in which case elementary convergence can be interpreted as convergence in this topology. Indeed, as there are only countably many predicates, this topology is metrisable.

To give an example, let us use the language of ordered fields $L$, with the model $\mathbf{R}$, and pick a transcendental number $x$, e.g. $x = \pi$. Then the sequence $x + \frac{1}{n}$ is elementarily convergent to $x$. The reason for this is that the language $L$ is fairly limited in nature, and as such it can only define a fairly small number of sets; in particular, if $P$ is a predicate of one variable, then the *Tarski-Seidenberg theorem* [**Ta1951**], [**Se1954**] tells us that the set $\{y \in \mathbf{R} : P(y) \text{ true}\}$ cut out by that set has to be a *semi-algebraic set* over the algebraic reals, i.e. a finite union of (possibly unbounded) intervals (which can be open, closed, or half-open) whose endpoints are algebraic reals. In particular, a transcendental number $x$, if it lies in such a set, lies in the interior of such a set, and so $x + \frac{1}{n}$ will also lie in such a set for $n$ large enough, and similarly if $x$ lies outside such a set.

In contrast, if one picks an algebraic number for $x$, such as $x = \sqrt{2}$, then $x + \frac{1}{n}$ does *not* converge in an elementary sense to $x$, because one can find a predicate such as $P(y) := (y^2 = 2)$ which is true for $x$ but not true for any of the $x + \frac{1}{n}$. So the language $L$ has sufficiently "poor vision" that it cannot

easily distinguish a transcendental number such as $\pi$ from nearby numbers such as $\pi + \frac{1}{n}$, but its vision is significantly better at algebraic numbers, and in particular can distinguish $\sqrt{2}$ from $\sqrt{2} + \frac{1}{n}$ easily. So we see that elementary convergence is, in this case, a slightly stronger concept than the usual topological or metric notion of convergence on $\mathbf{R}$.

In the case of the real model $\mathbf{R}$ of ordered fields, elementary limits are unique, but this is not the case in general. For instance, in the language of fields, and using the complex model $\mathbf{C}$, any given complex number $z$ is elementarily indistinguishable[16] from its complex conjugate $\overline{z}$, and so any sequence $z_n$ of complex numbers that would converge elementarily to $z$, would also converge elementarily to $\overline{z}$.

A related problem is that the operations on a structure $\mathfrak{U}$ are not necessarily continuous with respect to these elementary limits. For instance, if $x_n, y_n$ are sequences of real numbers that converge elementarily to $x, y$ respectively, it is not necessarily the case that $x_n + y_n$ converge to $x + y$ (consider for instance the case when $x_n = \pi + 1/n$ and $y_n = -\pi + 1/n$).

One way to partially resolve these problem is to consider the convergence not just of sequences of individual objects $x_n$, but of sequences of families $(x_{n,\alpha})_{\alpha \in A}$ of objects:

**Definition 4.4.3** (Joint elementary convergence)**.** Let $\mathfrak{U}$ be a structure a language $L$, let $A$ be a set, and for each natural number $n$, let $(x_{n,\alpha})_{\alpha \in A}$ be a tuple of of elements in $\mathfrak{U}$, and let $(x_\alpha)_{\alpha \in A}$ be another tuple in $\mathfrak{U}$, with each $x_{n,\alpha}$ having the same type as $x_\alpha$.

- We say that the tuples $(x_{n,\alpha})_{\alpha \in A}$ are *jointly elementarily Cauchy* if, for every natural number $m$, every predicate $P(y_1, \ldots, y_m)$ of $m$ variables in $L$ of the appropriate type, and every $\alpha_1, \ldots, \alpha_m \in A$, the truth value of $P(x_{n,\alpha_1}, \ldots, x_{n,\alpha_m})$ is eventually constant.
- We say that the tuples $(x_{n,\alpha})_{\alpha \in A}$ are *jointly elementarily convergent* to $(x_\alpha)_{\alpha \in A}$ if, for every natural number $m$, every predicate $P(y_1, \ldots, y_m)$ of $m$ variables in $L$ of the appropriate type, and every $\alpha_1, \ldots, \alpha_m \in A$, the truth value of $P(x_{n,\alpha_1}, \ldots, x_{n,\alpha_m})$ converges to the truth value of $P(x_{\alpha_1}, \ldots, x_{\alpha_m})$ as $n \to \infty$.

For instance, using the complex model $\mathbf{C}$ of the language of fields, if $z_n$ converges elementarily to (say) $i$, then we cannot prevent $z_n$ from also converging elementarily to $-i$. (Indeed, it is not hard to see that $z_n$ converges elementarily to $i$ if and only $z_n \in \{-i, +i\}$ for all sufficiently large $n$.) But if we ask that $(z_n, i)$ jointly converges to $(i, i)$, then $(z_n, i)$ will not also jointly converge to $(-i, i)$ (though it does jointly converge to $(-i, -i)$).

[16] In fact, there is an enormous Galois group $\mathrm{Gal}(\mathbf{C}/\overline{\mathbf{Q}})$, the action of which is completely undetectable with regards to elementary convergence.

In a similar fashion, if $x_n, y_n$ are reals that converge *jointly* elementarily to $x, y$, then $x_n + y_n$ will converge elemntarily to $x + y$ also.

Now we give a more sophisticated example. Here, $L$ is the language of set theory, and $\mathfrak{U}$ is a model of ZFC. In ZFC set theory, we can of course construct most of the objects we are used to in modern mathematics, such as a copy $\mathbf{R}_{\mathfrak{U}}$ of the real line, a copy $\mathbf{N}_{\mathfrak{U}}$ of the natural numbers, and so forth. Note that $\mathfrak{U}$'s interpretation $\mathbf{N}_{\mathfrak{U}}$ of the natural numbers may be different from the "true" natural numbers $\mathbf{N}$; in particular, in non-standard models of set theory, $\mathbf{N}_{\mathfrak{U}}$ may be much larger than $\mathbf{N}$ (e.g. it may be an ultrapower ${}^*\mathbf{N}$ of $\mathbf{N}$). Because of this, we will be careful to subscript $\mathfrak{U}$'s copies of such objects in order to distinguish them from their true counterparts, though it will not make much difference for this immediate example.

We can also define in $\mathfrak{U}$ all the formal apparatus needed for probability theory, such as a probability space $(\Omega, \mathcal{B}, \mathbf{P})$ and a real-valued random variable $X : \Omega \to \mathbf{R}_{\mathfrak{U}}$ on that space.

Now suppose that inside $\mathfrak{U}$ we have a sequence $(\Omega_n, \mathcal{B}_n, \mathbf{P}_n)$ of probability spaces, and a sequence $X_n : \Omega_n \to \mathbf{R}_{\mathfrak{U}}$ of random variables on these probability spaces. Now suppose the quintuple $(\Omega_n, \mathcal{B}_n, \mathbf{P}_n, X_n, \mathbf{R}_{\mathfrak{U}})$ is jointly elementarily convergent to a limit $(\Omega, \mathcal{B}, \mathbf{P}, X, \mathbf{R}_{\mathfrak{U}})$. The axioms of being a probability space can be encoded inside the first order language of set theory, so the limit $(\Omega, \mathcal{B}, \mathbf{P})$ is also a probability space (as viewed inside $\mathfrak{U}$). Similarly, $X : \Omega \to \mathbf{R}_{\mathfrak{U}}$ is a random variable on this probability space.

Now let $q, r$ be rational numbers. If $\mathbf{P}(X > q) > r$, then by the definition of elementary convergence (and the fact that rational numbers can be defined using an expression of finite length in the language $L$), we see that $\mathbf{P}_n(X_n > q) > r$ holds for all sufficiently large $n$. From this, one can deduce that the $X_n$ *converge in distribution* to $X$. Thus we see that in this case, joint elementary convergence is at least as strong as convergence in distribution, though (much as with the example with elementary convergence in $\mathbf{R}$ using the language of ordered fields) the two notions of convergence are not equivalent.

We included $\mathbf{R}_{\mathfrak{U}}$ in the quintuple due to the use of real numbers such as $\mathbf{P}(X > q)$ in the above discussion, but it is not strictly necessary, because one can construct $\mathbf{R}_{\mathfrak{U}}$ uniquely in $\mathfrak{U}$ from the axioms of set theory by using one of the standard constructions of the real numbers. But note that while we may use the set $\mathbf{R}_{\mathfrak{U}}$ of real numbers in the above elementary convergence, one cannot invoke specific real numbers unless they are "constructible" in the sense that they can be uniquely specified in the language $L$. If one wished to be able to use arbitrary real numbers as constants, one would not only place $\mathbf{R}_{\mathfrak{U}}$ into the quintuple, but also place in every element $x$ of $\mathbf{R}_{\mathfrak{U}}$ into the tuple (thus making the tuple quite large, and most likely uncountable, though note

from *Skolem's paradox* that it is possible for $\mathbf{R}_{\mathfrak{U}}$ to be (externally) countable even as it is uncountable from the internal perspective of $\mathfrak{U}$).

As we see from the above discussion, joint elementary convergence is a useful notion even when some of the elements in the tuple are constant. We isolate this case with another definition:

**Definition 4.4.4** (Joint relative elementary convergence)**.** Let $\mathfrak{U}$ be a structure for a language $L$, let $A, B$ be sets, and for each natural number $n$, let $(x_{n,\alpha})_{\alpha\in A}$ be a tuple of of elements in $\mathfrak{U}$, and let $(x_\alpha)_{\alpha\in A}$ and $(c_\beta)_{\beta\in B}$ be further tuples in $\mathfrak{U}$. We say that the tuples $(x_{n,\alpha})_{\alpha\in A}$ are *jointly elementarily convergent* to $(x_\alpha)_{\alpha\in A}$ *relative to the constants* $(c_\beta)_{\beta\in B}$ if the disjoint union $(x_{n,\alpha})_{\alpha\in A} \uplus (c_\beta)_{\beta\in B}$ is jointly elementarily convergent to $(x_\alpha)_{\alpha\in A} \uplus (c_\beta)_{\beta\in B}$.

We define the notion of $(x_{n,\alpha})_{\alpha\in A}$ are *jointly elementarily Cauchy relative to the constants* $(c_\beta)_{\beta\in B}$ similarly.

Informally, if $(x_{n,\alpha})_{\alpha\in A}$ is jointly elementarily convergent to $(x_\alpha)_{\alpha\in A}$ relative to $(c_\beta)_{\beta\in B}$ if the language $L$ is unable to asymptotically distinguish the $x_{n,\alpha}$ from the $x_\alpha$, even if it "knows" about the constants $c_\beta$.

**4.4.2. Elementary completion.** Not every sequence in a structure is elementarily Cauchy or elementarily convergent. However, we have the following simple fact:

**Proposition 4.4.5** (Arzelá-Ascoli)**.** *Let $\mathfrak{U}$ be a structure for a language $L$, and let $x_n$ be a sequence of elements in $\mathfrak{U}$ (all of the same type). Then there is a subsequence of the $x_n$ that is elementarily Cauchy.*

**Proof.** There are at most countably many predicates $P_1(x), P_2(x), \ldots$ of a single variable of the right type in $L$. By the infinite pigeonhole principle, we can find a subsequence $x_{n_{1,1}}, x_{n_{1,2}}, \ldots$ of the $x_n$ such that $P_1(x_{n_{1,i}})$ is eventually constant. We can find a further subsequence $x_{n_{2,1}}, x_{n_{2,2}}, \ldots$ of that sequence for which $P_2(x_{n_{2,i}})$ is eventually constant. We continue extracting subsequences $x_{n_{k,1}}, x_{n_{k,2}}, \ldots$ of this nature for each $k = 1, 2, \ldots$, and then the diagonal sequence $k \mapsto x_{n_{k,k}}$ is elementarily Cauchy, as desired. □

The same argument works when considering countably many variables and countably many constants (as there are still only countably many predicates to deal with):

**Proposition 4.4.6** (Arzelá-Ascoli)**.** *Let $\mathfrak{U}$ be a structure for a language $L$, let $A, B$ be at most countable sets, and for each natural number $n$, let $(x_{n,\alpha})_{\alpha\in A}$ be a tuple of of elements in $\mathfrak{U}$, and let $(c_\beta)_{\beta\in B}$ be another tuple in $\mathfrak{U}$. Then there is a subsequence $(x_{n_j,\alpha})_{\alpha\in A}$ which is jointly elementarily Cauchy relative to $(c_\beta)_{\beta\in B}$.*

As in the metric case, not every elementarily Cauchy sequence is elementarily convergent, and not every sequence has an elementarily convergent subsequence. For instance, in the language of ordered fields, using the structure $\mathbf{R}$, the sequence $\frac{1}{n}$ has no elementarily convergent subsequence (because any limit of such a subsequence would be positive but also less than $1/n$ for arbitrarily large $n$, contradicting the *Archimedean property* of the reals). From Proposition 4.4.5, we conclude that the reals are elementarily incomplete; there must exist some subsequence of $1/n$ that is elementarily Cauchy, but not elementarily convergent.

However, we can always complete any structure $\mathfrak{U}$ by passing to the *ultrapower* ${}^*\mathfrak{U}$, which we now briefly review. For the rest of this post, we fix a single *non-principal ultrafilter*[17] $\alpha_\infty \in \beta\mathbf{N}\backslash\mathbf{N}$ on the (standard) natural numbers $\mathbf{N}$. A property $P(\alpha)$ of a natural number $\alpha$ is said to hold *for all $\alpha$ sufficiently close to $\alpha_\infty$* if the set of $\alpha$ for which $P(\alpha)$ holds lies in the ultrafilter $\alpha_\infty$.

**Definition 4.4.7** (Ultrapower)**.** Let $\mathfrak{U}$ be a structure for some language $L$. Given two sequences $(x_\alpha)_{\alpha\in\mathbf{N}}$ and $(y_\alpha)_{\alpha\in\mathbf{N}}$ of objects in $\mathfrak{U}$, we say that the sequences are *equivalent* if one has $x_\alpha = y_\alpha$ for all $\alpha$ sufficiently close to $\alpha_\infty$. The equivalence class associated to a given sequence $(x_\alpha)_{\alpha\in\mathbf{N}}$ will be called the *ultralimit* of the $x_\alpha$ and denoted $\lim_{\alpha\to\alpha_\infty} x_\alpha$. The *ultrapower* ${}^*\mathfrak{U}$ of $\mathfrak{U}$ is the collection of all ultralimits $\lim_{\alpha\to\alpha_\infty} x_\alpha$ of sequences of objects in $\mathfrak{U}$. By identifying ${}^*x := \lim_{\alpha\to\alpha_\infty} x$ with $x$ for every object $x$ in $\mathfrak{U}$, we see that every object in $\mathfrak{U}$ can be identified with an object in ${}^*\mathfrak{U}$. We refer to elements of $\mathfrak{U}$ as *standard objects*, and elements of ${}^*\mathfrak{U}$ as *non-standard objects*.

Every relation and operation in $\mathfrak{U}$ can be extended to ${}^*\mathfrak{U}$ by taking ultralimits. For instance, given a $k$-ary relation $R(y_1,\ldots,y_k)$, and non-standard objects $x_i = \lim_{\alpha\to\alpha_\infty} x_{i,\alpha}$ for $i = 1,\ldots,k$, we say that $R(x_1,\ldots,x_k)$ holds in ${}^*\mathfrak{U}$ if and only if $R(x_{1,\alpha},\ldots,x_{k,\alpha})$ holds in $\mathfrak{U}$ for all $\alpha$ sufficiently close to $\alpha_\infty$. Similarly, given a $k$-ary operation $f(y_1,\ldots,y_k)$ and non-standard objects $x_i = \lim_{\alpha\to\alpha_\infty} x_{i,\alpha}$, we define the non-standard object $f(x_1,\ldots,x_k)$ to be the ultralimit $\lim_{\alpha\to\alpha_\infty} f(x_{1,\alpha},\ldots,x_{k,\alpha})$ of the standard objects $f(x_{1,\alpha},\ldots,x_{k,\alpha})$.

A fundamental theorem of Łos [**Lo1955**] asserts that the ultrapower ${}^*\mathfrak{U}$ is *elementarily equivalent* to $\mathfrak{U}$: any sentence in $L$ which is true in $\mathfrak{U}$, is also true in ${}^*\mathfrak{U}$, and vice versa; this fact is also known as the *transfer principle* for nonstandard analysis. For instance, the ultrapower of an ordered field is an ordered field, the ultrapower of an algebraically closed field is an algebraically closed field, and so forth. One must be slightly careful, though,

[17] See [**Ta2008**, §1.5] for some basic discussion of what non-principal ultrafilters are, and how they are used in non-standard analysis.

with models $\mathfrak{U}$ that involve standard objects such as a copy $\mathfrak{U}$ of the natural numbers, or a copy $\mathbf{R}_{\mathfrak{U}}$ of the real numbers; the ultrapower ${}^*\mathfrak{U}$ will have their own non-standard copy $\mathbf{N}_{{}^*\mathfrak{U}} = {}^*\mathbf{N}_{\mathfrak{U}}$ and $\mathbf{R}_{{}^*\mathfrak{U}} = {}^*\mathbf{R}_{\mathfrak{U}}$ of these objects, which are considerably larger than their standard counterparts, in the sense that they contain many more elements. Thus, for instance, if one is taking the ultrapower of a standard probability space $(\Omega, \mathcal{B}, \mathbf{P})$, in which the probability measure $\mathbf{P} : \mathcal{B} \to \mathbf{R}$ takes values in the standard reals, the ultrapower $({}^*\Omega, {}^*\mathcal{B}, {}^*\mathbf{P})$ is a non-standard probability space, in which the non-standard probability measure ${}^*\mathbf{P} : {}^*\mathcal{B} \to {}^*\mathbf{R}$ now takes values in the non-standard reals.

One can view the ultrapower ${}^*\mathfrak{U}$ as the completion of $\mathfrak{U}$, in much the same way as the reals are a completion of the rationals:

**Theorem 4.4.8** (Elementary completeness)**.** *Every elementarily Cauchy sequence $x_n$ in an ultrapower ${}^*\mathfrak{U}$ is elementarily convergent.*

This property is also known as *countable saturation*.

**Proof.** We can write $x_n = \lim_{\alpha \to \alpha_\infty} x_{n,\alpha}$ for each natural number $n \in \mathbf{N}$ and a sequence $x_{n,\alpha}$ of standard objects in $\mathfrak{U}$. As before, we enumerate the predicates $P_1, P_2, P_3, \ldots$ of one variable. For each natural number $m \in \mathbf{N}$, the truth value of $P_m(x_n)$ becomes eventually constant; we will call this constant $\lim_{n\to\infty} P_m(x_n)$.

Now let $M$ be a standard natural number. By construction, there exists an $n_M$ such that

$$P_m(x_{n_M}) = \lim_{n\to\infty} P_m(x_n)$$

for all $1 \leq m \leq M$. As $x_{n_M}$ is the ultralimit of the $x_{n_M,\alpha}$, there thus exists a set $E_M \in p$ such that

$$P_m(x_{n_M,\alpha}) = \lim_{n\to\infty} P_m(x_n)$$

for all $\alpha \in E_M$. By replacing each $E_M$ with $\bigcap_{M' \leq M} E_{M'}$ if necessary, we may assume that the $E_M$ are decreasing: $E_1 \supset E_2 \supset \ldots$.

For each $\alpha \in \mathbf{N}$, let $M_\alpha$ be the largest integer in $\{0, \ldots, \alpha\}$ such that $\alpha \in E_{M_\alpha}$, or $M_\alpha = 0$ if no such integer exists. By construction, we see that for any $m \in \mathbf{N}$, we haved

$$P_m(x_{n_{M_\alpha},\alpha}) = \lim_{n\to\infty} P_m(x_n)$$

whenever $\alpha \in E_m$ and $\alpha \geq m$. If we then set $x$ to be the non-standard object $x := \lim_{\alpha\to\alpha_\infty} x_{n_{M_\alpha},\alpha}$, we thus have

$$P_m(x) = \lim_{n\to\infty} P_m(x_n)$$

for each $m \in \mathbf{N}$, and thus $x_n$ converges elementarily to $x$ as required. □

Combining this theorem with Proposition 4.4.5 we conclude an analogue of the Bolzano-Weierstrass theorem for ultrapowers:

**Corollary 4.4.9** (Bolzano-Weierstrass for ultrapowers)**.** *In an ultrapower ${}^*\mathfrak{U}$, every sequence $x_n$ of non-standard objects in ${}^*\mathfrak{U}$ has an elementarily convergent subsequence $x_{n_j}$.*

The same argument works (but with more complicated notation) for countable families of objects, and with countably many constants:

**Theorem 4.4.10** (Bolzano-Weierstrass for ultrapowers, II)**.** *Let ${}^*\mathfrak{U}$ be an ultrapower, let $A, B$ be at most countable, let $n \mapsto (x_{n,\alpha})_{\alpha\in A}$ be a sequence of tuples of nonstandard objects in ${}^*\mathfrak{U}$, and let $(c_\beta)_{\beta\in B}$ be another sequence of tuples of nonstandard objects. Then there is a subsequence $(x_{n_j,\alpha})_{\alpha\in A}$ which converges jointly elementarily to a limit $(x_\alpha)_{\alpha\in A}$ relative to the constants $(c_\beta)_{\beta\in B}$.*

The proof of this theorem proceeds almost exactly as in the single variable case, the key point being that the number of predicates that one has to stabilise remains countable.

**Remark 4.4.11.** If one took the ultrafilter $\alpha_\infty$ over a larger set than the natural numbers $\mathbf{N}$, then one could make the sets $A, B$ larger as well. Such larger saturation properties, beyond countable saturation, are useful in model theory (particularly when combined with the use of large cardinals, such as inaccessible cardinals), but we will not need them here.

Conversely, every nonstandard object can be viewed as the elementary limit of standard objects:

**Proposition 4.4.12.** *Let $x \in {}^*\mathfrak{U}$ be a nonstandard object. Then there is a sequence $x_n \in \mathfrak{U}$ of standard objects that converges elementarily to $x$.*

**Proof.** Let $P_1, P_2, \ldots$ be an enumeration of the predicates of one variable. For any natural number $n$, there exists a nonstandard object $y \in {}^*\mathfrak{U}$ such that $P_i(y)$ has the same truth value as $P_i(x)$ for all $i = 1, \ldots, n$, namely $y = x$. By transfer, there must therefore exist a standard object $x_n \in \mathfrak{U}$ such that $P_i(x_n)$ has the same truth value as $P_i(x)$. Thus $x_n$ converges elementarily to $x$, and the claim follows. □

**Exercise 4.4.1.** If $x$ is the ultralimit of a sequence $x_n$ of standard objects, show that there is a subsequence $x_{n_j}$ that converges elementarily to $x$.

**Exercise 4.4.2** (Heine-Borel theorem for structures)**.** Given any structure $\mathfrak{U}$, show that the following four statements are equivalent:

- (Countable saturation) If $P_1(x), P_2(x), \ldots$ are a countable family of predicates, such that if any finite number of $P_i$ are simultaneously satisfiable in $\mathfrak{U}$ (i.e. for each $n$ there exists $x_n \in \mathfrak{U}$ such that $P_i(x_n)$ holds for all $i = 1, \ldots, n$), then the entire family of $P_i$ are simultaneously satisfiable (i.e. there exists $x \in \mathfrak{U}$ such that $P_i(x)$ holds for all $i$).
- (Countable compactness) Every countable cover of $\mathfrak{U}$ by sets of the form $\{x \in \mathfrak{U} : P(x) \text{ true}\}$ for some predicate $P$, has a finite subcover.
- (Elementary completeness) Every elementarily Cauchy sequence in $\mathfrak{U}$ has an elementarily convergent subsequence.
- (Bolzano-Weierstrass property) Every elementary sequence in $\mathfrak{U}$ has an elementarily convergent subsequence.

From Proposition 4.4.12 and Theorem 4.4.8 we see that ${}^*\mathfrak{U}$ can be viewed as an elementary completion of $\mathfrak{U}$, though the analogy with metric completion is not perfect because elementary limits are not unique.

The Bolzano-Weierstrass theorem for ultrapowers can then be used to derive the foundational properties of nonstandard analysis. For instance, consider the standard natural numbers $1, 2, 3, \ldots$ in $\mathbf{N}$, and hence in ${}^*\mathbf{N}$. Applying the Bolzano-Weierstrass theorem for ultraproducts, we conclude that some subsequence $n_j$ of natural numbers will converge elementarily to a non-standard natural number $N \in {}^*\mathbf{N}$. For any standard natural number $m \in \mathbf{N}$, we have $n_j > m$ for all sufficiently large $j$, and hence on taking elementary limits we have $N > m$ (since $m$ is constructible). Thus we have constructed an *unbounded* nonstandard natural number, i.e. a number which is larger than all standard natural numbers.

In a similar spirit, we can also use the Bolzano-Weierstrass theorem construct *infinitesimal* nonstandard real numbers $0 < \varepsilon = o(1)$ which are positive, but less than every standard positive real number (and in particular less than $1/n$ for any standard $n$).

More generally, we have the *overspill principle*: if $P(n, c_1, \ldots, c_k)$ is a predicate involving some non-standard constants $c_1, \ldots, c_k$, such that $P(n, c_1, \ldots, c_k)$ is true for arbitrarily large standard natural numbers $n \in \mathbf{N}$, then it must also be true for at least one unbounded nonstandard natural number $n \in {}^*\mathbf{N}$. Indeed, one simply takes a sequence $n_j \in \mathbf{N}$ of standard natural numbers for which $P(n_j, c_1, \ldots, c_k)$, and extracts a subsequence of these $n_j$ which converges elementarily to a non-standard limit $n$ (relative to $c_1, \ldots, c_k$), which must then be unbounded. Contrapositively, if $P(n, c_1, \ldots, c_k)$ holds for all unbounded $n$, then it must also hold for all sufficiently large standard $n$.

Similarly, we have the *underspill principle*: if a predicate $P(x, c_1, \ldots, c_k)$ is true for arbitrarily small positive standard real $x$, then it must also be true for at least one infinitesimal positive non-standard real $x$; and contrapositively, if it is true for all infinitesimal positive non-standard real $x$, then it is also true for all sufficiently small standard real $x$.

A typical application of these principles is in the nonstandard formulation of continuity:

**Proposition 4.4.13** (Nonstandard formulation of continuity)**.** *Let $f : \mathbf{R} \to \mathbf{R}$ be a standard function, which can then be extended by ultralimits to the nonstandard completion ${}^*f : {}^*\mathbf{R} \to {}^*\mathbf{R}$. Let $x \in \mathbf{R}$. Then the following are equivalent:*

- *$f$ is continuous at $x$.*
- *One has ${}^*f(y) = f(x) + o(1)$ whenever $y$ is a nonstandard real such that $y = x + o(1)$.*

**Proof.** If $f$ is continuous, then the "epsilon-delta" definition implies that whenever $y = x + o(1)$ (so that $|y - x| < \delta$ for every standard $\delta > 0$), one has $|{}^*f(y) - f(x)| < \varepsilon$ for every standard $\varepsilon$ (by transfer), and thus ${}^*f(y) = f(x) + o(1)$. Conversely, if $f$ is discontinuous at $x$, then there exists a sequence $x_n$ of standard reals converging to $x$ such that $|f(x_n) - f(x)| > \varepsilon$ for some standard $\varepsilon > 0$; taking ultralimits using Bolzano-Weierstrass to extract a subsequence that is elementarily convergent relative to $f$, we thus have a non-standard $y = x + o(1)$ with $|{}^*f(y) - f(x)| > \varepsilon$, so that ${}^*f(y) \neq f(x) + o(1)$. □

**Exercise 4.4.3.** With the notation as above, show that $f$ is uniformly continuous if and only if ${}^*f(y) = {}^*f(x) + o(1)$ whenever $x, y \in {}^*\mathbf{R}$ are such that $y = x + o(1)$.

**Exercise 4.4.4.** With the notation as above, show that $f$ is differentiable at a standard real $x$ with derivative $f'(x)$ if and only if ${}^*f(x + h) = f(x) + hf'(x) + o(h)$ for all nonstandard reals $h = o(1)$.

**4.4.3. The correspondence principle.** One can use the Bolzano-Weierstrass theorem for ultrapowers to establish various versions of the correspondence principle, which was discussed extensively in Section 4.3. A simple example occurs when demonstrating the equivalence of colouring theorems, such as the following:

**Theorem 4.4.14** (van der Waerden theorem, infinitary version)**.** *Suppose the integers are coloured by finitely many colours. Then there exist arbitrarily long monochromatic arithmetic progressions.*

**Theorem 4.4.15** (van der Waerden theorem, finitary version)**.** *For every $c$ and $k$ there exists $N$ such that whenever $\{-N, \ldots, N\}$ is coloured by $c$ colours, there exists a monochromatic arithmetic progression of length $k$.*

It is clear that Theorem 4.4.15 implies Theorem 4.4.14. To deduce Theorem 4.4.14 from Theorem 4.4.15, we can argue as follows. Suppose Theorem 4.4.15 fails, then there exists $c$ and $k$, and arbitrarily large standard natural number $N$ for which there exists a $c$-colouring of $\{-N, \ldots, N\}$ without any monochromatic arithmetic progressions of length $k$. Applying the overspill principle (or the Bolzano-Weierstrass theorem), there must then also exist an unbounded nonstandard natural number for which there exists a $c$-colouring of $\{-N, \ldots, N\}$ without any monochromatic arithmetic progressions of length $k$. But the nonstandard interval $\{-N, \ldots, N\}$ contains the standard integers $\mathbf{Z}$ as a subset, thus the integers can now also be $c$-coloured without any monochromatic arithmetic progressions of length $k$, contradicting Theorem 4.4.14.

As another example, we can relate qualitative and quantitative results in algebraic geometry. For instance, the following basic result in algebraic geometry,

**Theorem 4.4.16** (Qualitative decomposition into varieties)**.** *Every algebraic set over an algebraically closed field can be decomposed into finitely many algebraic varieties.*

is equivalent to the following more quantitative version:

**Theorem 4.4.17** (Quantitative decomposition into varieties)**.** *Every algebraic set $A$ of complexity at most $M$ over an algebraically closed field $k$ can be decomposed at most $C_M$ algebraic varieties, each of complexity at most $C_M$, where $C_M$ depends only on $M$.*

Here, we say that an (affine) algebraic set has complexity at most $M$ if it lies in a space $k^n$ of dimension at most $M$, and is defined as the zero locus of at most $M$ polynomials, each of degree at most $M$.

Clearly Theorem 4.4.17 implies Theorem 4.4.16. To show the converse implication, suppose that Theorem 4.4.17 failed, then there exists $M$ such that for every standard natural number $N$ there exists an algebraic set $A_N$ of complexity at most $M$ over some field $k_N$ that cannot be decomposed into fewer than $N$ algebraic varieties. We use the Bolzano-Weierstrass theorem for ultrapowers to extract a subsequence $A_{N_j}, k_{N_j}$ that converges jointly elementarily to some limit $A, k$. As each of the $k_{N_j}$ are algebraically closed fields, the elementary limit $k$ is also. As the $A_{N_j}$ were algebraic sets over $k_{N_j}$ of uniformly bounded complexity, $A$ is an algebraic set over $k$, and thus by Theorem 4.4.16 is decomposable into at most $N_0$ algebraic varieties

for some finite $N_0$. The property of being decomposable into at most $N_0$ algebraic varieties can be phrased in an elementarily open manner, i.e. as the disjunction of sentences in first-order logic; see [**Ta2011b**, §2.1], with the key point being that any top-dimensional component of an algebraic set has a lesser degree than that of the original set, and so has a uniform bound on complexity. Thus, we see that for all sufficiently large $j$, $A_{N_j}$ must also be decomposable into at most $N_0$ algebraic varieties, a contradiction.

Our final example, namely the Furstenberg correspondence principle, is a bit more sophisticated. Here, we are demonstrating the equivalence of the following two statements:

**Theorem 4.4.18** (Furstenberg recurrence theorem)**.** *Let $(X, \mathcal{B}, \mu, T)$ be a measure-preserving system, and let $A \subset X$ have positive measure. Let $k \geq 1$. Then there exists $r > 0$ such that $A \cap T^r A \cap \ldots \cap T^{(k-1)r} A$ has positive measure.*

**Theorem 4.4.19** (Szemerédi's theorem)**.** *Every set of integers of positive upper density contains arbitrarily long arithmetic progressions.*

It is easy to use Theorem 4.4.19 to show Theorem 4.4.18, so we focus on the reverse inclusion. Suppose that Theorem 4.4.19 failed, then there exists a standard integer $k \geq 1$, a standard real $\delta > 0$, and a standard set $A \subset \mathbf{Z}$ of integers of positive upper density at least $\delta$ that has no arithmetic progressions of length $k$. In particular, for arbitrarily large standard $N$, there exists a subset of $\{-N, \ldots, N\}$ of density at least $\delta$ without any progressions of length $k$. Applying the overspill principle, there thus exists an unbounded nonstandard $N$ and a nonstandard subset $A$ of $\{-N, \ldots, N\}$ of density at least $\delta$ without any progressions of length $k$.

Let $\mathcal{A}$ be the collection of all nonstandard subsets[18] of $\{-N, \ldots, N\}$. This is a Boolean algebra. From countable saturation we see that this Boolean algebra has the following special property: if any set $E$ in this Boolean algebra is partitioned into an (externally) countable family $(E_n)_{n \in \mathbf{N}}$ of further elements in this Boolean algebra, then all but finitely many of the $E_n$ are empty. For if this were not the case, then by the axiom of choice, one could find a subsequence $n_j$ and a set of elements $x_{n_j}$ of $E_{n_j}$. Passing to a further subsequence, we can assume that the $x_{n_j}$ converge elementarily to a limit $x$. But then this limit lies in $E$ but not in any of the $E_n$, a contradiction.

From the above property, we see that any (external) finitely additive measure $\mu : \mathcal{A} \to [0, 1]$ on $\mathcal{A}$ is automatically a *premeasure*, and thus by the

[18] Note that not every subset of a nonstandard set remains nonstandard; it may instead merely be an external subset. See [**Ta2008**, §1.5] for further discussion.

*Hahn-Kolmogorov extension theorem* (see e.g. [**Ta2011**, §1.7]), can be extended to a countably additive measure on the *measure-theoretic completion* $\overline{\langle \mathcal{A} \rangle}$ of the (external) $\sigma$-algebra $\langle \mathcal{A} \rangle$ generated by $\mathcal{A}$.

In particular, if we consider the nonstandard normalised counting measure

$$E \mapsto \frac{\#E}{2N+1}$$

on $\mathcal{A}$ and take its standard part,

$$\mu : E \mapsto \mathrm{st}\left(\frac{\#E}{2N+1}\right)$$

this is a finitely additive probability measure on $\mathcal{A}$, and hence extends to a probability measure in $\mathcal{B} := \overline{\langle \mathcal{A} \rangle}$, which we will continue to call $\mu$. This measure is known as the *Loeb measure* on the nonstandard set $\{-N, \ldots, N\}$. Observe that any nonstandard subset of $\{-N, \ldots, N\}$ of infinitesimal density will have Loeb measure zero. On the other hand, the set $A$ had density at least $\delta$, and so will have Loeb measure at least $\delta$ also.

Next, we define the shift $T : \{-N, \ldots, N\} \to \{-N, \ldots, N\}$ by $Tx := x+1$, leaving $T$ undefined for $x = N$. But observe that $\{N\}$ has an infinitesimal density, hence has Loeb measure zero. So $T$ is defined almost everywhere, and is easily seen to be measurable and measure-preserving; it has an (almost everywhere defined) inverse that is also measurable and measure-preserving. Thus $\{-N, \ldots, N\}$ with Loeb measure and the shift $T$ becomes a measure-preserving system. Applying Theorem 4.4.18, we can thus find a standard $r > 0$ such that $A \cap T^r A \cap \ldots \cap T^{(k-1)r} A$ has positive Loeb measure, so $A$ contains a $k$-term arithmetic progression, a contradiction.

**Remark 4.4.20.** As stated, the measure space structure on $\{-N, \ldots, N\}$ is not separable (i.e. countably generated) or regular (coming from a metric space). However, this can be fixed by restricting attention to the much smaller $\sigma$-algebra generated by $A$ and its shifts (after dealing with the null sets on which $T$ is not defined, e.g. by cutting out the $o(N)$ neighbourhood of $\{-N, N\}$). We omit the details.

**4.4.4. The Szemerédi regularity lemma.** Finally, we use nonstandard analysis to give a proof of the *Szemerédi regularity lemma* (see e.g. [**Sz1978**]), which we phrase as follows:

**Lemma 4.4.21** (Szemerédi regularity lemma)**.** *Let $G = (V, E)$ be a finite graph, and let $\varepsilon > 0$. Then there exists a vertex partition $V = V_1 \cup \ldots \cup V_m$ with $m \leq C(\varepsilon)$ such that for all pairs $(i, j)$ outside of a bad set $S$ with $\sum_{(i,j) \in S} |V_i||V_j| \leq \varepsilon |V|^2$, there exists a density $d_{ij}$ such that*

$$|E(A, B) - d_{ij}|V_i||V_j|| \leq \varepsilon |V_i||V_j|$$

*for all $A \subset V_i$ and $B \subset V_j$, where $E(A,B) := |E \cap (A \times B)|$, viewing $E$ as a symmetric subset of $V \times V$.*

Here we do not make the cells $V_i$ in the partition of equal size, as is customary, but it is not too difficult to obtain this additional property from the above formulation of the lemma. The ability of nonstandard analysis to establish regularity lemmas was first observed by Elek and Szegedy [**ElSz2007**].

An application of the Bolzano-Weierstrass theorem for ultraproducts shows that this lemma is equivalent to the following nonstandard version:

**Lemma 4.4.22** (Szemerédi regularity lemma, nonstandard formulation)**.** *Let $G = (V,E)$ be a nonstandard finite graph, and let $\varepsilon > 0$. Then there exists a vertex partition $V = V_1 \cup \ldots \cup V_m$, where $m$ is a standard natural number and $V_1, \ldots, V_m$ are nonstandard subsets of $V$ such that for all pairs $(i,j)$ outside of a bad set $S$ with $\sum_{(i,j)\in S} |V_i||V_j| \leq \varepsilon |V|^2$, there exists a standard density $d_{ij}$ such that*

$$|E(A,B) - d_{ij}|V_i||V_j|| \leq \varepsilon |V_i||V_j|$$

*for all $A \subset V_i$ and $B \subset V_j$.*

To see why Lemma 4.4.22 implies Lemma 4.4.21, suppose that Lemma 4.4.21 failed. Then there is a standard $\varepsilon > 0$ such that for every standard $m$ one could find a standard finite graph $G_m = (V_m, E_m)$ which could not be regularised into $m$ or fewer cells as required by the lemma. Applying the Bolzano-Weierstrass theorem for ultraproducts, we can assume that $G_m$ converges elementarily (relative to $\varepsilon$) to a limit $G = (V,E)$, which is then a nonstandard finite graph $G = (V,E)$ which cannot be regularised into any standard finite number of cells. But this contradicts Lemma 4.4.22.

It remains to prove Lemma 4.4.22. Let $\mu_V : \mathcal{B}_V \to [0,1]$ be Loeb measure on $V$ (as constructed in the previous section), and $\mu_{V\times V} : \mathcal{B}_{V\times V} \to [0,1]$ be Loeb measure on $V \times V$. It is easy to see that $\mathcal{B}_{V\times V}$ contains the product $\sigma$-algebra $\mathcal{B}_V \times \mathcal{B}_V$ as a subalgebra, and that the product measure $\mu_V \times \mu_V$ is the restriction of $\mu_{V\times V}$ to $\mathcal{B}_V \times \mathcal{B}_V$. The edge set $E$, viewed as a symmetric nonstandard subset of $V \times V$, is measurable in $\mathcal{B}_{V\times V}$, but is not necessarily measurable in $\mathcal{B}_V \times \mathcal{B}_V$. One can then form the conditional expectation $f := \mathbf{E}(1_E|\mathcal{B}_V \times \mathcal{B}_V)$, which is a $\mathcal{B}_V \times \mathcal{B}_V$-measurable function that is defined up to $\mu_V \times \mu_V$-almost everywhere equivalence, and takes values in $[0,1]$.

The $\sigma$-algebra $\mathcal{B}_V \times \mathcal{B}_V$ is generated by product sets $A \times B$ of $\mathcal{B}_V$-measurable functions, which in turn can be approximated in measure to arbitrary accuracy by product sets of nonstandard sets. As $f$ is $\mathcal{B}_V \times \mathcal{B}_V$-measurable, we can approximate it to less than $\varepsilon^2$ in $L^1(\mu_V \times \mu_V)$ norm by a finite linear combination of indicator functions of products of nonstandard

sets. Organising these products, we thus see that

$$\left\| f - \sum_{i=1}^{m} \sum_{j=1}^{m} d_{ij} 1_{V_i \times V_j} \right\|_{L^1(\mu_V \times \mu_V)} < \varepsilon^2$$

for some finite partition of $V$ into nonstandard sets $V_1, \ldots, V_m$ and some standard real numbers $d_{ij} \in [0,1]$. By Markov's inequality, we thus see that

$$\|f - d_{ij}\|_{L^1(V_i \times V_j)} < \varepsilon \mu_V(V_i) \mu_V(V_j)$$

for all $(i,j)$ outside of a bad set $S$ with

$$\sum_{S} \mu_V(V_i) \mu_V(V_j) \leq \varepsilon.$$

Now let $A \subset V_i$ and $B \subset V_j$ be nonstandard sets, with $(i,j)$ outside of $S$. Then

$$\|f 1_{A \times B} - d_{ij} 1_{A \times B}\|_{L^1(V_i \times V_j)} < \varepsilon \mu_V(V_i) \mu_V(V_j).$$

On the other hand, $1_E - f$ is orthogonal to all $\mathcal{B}_V \times \mathcal{B}_V$ functions, and in particular to $1_{A \times B}$, and thus

$$\int_{V \times V} 1_E 1_{A \times B} \, d\mu_{V \times V} = \int_{V \times V} f 1_{A \times B} \, d\mu_{V \times V}.$$

Since

$$\int_{V \times V} 1_E 1_{A \times B} \, d\mu_{V \times V} = \frac{|E(A,B)|}{|V|^2}$$

and

$$\int_{V \times V} d_{ij} 1_{A \times B} \, d\mu_{V \times V} = d_{ij} \frac{|A||B|}{|V|^2}$$

and

$$\mu_V(V_i) = |V_i|/|V|; \mu_V(V_j) = |V_j|/|V|$$

we thus see that

$$\left| \frac{|E(A,B)|}{|V|^2} - d_{ij} \frac{|A||B|}{|V|^2} \right| < \varepsilon \frac{|V_i||V_j|}{|V|^2}.$$

Thus $G$ has been regularised using a finite number of cells, as required.

## 4.5. Concentration compactness via nonstandard analysis

One of the key difficulties in performing analysis in infinite-dimensional function spaces, as opposed to finite-dimensional vector spaces, is that the *Bolzano-Weierstrass theorem* no longer holds: a bounded sequence in an infinite-dimensional function space need not have any convergent subsequences (when viewed using the strong topology). To put it another way, the closed unit ball in an infinite-dimensional function space usually fails to be (sequentially) compact.

As compactness is such a useful property to have in analysis, various tools have been developed over the years to try to salvage some sort of substitute for the compactness property in infinite-dimensional spaces. One of these tools is *concentration compactness*, which was discussed in [**Ta2009b**, §1.6]. This can be viewed as a compromise between weak compactness (which is true in very general circumstances, but is often too weak for applications) and strong compactness (which would be very useful in applications, but is usually false), in which one obtains convergence in an intermediate sense that involves a group of symmetries acting on the function space in question.

Concentration compactness is usually stated and proved in the language of standard analysis: epsilons and deltas, limits and supremas, and so forth. In this post, I wanted to note that one could also state and prove the basic foundations of concentration compactness in the framework of nonstandard analysis, in which one now deals with infinitesimals and ultralimits instead of epsilons and ordinary limits. This is a fairly mild change of viewpoint, but I found it to be informative to view this subject from a slightly different perspective. The nonstandard proofs require a fair amount of general machinery to set up, but conversely, once all the machinery is up and running, the proofs become slightly shorter, and can exploit tools from (standard) infinitary analysis, such as orthogonal projections in Hilbert spaces, or the continuous-pure point decomposition of measures. Because of the substantial amount of setup required, nonstandard proofs tend to have significantly more net complexity than their standard counterparts when it comes to basic results (such as those presented in this section), but the gap between the two narrows when the results become more difficult, and for particularly intricate and deep results it can happen that nonstandard proofs end up being simpler overall than their standard analogues, particularly if the nonstandard proof is able to tap the power of some existing mature body of infinitary mathematics (e.g. ergodic theory, measure theory, Hilbert space theory, or topological group theory) which is difficult to directly access in the standard formulation of the argument.

**4.5.1. Weak sequential compactness in a Hilbert space.** Before turning to concentration compactness, we will warm up with the simpler situation of weak sequential compactness in a Hilbert space. For sake of notation we shall only consider complex Hilbert spaces, although all the discussion here works equally well for real Hilbert spaces.

Recall that a bounded sequence $x_n$ of vectors in a Hilbert space $H$ is said to *converge weakly* to a limit $x$ if one has $\langle x_n, y\rangle \to \langle x, y\rangle$ for all $y \in H$. We have the following basic theorem:

**Theorem 4.5.1** (Sequential Banach-Alaoglu theorem)**.** *Every bounded sequence $x_n$ of vectors in a Hilbert space $H$ has a weakly convergent subsequence.*

The usual (standard analysis) proof of this theorem runs as follows:

**Proof.** (Sketch) By restricting to the closed span of the $x_n$, we may assume without loss of generality that $H$ is separable. Letting $y_1, y_2, \ldots$ be a dense subet of $H$, we may apply the Bolzano-Weierstrass theorem iteratively, followed by the *Arzelá-Ascoli diagonalisation argument*, to find a subsequence $x_{n_j}$ for which $\langle x_{n_j}, y_m \rangle$ converges to a limit for each $m$. Using the boundedness of the $x_{n_j}$ and a density argument, we conclude that $\langle x_{n_j}, y \rangle$ converges to a limit for each $y$; applying the *Riesz representation theorem for Hilbert spaces*, (see e.g. [**Ta2010**, §1.4]) the limit takes the form $\langle x, y \rangle$ for some $x$, and the claim follows. □

However, this proof does not extend easily to the concentration compactness setting, when there is also a group action. For this, we need a more "algorithmic" proof based on the "energy increment method". We give one such (standard analysis) proof as follows:

**Proof.** As $x_n$ is bounded, we have some bound of the form

$$\limsup_{n\to\infty} \|x_n\|^2 \leq E$$

for some finite $E$. Of course, this bound would persist if we passed from $x_n$ to a subsequence.

Suppose for contradiction that no subsequence of $x_n$ was weakly convergent. In particular, $x_n$ itself was not weakly convergent, which means that there exists $y_1 \in H$ for which $\langle x_n, y_1 \rangle$ did not converge. We can take $y_1$ to be a unit vector. Applying the Bolzano-Weierstrass theorem, we can pass to a subsequence (which, by abuse of notation, we continue to call $x_n$) in which $\langle x_n, y_1 \rangle$ converged to some non-zero limit $c_1$. We can choose $c_1$ to be nearly maximal in magnitude among all possible choices of subsequence and of $y_1$; in particular, we have

$$\limsup_{n\to\infty} |\langle x_n, y \rangle| \leq 2|c_1|$$

(say) for all other choices of unit vector $y$.

We may now decompose

$$x_n = c_1\phi_1 + x'_{1,n} + w_{1,n}$$

where $x'_{1,n}$ is orthogonal to $\phi_1$ and $w_{1,n}$ converges strongly to zero. From Pythagoras theorem we see that $x'_{1,n}$ asymptotically has strictly less energy

than $E$:

$$\limsup_{n\to\infty} \|x'_{1,n}\|^2 \leq E - |c_1|^2.$$

If $x'_{1,n}$ was weakly convergent, then $x_n$ would be too, so we may assume that it is not weakly convergent. Arguing as before, we may find a unit vector $\phi_2$ (which we can take to be orthogonal to $\phi_1$) and a constant $c_2$ such that (after passing to a subsequence, and abusing notation once more) one had a decomposition

$$x'_{1,n} = c_2\phi_2 + x'_{2,n} + w_{2,n}$$

in which $x'_{2,n}$ is orthogonal to both $\phi_1, \phi_2$ and $w_{2,n}$ converges strongly to zero, and such that

$$\limsup_{n\to\infty} |\langle x'_{1,n}, y\rangle| \leq 2|c_2|$$

for all unit vectors $y$. From Pythagoras, we have

$$\limsup_{n\to\infty} \|x'_{2,n}\|^2 \leq E - |c_1|^2 - |c_2|^2.$$

We iterate this process to obtain an orthonormal sequence $\phi_1, \phi_2, \ldots$ and constants $c_1, c_2, \ldots$ obeying the Bessel inequality

$$\sum_{k=1}^{\infty} |c_k|^2 \leq E$$

(which, in particular, implies that the $c_k$ go to zero as $k \to \infty$) such that, for each $k$, one has a subsequence of the $x_n$ for which one has a decomposition of the form

$$x_n = \sum_{i=1}^{k-1} c_i\phi_i + x'_{k,n} + w_{k,n}$$

where $w_{k,n}$ converges strongly to zero, and for which

$$\limsup_{n\to\infty} |\langle x'_{k,n}, y\rangle| \leq 2|c_{k+1}|$$

for all unit vectors $y$. The series $\sum_{i=1}^{\infty} c_i\phi_i$ then converges (conditionally in the strong topology) to a limit $x$, and by diagonalising all the subsequences we obtain a final subsequence $x_{n_j}$ which converges weakly to $x$. □

Now we give a third proof, which is a nonstandard analysis proof that is analogous to the second standard analysis proof given above.

The basics of nonstandard analysis are reviewed in [**Ta2008**, §1.5] (see also [**Ta2011b**, §2.1]), as well as Section 4.4. Very briefly, we will need to fix a non-principal *ultrafilter* $p \in \beta\mathbf{N}\backslash\mathbf{N}$ on the natural numbers. Once one fixes this ultrafilter, one can define the *ultralimit* $\lim_{n\to p} x_n$ of any sequence of standard objects $x_n$, defined as the equivalence class of all sequences $(y_n)_{n\in\mathbf{N}}$ such that $\{n \in \mathbf{N} : x_n = y_n\} \in p$. We then define the *ultrapower* ${}^*X$ of a standard set $X$ to be the collection of all ultralimits $\lim_{n\to p} x_n$ of sequences

$x_n$ in $X$. We can interpret ${}^*X$ as the space of all nonstandard elements of $X$, with the standard space $X$ being embedded in the nonstandard one ${}^*X$ by identifying $x$ with its nonstandard counterpart ${}^*x := \lim_{n\to p} x$. One can extend all (first-order) structures on $X$ to ${}^*X$ in the obvious manner, and a famous theorem of Łos [**Lo1955**] asserts that all first-order sentences that are true about a standard space $X$, will also be true about the nonstandard space ${}^*X$. Thus, for instance, the ultrapower ${}^*H$ of a standard Hilbert space $H$ over the standard complex numbers $\mathbf{C}$ will be a nonstandard Hilbert space ${}^*H$ over the nonstandard reals ${}^*\mathbf{R}$ or the nonstandard complex numbers ${}^*\mathbf{C}$. It has a nonstandard inner product $\langle , \rangle : {}^*H \times {}^*H \to {}^*\mathbf{C}$ instead of a standard one, which obeys the nonstandard analogue of the Hilbert space axioms. In particular, it is complete in the nonstandard sense: any nonstandard Cauchy sequence $(x_n)_{n\in{}^*\mathbf{N}}$ of nonstandard vectors $x_n \in {}^*H$ indexed by the nonstandard natural numbers ${}^*\mathbf{N}$ will converge (again, in the nonstandard sense) to a limit $x \in {}^*H$.

The ultrapower ${}^*H$ - the space of ultralimits $\lim_{n\to p} x_n$ of *arbitrary* sequences $x_n$ in $H$ - turns out to be too large and unwieldy to be helpful for us. We will work instead with a more tractable subquotient, defined as follows. Let $O(H)$ be the space of ultralimits $\lim_{n\to p} x_n$ of *bounded* sequences $x_n \in H$, and let $o(H)$ be the space of ultralimits[19] $\lim_{n\to p} x_n$ of sequences $x_n \in H$ that converge to zero. It is clear that $o(H)$, $O(H)$ are vector spaces over the *standard* complex numbers $\mathbf{C}$, with $o(H)$ being a subspace of $O(H)$. We define the quotient space $\tilde{H} := O(H)/o(H)$, which is then also a vector space over $\mathbf{C}$. One easily verifies that $H$ is a subspace of $O(H)$ that is disjoint from $o(H)$, so we can embed $H$ as a subspace of $\tilde{H}$.

**Remark 4.5.2.** When $H$ is finite dimensional, the Bolzano-Weierstrass theorem (or more precisely, the *proof* of this theorem) shows that $H = \tilde{H}$. For infinite-dimensional spaces, though, $\tilde{H}$ is larger than $H$, basically because there exist bounded sequences in $H$ with no convergent subsequences. Thus we can view the quotient $\tilde{H}/H$ as measuring the failure of the Bolzano-Weierstrass theorem (a sort of "Bolzano-Weierstrass cohomology", if you will).

Now we place a Hilbert space structure on $\tilde{H}$. Observe that if $x = \lim_{n\to p} x_n$ and $y = \lim_{n\to p} y_n$ are elements of $O(H)$ (so that $x_n, y_n$ are bounded), then the nonstandard inner product $\langle x, y\rangle = \lim_{n\to p}\langle x_n, y_n\rangle$ is a nonstandard complex number which is bounded (i.e. it it lies in $O(\mathbf{C})$). Since $\mathbf{C} = O(\mathbf{C})/o(\mathbf{C})$, we can thus extract a *standard part* $\mathrm{st}\langle x, y\rangle$, defined as the unique standard complex number such that

$$\langle x, y\rangle = \mathrm{st}\langle x, y\rangle + o(1)$$

[19] The space $o(H)$ is also known as the *monad* of the origin of $H$.

where $o(1)$ denotes an infinitesimal, i.e. a non-standard quantity whose magnitude is less than any standard positive real $\varepsilon > 0$. From the Cauchy-Schwarz inequality we see that if we modify either $x$ or $y$ by an element of $o(H)$, then the standard part $\mathrm{st}\langle x, y\rangle$ does not change. Thus, we see that the map $x, y \mapsto \mathrm{st}\langle x, y\rangle$ on $O(H)$ descends to a map $x, y \mapsto \langle x, y\rangle$ on $\tilde{H}$. One easily checks that this map is[20] a standard Hermitian inner product on $\tilde{H}$ that extends the one on the subspace $H$. Furthermore, by using the countable saturation (or Bolzano-Weierstrass) property of nonstandard analysis (see Section 4.4), we can also show that $\tilde{H}$ is complete with respect to this inner product; thus $\tilde{H}$ is a standard Hilbert space[21] that contains $H$ as a subspace.

After all this setup, we can now give the third proof of Theorem 4.5.1:

**Proof.** Let $z := \lim_{n\to p} x_n$ be the ultralimit of the $x_n$, then $z$ is an element of $O(H)$. Let $\tilde{z}$ be the image of $z$ in $\tilde{H}$, and let $x$ be the orthogonal projection of $\tilde{z}$ to $H$. We claim that a subsequence of $x_n$ converges weakly to $x$.

For any $y \in H$, $\tilde{z} - x$ is orthogonal to $y$, and thus $\langle z - x, y\rangle = o(1)$. In other words,

$$\lim_{n\to p} \langle x_n, y\rangle = \langle x, y\rangle + o(1) \tag{4.2}$$

for all $y \in H$. This is already the nonstandard analogue of weak convergence along a subsequence, but we can get to weak convergence itself with only a little more argument. Indeed, from (4.2) we can easily construct a subsequence $x_{n_j}$ such that

$$|\langle x_{n_j}, x_i\rangle - \langle x, x_i\rangle| \leq \frac{1}{j}$$

and

$$|\langle x_{n_j}, x\rangle - \langle x, x\rangle| \leq \frac{1}{j}$$

for all $1 \leq i \leq j$, which implies that

$$\lim_{j\to\infty} \langle x_{n_j}, y\rangle = \langle x, y\rangle$$

whenever $y$ is a finite linear combination of the $x_i$ and $x$. Applying a density argument using the boundedness of the $x_n$, this is then true for all $y$ in the closed span of the $x_i$ and $x$; it is also clearly true for $y$ in the orthogonal complement, and the claim follows. □

[20] If one prefers to think in terms of commutative diagrams, one can think of the inner product as a bilinear map from the short exact sequence $0 \to o(H) \to O(H) \to \tilde{H} \to 0$ to the short exact sequence $0 \to o(\mathbf{C}) \to O(\mathbf{C}) \to \mathbf{C} \to 0$.

[21] One can view $\tilde{H}$ as a sort of nonstandard completion of $H$, in a manner somewhat analogous to how the Stone-Cech compactification $\beta X$ of a space can be viewed as a topological completion of $X$. This is of course consistent with the philosophy of Section 4.4.

Observe that in contrast with the first two proofs, the third proof gave a "canonical" choice for the subsequence limit $x$. This is ultimately because the ultrafilter $p$ already "made all the choices beforehand", in some sense.

Observe also that we used the existence of orthogonal projections in Hilbert spaces in the above proof. If one unpacks the usual proof that these projections exist, one will find an energy increment argument that is not dissimilar to that used in the second proof of Theorem 4.5.1. Thus we see that the somewhat intricate energy increment argument from that second proof has in some sense been encapsulated into a general-purpose package in the nonstandard setting, namely the existence of orthogonal projections.

**4.5.2. Concentration compactness for unitary group actions.** Now we generalise the sequential Banach-Alaoglu theorem to allow for a group of symmetries. The setup is now that of a (standard) complex vector space $H$, together with a locally compact group $G$ acting unitarily on $H$ in a jointly continuous manner, thus the map $(g,x) \mapsto gx$ is jointly continuous from $G \times H$ to $H$ (or equivalently, the representation map from $G$ to $U(H)$ is continuous if we give $U(H)$ the strong operator topology). We also assume that $G$ is a group of *dislocations*, which means that $g_n x$ converges weakly to zero in $H$ whenever $x \in H$ and $g_n$ goes to infinity in $G$ (which means that $g_n$ eventually escapes any given compact subset of $G$). A typical example of such a group is the translation action $h : f(\cdot) \mapsto f(\cdot - h)$ of $\mathbf{R}^d$ on $L^2(\mathbf{R}^d)$, another example is the scaling action $\lambda : f(\cdot) \mapsto \frac{1}{\lambda^{d/2}} f(\frac{\cdot}{\lambda})$ of $\mathbf{R}^+$ on $L^2(\mathbf{R}^d)$. (One can also combine these two actions to give an action of the semidirect product $\mathbf{R}^+ \ltimes \mathbf{R}^d$ on $L^2(\mathbf{R}^d)$.)

The basic theorem here is

**Theorem 4.5.3** (Profile decomposition)**.** *Let $G, H$ be as above. Let $x_n$ be a bounded sequence in $H$ obeying the energy bound*

$$\limsup_{n\to\infty} \|x_n\|^2 \leq E.$$

*Then, after passing to a subsequence, one can find a sequence $\phi_1, \phi_2, \ldots \in H$ with the Bessel inequality*

$$\sum_{k=1}^{\infty} \|\phi_k\|^2 \leq E$$

*and group elements $g_{k,n} \in G$ for $k, n \in \mathbf{N}$ such that*

$$g_{k',n}^{-1} g_{k,n} \to \infty \text{ as } n \to \infty$$

*whenever $k \neq k'$ and $\phi_k, \phi_{k'}$ are non-zero, such that for each $K \in \mathbf{N}$ one has the decomposition*

$$x_n = \sum_{k=1}^{K} g_{k,n}\phi_k + w_{K,n}$$

*such that*

$$\limsup_{n\to\infty} \|w_{K,n}\|^2 \leq E - \sum_{k=1}^{K} \|\phi_k\|^2$$

*and*

$$\limsup_{n\to\infty} \sup_{g\in G} |\langle g^{-1} w_{K,n}, y\rangle|^2 \leq \sum_{k=K+1}^{\infty} \|\phi_k\|^2$$

*for all unit vectors $y$, and such that $g_{k,n}^{-1} w_{K,n}$ converges weakly to zero for every $1 \leq k \leq K$.*

Note that Theorem 4.5.1 is the case when $G$ is trivial.

There is a version of the conclusion available in which $K$ can be taken to be infinite, and also one can generalise $G$ to be a more general object than a group by modifying the hypotheses somewhat; see [**ScTi2002**]. The version with finite $K$ is slightly more convenient though for applications to nonlinear dispersive and wave equations; see [**KiVa2008**] for some applications of this type of decomposition. In order for this theorem to be useful for applications, one needs to exploit some sort of *inverse theorem* that controls other norms of a vector $w$ in terms of expressions such as $\sup_{g\in G} |\langle gw, y\rangle|$; these theorems tend to require "hard" harmonic analysis and cannot be established purely by such "soft" analysis tools as nonstandard analysis.

One can adapt the second proof of Theorem 4.5.1 to give a standard analysis proof of Theorem 4.5.3:

**Proof.** (Sketch) Applying Theorem 4.5.1 we can (after passing to a subsequence) find group elements $g_{1,n}$ such that $g_{1,n}^{-1} x_n$ converges weakly to a limit $\phi_1 \in H$, which we can choose to be nearly maximal in the sense that

$$\|\phi_1'\| \leq 2\|\phi_1\|$$

(say) whenever $\phi_1$ is the weak limit of $g_{n_j}^{-1} x_{n_j}$ for some subsequence $x_{n_j}$ and some collection of group elements $g_{n_j}$. In particular, this implies (from further application of Theorem 4.5.1, and an argument by contradiction) that

$$\limsup_{n\to\infty} \sup_{g\in G} |\langle g^{-1} x_n, y\rangle| \leq 2\|\phi_1\|$$

for any unit vector $y$.

We may now decompose

$$x_n = g_{1,n}\phi_1 + w_{1,n}$$

where $g_{1,n}^{-1}w_{1,n}$ converges weakly to zero. From Pythagoras' theorem we see that $w_{1,n}$ asymptotically has strictly less energy than $E$:

$$\limsup_{n\to\infty} \|w_{1,n}\|^2 \leq E - \|\phi_1\|^2.$$

We then repeat the argument, passing to a further subsequence and finding group elements $g_{2,n}$ such that $g_{2,n}^{-1}w_{1,n}$ converges weakly to $\phi_2 \in H$, with

$$\limsup_{n\to\infty} \sup_{g\in G} |\langle g^{-1}x_{1,n}, y\rangle| \leq 2\|\phi_2\|$$

for any unit vector $y$.

Note that $g_{1,n}^{-1}w_{1,n}$ converges weakly to zero, while $g_{2,n}^{-1}w_{1,n}$ converges weakly to $\phi_2$. If $\phi_2$ is non-zero, this implies that $g_{1,n}^{-1}g_{2,n}$ must go to infinity (otherwise it has a convergent subsequence, and this soon leads to a contradiction).

If one iterates the above construction and passes to a diagonal subsequence one obtains the claim. □

Now we give the nonstandard analysis proof. As before, we introduce the short exact sequence of Hilbert spaces:

$$0 \to o(H) \to O(H) \to \tilde{H} \to 0.$$

We will also need an analogous short exact sequence of groups

$$0 \to o(G) \to O(G) \to G \to 0$$

where $O(G) \leq {}^*G$ is the space of ultralimits $\lim_{n\to p} g_n$ of sequences $g_n$ in $G$ that lie in a compact subset of $G$, and $o(G) \leq O(G)$ is the space of ultralimits of $\lim_{n\to p} g_n$ of sequences $g_n$ that converge to the identity element (i.e. $o(G)$ is the monad of the group identity). One easily verifies that $o(G)$ is a normal subgroup of $O(G)$, and that the quotient is isomorphic[22] to $G$.

The group ${}^*G$ acts unitarily on ${}^*H$, and so preserves both $o(H)$ and $O(H)$. As such, it also acts unitarily on $\tilde{H}$. The induced action of the subgroup $o(G)$ is trivial; and the induced action of the subgroup $O(G)$ preserves $H$.

Let $\langle({}^*G)H\rangle$ be the closed span of the set $\{gx : g \in {}^*G; h \in H\}$ in $\tilde{H}$; this is a Hilbert space. Inside this space we have the subspaces $gH$ for $g \in {}^*G$. As $O(G)$ preserves $H$, we see that $gH = g'H$ whenever $g, g'$ lie in the same coset of ${}^*G/O(G)$, so we can define $\gamma H$ for any $\gamma \in {}^*G/O(G)$ in a well-defined manner. On the other hand, if $g, g'$ do not lie in the same coset

[22] Indeed, $O(G)$ can be expressed as a semi-direct product $G \ltimes o(G)$, though we will not need this fact here.

of ${}^*H$, then we have $g' = g \lim_{n\to p} h_n$ for some sequence $h_n$ in $G$ that goes to infinity. As $G$ is a group of dislocations, we conclude that $g'H$ and $gH$ are now orthogonal. In other words, $\gamma' H$ and $\gamma H$ are orthogonal whenever $\gamma, \gamma' \in {}^*G/O(G)$ are distinct. We conclude that we have the decomposition

$$\langle ({}^*G)H \rangle = \bigoplus_{\gamma \in {}^*G/O(G)} \gamma H \tag{4.3}$$

where $\bigoplus$ is the Hilbert space direct sum.

Now we can prove Theorem 4.5.3. As in the previous section, starting with a bounded sequence $x_n$ in $H$, we form the ultralimit $z := \lim_{n\to p} x_n \in O(H)$ and the image $\tilde{z} \in \tilde{H}$. We let $x$ be the orthogonal projection of $\tilde{z}$ to $\langle ({}^*G)H \rangle$. By (4.3), we can write

$$x = \sum_k g_k \phi_k$$

for some at most countable sequence of vectors $\phi_k \in H$ and $g_k \in {}^*G$, with the $g_n$ lying in distinct cosets of $O(G)$. In particular, for any $k \neq k'$, $g_{k'}^{-1} g_k$ is the ultralimit of a sequence of vectors going to infinity. By adding dummy values of $g_k, \phi_k$ if necessary we may assume that $k$ ranges from 1 to infinity. Also, one has the Bessel inequality

$$\sum_k \|\phi_k\|^2 = \|x\|^2 \leq \|z\|^2 \leq E$$

and from Cauchy-Schwarz and Bessel one has

$$|\langle z - \sum_{k=1}^K g_k \phi_k, gy \rangle| \leq \sum_{k=K+1}^\infty \|\phi_k\|^2.$$

for any unit vector $y \in H$ and $g \in G$. From this we can obtain the required conclusions by arguing as in the previous section.

**4.5.3. Concentration compactness for measures.** We now give a variant of the profile decomposition, for Borel probability measures $\mu_n$ on $\mathbf{R}^d$. Recall that such a sequence is said to be *tight* if, for every $\varepsilon > 0$, there is a ball $B(0,R)$ such that $\limsup_{n\to\infty} \mu_n(\mathbf{R}^d \backslash B(0,R)) \leq \varepsilon$. Given any Borel probability measure $\mu$ on $\mathbf{R}^d$ and any $x \in \mathbf{R}^d$, define the translate $\tau_x \mu$ to be the Borel probability measure given by the formula $\tau_x \mu(E) := \mu(E - x)$.

**Theorem 4.5.4** (Profile decomposition for probability measures on $\mathbf{R}^d$)**.** *Let $\mu_n$ be a sequence of Borel probability measures on $\mathbf{R}^d$. Then, after passing to a subsequence, one can find a sequence $c_k$ of non-negative real numbers with $\sum_k c_k \leq 1$, a tight sequence $\nu_{k,n}$ of positive measures whose mass converges to 1 as $n \to \infty$ for fixed $k$, and shifts $x_{k,n} \in \mathbf{R}^d$ such that*

$$x_{k,n} - x_{k',n} \to \infty \text{ as } n \to \infty$$

*for all $k \neq k'$, and such that for each $K$, one has the decomposition*

$$\mu_n = \sum_{k=1}^{K} c_k \tau_{k,n} \nu_{k,n} + \rho_{K,n}$$

*where the error $\rho_{K,n}$ obeys the bounds*

$$\limsup_{n\to\infty} \sup_{x\in\mathbf{R}^d} \rho_{K,n}(B(x,R)) \leq \sup_{k\geq K} c_k$$

*and*

$$\lim_{n\to\infty} \rho_{K,n}(B(x_{k,n},R)) = 0$$

*for all radii $R$ and $1 \leq k \leq K$.*

*Furthermore, one can ensure that for each $k$, $\nu_{k,n}$ converges in the* vague topology *to a probability measure $\nu_k$.*

We first give the standard proof of this theorem:

**Proof.** (Sketch) Suppose first that

$$\limsup_{n\to\infty} \sup_{x\in\mathbf{R}^d} \mu_n(B(x,R)) = 0$$

for all $R$. Then we are done by setting all the $c_k$ equal to zero, and $\rho_{K,n} = \mu_n$. So we may assume that we can find $R$ such that

$$\limsup_{n\to\infty} \sup_{x\in\mathbf{R}^d} \mu_n(B(x,R)) = \alpha$$

for some $\alpha > 0$; we may also assume that $\alpha$ is approximately maximal in the sense that

$$\limsup_{n\to\infty} \sup_{x\in\mathbf{R}^d} \mu_n(B(x,R')) \leq 2\alpha$$

(say) for all other radii $R'$. By passing to a subsequence, we may thus find $x_{1,n} \in \mathbf{R}^d$ such that

$$\lim_{n\to\infty} \mu_n(B(x_{1,n},R)) = \alpha;$$

By passing to a further subsequence using the Helly selection principle (or the sequential Banach-Alaoglu theorem), we may assume that the translates $\tau_{-x_{1,n}}\mu_n$ converge in the vague topology to a limit of total mass at most 1 and at least $\alpha$, and which can be expressed as $c_1\nu_1$ for some $c_1 \geq \alpha$ and a probability measure $\nu_1$.

As $\tau_{-x_{1,n}}\mu_n$ converges vaguely to $c_1\nu_1$, we have

$$\limsup_{n\to\infty} \tau_{-x_{1,n}}\mu_n(B(0,R')\backslash B(0,R)) \leq c_1\nu_1(\mathbf{R}^d\backslash B(0,R/2))$$

for any $0 < R < R'$. By making $R'_n$ grow sufficiently slowly to infinity with respect to $n$, we may thus ensure that

$$\limsup_{n\to\infty} \tau_{-x_{1,n}}\mu_n(B(0,R'_n)\backslash B(0,R)) \leq c_1\nu_1(\mathbf{R}^d\backslash B(0,R/2))$$

for all integers $R > 0$. If we then set $c_1\tilde{\nu}_{1,n}$ to be the restriction of $\tau_{-x_{1,n}}\mu_n$ to $B(0, R'_n)$, we see that $\tilde{\nu}_{1,n}$ is tight, converges vaguely to $\nu_{1,n}$, and has total mass converging to 1. We can thus split

$$\mu_n = c_1\tau_{x_{1,n}}\tilde{\nu}_{1,n} + \rho_{1,n}$$

for some residual positive measure $\rho_{1,n}$ of total mass converging to $1 - c_1$, and such that $\rho_{1,n}(B(x_{1,n}, R)) \to 0$ as $n \to \infty$ for any fixed $R$. We can then iterate this procedure to obtain the claims of the theorem (after one last diagonalisation to combine together all the subsequences). □

Now we give the nonstandard proof. We take the ultralimit $\mu := \lim_{n\to p} \mu_n$ of the standard Borel probability measures $\mu_n$ on $\mathbf{R}^d$, resulting in a nonstandard Borel probability measure. What, exactly, is a nonstandard Borel probability measure? A standard Borel probability measure, such as $\mu_n$, is a map $\mu_n : \mathcal{B} \to [0,1]$ from the standard Borel $\sigma$-algebra $\mathcal{B}$ to the unit interval $[0,1]$ which is countably additive and maps $\mathbf{R}^n$ to 1. Thus, the nonstandard Borel probability measure is a nonstandard map $\mu : {}^*\mathcal{B} \to {}^*[0,1]$ from the nonstandard Borel $\sigma$-algebra (the collection of all ultralimits of standard Borel sets) to the nonstandard interval $[0,1]$ which is nonstandardly countably additive and maps ${}^*\mathbf{R}^n$ to 1. In particular, it is finitely additive.

There is an important subtlety here. The nonstandard Borel $\sigma$-algebra is closed under nonstandard countable unions: if $(E_n)_{n\in{}^*\mathbf{N}}$ is a nonstandard countable sequence of nonstandard Borel sets (i.e. an ultralimit of standard countable sequences $(E_{n,m})_{n\in\mathbf{N}}$ of standard Borel sets), then $\bigcup_n E_n$ is also nonstandard Borel, but this is not necessarily the case for external countable unions, thus if $(E_n)_{n\in\mathbf{N}}$ is an external countable sequence of nonstandard Borel sets, then $\bigcup_n E_n$ need not be nonstandard Borel. On the other hand, $\mathcal{B}$ is certainly still closed under finite unions and other finite Boolean operations, so it can be viewed (externally) as a Boolean algebra, at least.

Now we perform the Loeb measure construction (which was also introduced in Section 4.4). Consider the standard part $\mathrm{st}(\mu)$ of $\mu$; this is a finitely additive map from ${}^*\mathcal{B}$ to $[0,1]$. From the countable saturation property, one can verify that this map is a premeasure, and so (by the *Hahn-Kolmogorov theorem*, see [**Ta2011**, §1.7]) extends to a countably additive probability measure $\tilde{\mu}$ on the measure-theoretic completion $\tilde{\mathcal{B}} := \overline{\langle {}^*\mathcal{B}\rangle}$ of ${}^*\mathcal{B}$.

The measure $\tilde{\mu}$ is a measure on ${}^*\mathbf{R}^d$. We push it forward to the quotient space ${}^*\mathbf{R}^d/O(\mathbf{R}^d)$ by the obvious quotient map $\pi : {}^*\mathbf{R}^d/\mathbf{R}^d$ to obtain a pushforward measure $\pi_*\tilde{\mu}$ on the pushforward $\sigma$-algebra $\pi_*\tilde{\mathcal{B}}$, which consists of all (external) subsets $E$ of ${}^*\mathbf{R}^d/O(\mathbf{R}^d)$ whose preimage $\pi^{-1}(E)$ is measurable in $\tilde{\mathcal{B}}$.

We claim that every point in ${}^*\mathbf{R}^d/O(\mathbf{R}^d)$ is measurable in $\pi_*\tilde{\mathcal{B}}$, or equivalently that every coset $x + O(\mathbf{R}^d)$ in ${}^*\mathbf{R}^d$ is measurable in $\mathcal{B}$. Indeed, this coset is the union of the countable family of (nonstandard) balls $\{y \in {}^*\mathbf{R}^d : |x-y| < n\}$ for $n \in \mathbf{N}$, each one of which is a nonstandard Borel set and thus measurable in $\tilde{\mathcal{B}}$.

Because of this, we can decompose the measure $\pi_*\tilde{\mu}$ into pure point and singular components, thus

$$\pi_*\tilde{\mu} = \sum_k c_k \delta_{x_k + O(\mathbf{R}^d)} + \rho$$

where $c_k$ are standard non-positive reals, $k$ ranges over an at most countable set, $x_k + O(\mathbf{R}^d)$ are disjoint cosets in ${}^*\mathbf{R}^d/O(\mathbf{R}^d)$, and $\rho$ is a finite measure on $\pi_*\tilde{\mathcal{B}}$ such that

$$\sum_k c_k + \|\rho\| = 1$$

and

$$\rho(\{x + O(\mathbf{R}^d)\}) = 0$$

for every coset $x + O(\mathbf{R}^d)$.

Now we analyse the restriction of $\tilde{\mu}$ to a single coset $x_k + O(\mathbf{R}^d)$, which has total mass $c_k$. For any standard continuous, compactly supported function $f : \mathbf{R}^d \to \mathbf{R}$, one can form the integral

$$\int_{x_k + O(\mathbf{R}^d)} {}^*f(x - x_k)\ d\tilde{\mu}(x).$$

This is a non-negative continuous linear functional, so by the Riesz representation theorem there exists a non-negative Radon measure $\nu_k$ on $\mathbf{R}^d$ such that

$$\int_{x_k + O(\mathbf{R}^d)} {}^*f(x - x_k)\ d\tilde{\mu}(x) = c_k \int_{\mathbf{R}^d} f(y)\ d\nu_k(y)$$

for all such $f$. As $x_k + O(\mathbf{R}^d)$ has total mass $c_k$, $\nu_k$ is a probability measure. From definition of $\tilde{\mu}$, we thus have

$$\int_{{}^*\mathbf{R}^d} {}^*f(x - x_k)\ d\mu(x) = c_k \int_{\mathbf{R}^d} f(y)\ d\nu_k(y) + o(1)$$

for all $f$.

We have

$$\mu(B(x_k, R)) \leq c_k + o(1)$$

for every standard $R$, and thus by the overspill principle there exists an unbounded $R_k$ for which

$$\mu(B(x_k, R_k)) \leq c_k + o(1);$$

since $\mu(x_k + O(\mathbf{R}^d)) = c_k$, we thus have

$$\mu(B(x_k, R_k)) = c_k + o(1);$$

If we set $c_k\tilde{\nu}_k$ to be the restriction of $\tau_{-x_k}\mu$ to $B(0, R_k)$, we thus see that

$$\int_{{}^*\mathbf{R}^d} {}^*f(x)\ d\tilde{\nu}_k(y) = \int_{\mathbf{R}^d} f(y)\ d\nu_k(y) + o(1)$$

for all test functions $f$. Writing $\tilde{\nu}_k$ as the ultralimit of probability measures $\tilde{\nu}_{k,n}$, we thus see (upon passing to a subsequence) that $\tilde{\nu}_{k,k}$ converges vaguely to the probability measure $\nu_k$, and is in particular tight.

For any standard $K \geq 1$, we can write

$$\mu = \sum_{k=1}^{K} c_k \tau_{x_k} \nu_k + \rho_K$$

where $\rho_K$ is a finite measure. Letting $\tilde{\rho}_K$ be the Loeb extension of the standard part of $\rho_K$, we see that $\tilde{\rho}_K$ assigns zero mass to $x_k + O(\mathbf{R}^d)$ for $k \leq K$ and assigns a mass of at most $\sup_{k>K} c_k$ to any other coset of $O(\mathbf{R}^d)$. This implies that

$$\tilde{\rho}_K(B(x, R)) \leq \sup_{k>K} c_k + o(1)$$

for any standard $R$. Expressing $\rho_K$ as an ultralimit of $\rho_{K,n}$, we then obtain the claim.

*Chapter 5*

# Partial differential equations

## 5.1. Quasilinear well-posedness

When solving the initial value problem to an *ordinary differential equation*, such as

$$\partial_t u = F(u); \quad u(0) = u_0, \tag{5.1}$$

where $u : \mathbf{R} \to V$ is the unknown solution (taking values in some finite-dimensional vector space $V$), $u_0 \in V$ is the initial datum, and $F : V \to V$ is some nonlinear function (which we will take to be smooth for sake of argument), then one can construct a solution locally in time via the *Picard iteration method*. There are two basic ideas. The first is to use the *fundamental theorem of calculus* to rewrite the initial value problem (5.1) as the problem of solving an *integral equation*,

$$u(t) = u_0 + \int_0^t F(u(s))\ ds. \tag{5.2}$$

The second idea is to solve this integral equation by the *contraction mapping theorem*, showing that the integral operator $\mathcal{N}$ defined by

$$\mathcal{N}(u)(t) := u_0 + \int_0^t F(u(s))\ ds$$

is a contraction on a suitable complete metric space (e.g. a closed ball in the function space $C^0([0,T];V)$), and thus has a unique fixed point in this space. This method works as long as one only seeks to construct *local* solutions (for

time $t$ in $[0,T]$ for sufficiently small $T > 0$), but the solutions constructed have a number of very good properties, including

(1) **Local existence**: A solution $u$ exists in the space $C^0([0,T];V)$ (and even in $C^\infty([0,T];V)$) for $T$ sufficiently small.

(2) **Uniqueness**: There is at most one solution $u$ to the initial value problem in the space $C^0([0,T];V)$ (or in smoother spaces, such as $C^\infty([0,T];V)$). (For solutions in the weaker space $C^0([0,T];V)$ we use the integral formulation (5.2) to define the solution concept.)

(3) **Lipschitz continuous dependence on the data**: If $u_0^{(n)}$ is a sequence of initial data converging to $u_0$, then the associated solutions $u^{(n)}$ converge uniformly to $u$ on $[0,T]$ (possibly after shrinking $T$ slightly). In fact we have the Lipschitz bound $\|u^{(n)}(t) - u(t)\|_V \leq C\|u_0^{(n)} - u_0\|_V$ for $n$ large enough and $t \in [0,T]$, where $C$ is an absolute constant.

This package of properties is referred to as *(local Lipschitz) wellposedness*.

This method extends to certain *partial differential equations*, particularly those of a *semilinear* nature (linear except for lower order nonlinear terms). For instance, if trying to solve an initial value problem of the form

$$\partial_t u + Lu = F(u); \quad u(0,x) = u_0(x),$$

where now $u : \mathbf{R} \to V$ takes values in a function space $V$ (e.g. a Sobolev space $H^k(\mathbf{R}^d)$), $u_0 \in V$ is an initial datum, $L$ is some (differential) operator (*independent* of $u$) that is (densely) defined on $V$, and $F$ is a nonlinearity which is also (densely) defined on $V$, then (formally, at least) one can solve this problem by using *Duhamel's formula* to convert the problem to that of solving an integral equation

$$u(t) = e^{-tL}u_0 + \int_0^t e^{-(t-s)L}F(u(s))\ ds$$

and one can then hope to show that the associated nonlinear integral operator

$$u \mapsto e^{-tL}u_0 + \int_0^t e^{-(t-s)L}F(u(s))\ ds$$

is a contraction in a subset of a suitably chosen function space.

This method turns out to work surprisingly well for many semilinear partial differential equations, and in particular for semilinear parabolic, semilinear dispersive, and semilinear wave equations. As in the ODE case, when the method works, it usually gives the entire package of Lipschitz wellposedness: existence, uniqueness, and Lipschitz continuous dependence on the initial data, for short times at least.

However, when one moves from semilinear initial value problems to *quasilinear* initial value problems such as

$$\partial_t u + L_u u = F(u); \quad u(0,x) = u_0(x)$$

in which the top order operator $L_u$ now depends on the solution $u$ itself, then the nature of well-posedness changes; one can still hope to obtain (local) existence and uniqueness, and even continuous dependence on the data, but one usually is forced to give up Lipschitz continuous dependence at the highest available regularity (though one can often recover it at lower regularities). As a consequence, the Picard iteration method is not directly suitable for constructing solutions to such equations.

One can already see this phenomenon with a very simple equation, namely the one-dimensional constant-velocity *transport equation*

$$\partial_t u + c\partial_x u = 0; \quad u(0,x) = u_0(x) \tag{5.3}$$

where we consider $c = c_0$ as part of the initial data. (If one wished, one could view this equation as a rather trivial example of a system, namely

$$\begin{aligned} \partial_t u + c\partial_x u &= 0 \\ \partial_t c &= 0 \\ u(0,x) &= u_0(x); \\ c(0) &= c_0, \end{aligned}$$

to emphasise this viewpoint, but this would be somewhat idiosyncratic.) One can solve this equation explicitly of course to get the solution

$$u(t,x) = u_0(x - ct).$$

In particular, if we look at the solution just at time $t = 1$ for simplicity, we have

$$u(1,x) = u_0(x - c).$$

Now let us see how this solution $u(1,x)$ depends on the parameter $c$. One can ask whether this dependence is Lipschitz in $c$, in some function space $V$:

$$\|u_0(\cdot - c) - u_0(\cdot - c')\|_V \leq A|c - c'|$$

for some finite $A$. But using the Newton approximation

$$u_0(\cdot - c) - u_0(\cdot - c') \approx (c - c')\partial_x u_0(\cdot - c)$$

we see that we should only expect such a bound when $\partial_x u_0$ (and its translates) lie in $V$. Thus, we see a *loss of derivatives* phenomenon with regard to Lipschitz well-posedness; if the initial data $u_0$ is in some regularity space, say $C^3$, then one only obtains Lipschitz dependence on $c$ in a lower regularity space such as $C^2$.

We have just seen that if all one knows about the initial data $u_0$ is that it is bounded in a function space $V$, then one usually cannot hope to make the dependence of $u$ on the velocity parameter $c$ Lipschitz continuous. Indeed, one cannot even make it *uniformly* continuous[1] in $V$. Given two values of $c$ that are close together, e.g. $c = 0$ and $c = \varepsilon$, and a reasonable function space $V$ (e.g. a Sobolev space $H^k$, or a classical regularity space $C^k$) one can easily cook up a function $u_0$ that is bounded in $V$ but whose two solutions $u_0(\cdot)$ and $u_0(\cdot - \varepsilon)$ separate in the $V$ norm at time 1, simply by choosing $u_0$ to be supported on an interval of width $\varepsilon$.

On the other hand, one still has *non-uniform* continuous dependence on the initial parameters: if $u_0$ lies in some reasonable function space $V$, then the map $c \mapsto u_0(\cdot - c)$ is continuous[2] in the $V$ topology, even if it is not uniformly continuous with respect to $v_0$. The reason for this is that we already have established this continuity in the case when $u_0$ is so smooth that an additional derivative of $u_0$ lies in $V$; and such smooth functions tend to be dense in the original space $V$, so the general case can then be established by a limiting argument, approximating a general function in $V$ by a smoother function. We then see that the non-uniformity ultimately comes from the fact that a given function in $V$ may be arbitrarily rough (or concentrated at an arbitrarily fine scale), and so the ability to approximate such a function by a smooth one can be arbitrarily poor.

In many quasilinear PDE, one often encounters qualitatively similar phenomena. Namely, one often has local well-posedness in sufficiently smooth function spaces $V$ (so that if the initial data lies in $V$, then for short times one has existence, uniqueness, and continuous dependence on the data in the $V$ topology), but Lipschitz or uniform continuity in the $V$ topology is usually false. However, if the data (and solution) is known to be in a high-regularity function space $V$, one can often recover Lipschitz or uniform continuity in a lower-regularity topology.

Because the continuous dependence on the data in quasilinear equations is necessarily non-uniform, the arguments needed to establish this dependence can be remarkably delicate. As with the simple example of the transport equation, the key is to approximate a rough solution by a smooth solution first, by smoothing out the data (this is the non-uniform step, as it

[1] Part of the problem here is that using a subtractive method $\|u - v\|_V$ to determine the distance between two solutions $u, v$ is not a physically natural operation when transport mechanisms are present that could cause the key features of $u, v$ (such as singularities) to be situated in slightly different locations. In such cases, the correct notion of distance may need to take transport into account, e.g. by using metrics such as the *Wasserstein metric*.

[2] More succinctly: translation is a continuous but not uniformly continuous operation in most function spaces.

depends on the physical scale (or wavelength) that the data features are located). But for quasilinear equations, keeping the rough and smooth solution together can require a little juggling of function space norms, in particular playing the low-frequency nature of the smooth solution against the high-frequency nature of the residual between the rough and smooth solutions.

In this section I will illustrate this phenomenon with one of the simplest quasilinear equations, namely the initial value problem for the *inviscid Burgers' equation*

$$\partial_t u + u u_x = 0; \quad u(0,x) = u_0(x) \tag{5.4}$$

which is a modification of the transport equation (5.3) in which the velocity $c$ is no longer a parameter, but now depends (and is, in this case, actually *equal* to) the solution. To avoid technicalities we will work only with the classical function spaces $C^k$ of $k$ times continuously differentiable functions, though one can certainly work with other spaces (such as Sobolev spaces) by exploiting the *Sobolev embedding theorem*. To avoid having to distinguish continuity from uniform continuity, we shall work in a compact domain by assuming periodicity in space, thus for instance restricting $x$ to the unit circle $\mathbf{R}/\mathbf{Z}$.

This discussion is inspired by the survey article [**Tz2006**] of Tzvetkov, which further explores the distinction between well-posedness and ill-posedness in both semilinear and quasilinear settings.

**5.1.1. A priori estimates.** To avoid technicalities let us make the *a priori* assumption that all solutions of interest are smooth.

The Burgers equation is a pure transport equation: it moves the solution $u$ around, but does not increase or decrease its values. As a consequence we obtain an *a priori* estimate for the $C^0$ norm:

$$\|u(t)\|_{C^0} \leq \|u(0)\|_{C^0}.$$

To deal with the $C^1$ norm, we perform the standard trick of *differentiating the equation*, obtaining

$$\partial_t u_x + u u_{xx} + u_x^2 = 0$$

which we rewrite as a forced transport equation

$$(\partial_t + u\partial_x) u_x = -u_x^2.$$

Inspecting what this equation does at local maxima in space, one is led (formally, at least) to the differential inequality

$$\partial_t \|u_x\|_{C^0} \leq \|u_x\|_{C^0}^2$$

which leads to an *a priori* estimate of the form

$$\|u(t)\|_{C^1} \leq C\|u(0)\|_{C^1} \tag{5.5}$$

for some absolute constant $C$, if $t$ is sufficiently small depending on $\|u(0)\|_{C^1}$. More generally, the same arguments give

$$\|u(t)\|_{C^k} \leq C_k \|u(0)\|_{C^k}$$

for $k = 1, 2, 3, \ldots$, where $C_k$ depends only on $k$, and $t$ is sufficiently small depending on[3] $\|u(0)\|_{C^k}$.

**Remark 5.1.1.** One can also obtain similar *a priori* estimates for the Sobolev scale of spaces $H^k$ by using differentiation under the integral sign, follows by integration by parts; this is an example of the *energy method*, which we will not elaborate further upon here.

The *a priori* estimates are not quite enough by themselves to establish local existence of solutions in the indicated function spaces, but in practice, once one has *a priori* estimates, one can usually work a little bit harder to then establish existence, for instance by using a compactness, viscosity, or penalty method. We will not discuss this topic here.

**5.1.2. Lipschitz continuity at low regularity.** Now let us consider two solutions $u, v$ to Burgers' equation from two different initial data, thus

$$\partial_t u + u u_x = 0; \quad u(0) = u_0 \tag{5.6}$$

and

$$\partial_t v + v v_x = 0; \quad v(0) = v_0. \tag{5.7}$$

We want to say that if $u_0$ and $v_0$ are close in some sense, then $u$ and $v$ will stay close at later times. For this, the standard trick is to look at the difference $w := v - u$ of the two solutions. Subtracting (5.6) from (5.7) we obtain the *difference* equation for $w$:

$$\partial_t w + v w_x + w u_x = 0; \quad w(0) = w_0 := v_0 - u_0. \tag{5.8}$$

We can view the evolution equation in (5.8) as a forced transport equation:

$$(\partial_t + v\partial_x) w = -w u_x.$$

This leads to a bound for how the $C^0$ norm of $w$ grows:

$$\partial_t \|w\|_{C^0} \leq \|w u_x\|_{C^0} \leq \|w\|_{C^0} \|u\|_{C^1}.$$

Applying *Gronwall's inequality*, one obtains the *a priori* inequality

$$\|w(t)\|_{C^0} \leq \|w_0\|_{C^0} \exp\Big(\int_0^t \|u(s)\|_{C^1}\, ds\Big)$$

and hence by (5.5) we have

$$\|u(t) - v(t)\|_{C^0} \leq \|u_0 - v_0\|_{C^0} \exp(C\|u_0\|_{C^1}) \tag{5.9}$$

[3] Actually, if one works a little more carefully, one only needs $t$ sufficiently small depending on $\|u(0)\|_{C^1}$.

if $t$ is sufficiently small (depending on the $C^1$ norm of $u_0$). Thus we see that we have Lipschitz dependence in the $C^0$ topology... but only if at least one of the two solutions $u, v$ already had one higher derivative of regularity (i.e. one of $u, v$ was in $C^1$ and not just in $C^0$).

More generally, by using the trick of differentiating the equation, one can obtain an *a priori* inequality of the form

$$\|u(t) - v(t)\|_{C^k} \leq \|u_0 - v_0\|_{C^k} \exp(C_k \|u_0\|_{C^{k+1}})$$

for some $C_k$ depending only on $k$, for $t$ sufficiently small depending on $\|u_0\|_{C^{k+1}}$. Once again, to get Lipschitz continuity at some regularity $C^k$, one must first assume one higher degree $C^{k+1}$ of regularity on one of the solutions.

This loss of derivatives is unfortunate, but this is at least good enough to recover uniqueness: setting $u_0 = v_0$ in, say, (5.9) we obtain uniqueness of $C^1$ solutions (locally in time, at least), thanks to the trivial fact that two $C^1$ functions that agree in $C^0$ norm automatically agree in $C^1$ norm also. One can then boost local uniqueness to global uniqueness by a standard continuity argument which we omit.

**5.1.3. Non-uniform continuity at high regularity.** Let $u_0^{(n)}$ be a sequence of $C^1$ data converging in the $C^1$ topology to a limit $u_0 \in C^1$. As $u_0$ and $u_0^{(n)}$ are then uniformly bounded in $C^1$, existence theory then gives us $C^1$ solutions $u^{(n)}$, $u$ to the associated initial value problems

$$\partial_t u + u u_x = 0; \quad u(0) = u_0 \tag{5.10}$$

and

$$\partial_t u^{(n)} + u^{(n)} u_x^{(n)} = 0; \quad u^{(n)}(0) = u_0^{(n)} \tag{5.11}$$

for all $t$ in some uniform time interval $[0, T]$.

From (5.5) we know that the $u^{(n)}$ and $u$ are uniformly bounded in $C^1$ norm (for $T$ small enough). From the Lipschitz continuity (5.9) we know that $u^{(n)}$ converges to $u$ in $C^0$ norm. But does $u^{(n)}$ converge to $u$ in the $C^1$ norm?

The answer is yes, but the proof is remarkably delicate. A direct attempt to control the difference between $u^{(n)}$ and $u$ in $C^1$, following the lines of the previous argument, requires something to be bounded in $C^2$. But we only have $u^{(n)}$ and $u$ bounded in $C^1$.

However, note that in the arguments of the previous section, we don't need *both* solutions to be in $C^2$; it's enough for just *one* solution to be in $C^2$. Now, while neither $u^{(n)}$ nor $u$ are bounded in $C^2$ yet, what we can do is to introduce a *third* solution $v$, which is *regularised* to lie in $C^2$ and not just in $C^1$, while still being initially close to $u_0$ and hence to $u_0^{(n)}$ in $C^1$ norm.

The hope is then to show that $u$ and $u^{(n)}$ are both close to $v$ in $C^1$, which by the triangle inequality will make $u$ and $u^{(n)}$ close to each other.

Unfortunately, in order to get the regularised solution $v$ close to $u_0$ initially, the $C^2$ norm of $v(0)$ (and hence of $v$) may have to be quite large. But we can compensate for this by making the $C^0$ distance between $v(0)$ and $u_0$ quite small. The two effects turn out to basically cancel each other and allow one to proceed.

Let's see how this is done. We will use an argument of Bona and Smith [**BoSm1975**]. Consider a solution $v$ which is initially close to $u_0$ in $C^1$ norm (and *very* close in $C^0$ norm), and also has finite (but potentially large) $C^2$ norm; we will quantify these statements more precisely later.

Once again, we set $w = v - u$ and $w_0 = v_0 - u_0$, giving a difference equation which we now write as

$$\partial_t w + uw_x + wv_x = 0; \quad w(0) = w_0 \tag{5.12}$$

in order to take advantage of the higher regularity of $v$. For the $C^0$ norm, we have

$$\|w(t)\|_{C^0} = O(\|w_0\|_{C^0}) \tag{5.13}$$

for $t$ sufficiently small, thanks (5.9) and the uniform $C^1$ bounds. For the $C^1$ norm, we first differentiate (5.12) to obtain

$$(\partial_t + u\partial_x)w_x = -u_x w_x - w_x v_x - w v_{xx}$$

and thus

$$\partial_t \|w_x\|_{C^0} \leq \|u_x w_x\|_{C^0} + \|w_x v_x\|_{C^0} + \|w v_{xx}\|_{C^0}.$$

The first two terms on the RHS are $O(\|w_x\|_{C^0})$ thanks to the uniform $C^1$ bounds. The third term is $O(\|w_0\|_{C^0}\|v_0\|_{C^2})$ by (5.13) and *a priori* $C^2$ estimates (here we use the fact that the time of existence for $C^2$ bounds can be controlled by the $C^1$ norm). Using Gronwall's inequality, we conclude that

$$\|w_x(t)\|_{C^0} \ll \|\partial_x w_0\|_{C^0} + \|w_0\|_{C^0}\|v_0\|_{C^2}$$

and thus

$$\|v(t) - u(t)\|_{C^1} \ll \|v_0 - u_0\|_{C^1} + \|v_0 - u_0\|_{C^0}\|v_0\|_{C^2}.$$

Similarly one has

$$\|v(t) - u^{(n)}(t)\|_{C^1} \ll \|v_0 - u_0^{(n)}\|_{C^1} + \|v_0 - u_0^{(n)}\|_{C^0}\|v_0\|_{C^2},$$

and so by the triangle inequality we have

$$\|u^{(n)}(t) - u(t)\|_{C^1} \ll \|v_0 - u_0\|_{C^1} + \|v_0 - u_0\|_{C^0}\|v_0\|_{C^2} \tag{5.14}$$

for $n$ sufficiently large.

Note how the $C^2$ norm in the second term is balanced by the $C^0$ norm. We can exploit this balance as follows. Let $\varepsilon > 0$ be a small quantity, and let $v_0 := u_0 * P_\varepsilon$, where $P_\varepsilon = \frac{1}{\varepsilon} P(\frac{x}{\varepsilon})$ is a suitable approximation to the identity. A little bit of integration by parts using the $C^1$ bound on $u_0$ then gives the bounds

$$\|v_0 - u_0\|_{C^0} \ll \varepsilon$$

and

$$\|v_0 - u_0\|_{C^1} \ll 1$$

and

$$\|v_0\|_{C^2} \ll \frac{1}{\varepsilon}.$$

This is not quite enough to get anything useful out of (5.14). But to do better, we can use the fact that $\partial_x u_0$, being uniformly continuous, has some *modulus of continuity*, thus one has

$$\|\partial_x u_0(\cdot + t) - \partial_x u_0(\cdot)\|_{C^0} = o(1)$$

as $t \to 0$. Using this, one can soon get the improved estimates

$$\|v_0 - u_0\|_{C^0} = o(\varepsilon)$$

and

$$\|v_0 - u_0\|_{C^1} = o(1)$$

as $\varepsilon \to 0$. Applying (5.14), we thus see that

$$\|u^{(n)}(t) - u(t)\|_{C^1} \ll o(1)$$

for $n$ sufficiently large, and the continuity claim follows.

## 5.2. A type diagram for function spaces

In harmonic analysis and PDE, one often wants to place a function $f : \mathbf{R}^d \to \mathbf{C}$ on some domain (let's take a Euclidean space $\mathbf{R}^d$ for simplicity) in one or more function spaces in order to quantify its "size" in some sense. Examples include

(1) The *Lebesgue spaces* $L^p$ of functions $f$ whose norm $\|f\|_{L^p} := (\int_{\mathbf{R}^d} |f|^p)^{1/p}$ is finite, as well as their relatives such as the weak $L^p$ spaces $L^{p,\infty}$ (and more generally the *Lorentz spaces* $L^{p,q}$) and *Orlicz spaces* such as $L \log L$ and $e^L$;

(2) The classical regularity spaces $C^k$, together with their *Hölder continuous* counterparts $C^{k,\alpha}$;

(3) The *Sobolev spaces* $W^{s,p}$ of functions $f$ whose norm $\|f\|_{W^{s,p}} = \|f\|_{L^p} + \||\nabla|^s f\|_{L^p}$ is finite (other equivalent definitions of this norm exist, and there are technicalities if $s$ is negative or $p \notin (1, \infty)$), as

well as relatives such as homogeneous[4] *Sobolev spaces* $\dot{W}^{s,p}$, *Besov spaces* $B^{s,p}_q$, and *Triebel-Lizorkin spaces* $F^{s,p}_q$;

(4) *Hardy spaces* $\mathcal{H}^p$, the space BMO of functions of *bounded mean oscillation* (and the subspace VMO of functions of vanishing mean oscillation);

(5) The *Wiener algebra* $A$;

(6) *Morrey spaces* $M^p_q$;

(7) The space $M$ of finite measures;

(8) etc., etc.

As the above partial list indicates, there is an entire zoo of function spaces one could consider, and it can be difficult at first to see how they are organised with respect to each other. However, one can get some clarity in this regard by drawing a type diagram for the function spaces one is trying to study. A type diagram assigns a tuple (usually a pair) of relevant exponents to each function space. For function spaces $X$ on Euclidean space, two such exponents are the regularity $s$ of the space, and the integrability $p$ of the space. These two quantities are somewhat fuzzy in nature (and are not easily defined for all possible function spaces), but can basically be described as follows. We test the function space norm $\|f\|_X$ of a modulated rescaled bump function (or "wave packet")

$$f(x) := Ae^{ix\cdot\xi}\phi(\frac{x-x_0}{R}) \tag{5.15}$$

where $A > 0$ is an amplitude, $R > 0$ is a radius, $\phi \in C^\infty_c(\mathbf{R}^d)$ is a test function, $x_0$ is a position, and $\xi \in \mathbf{R}^d$ is a frequency of some magnitude $|\xi| \sim N$. One then studies how the norm $\|f\|_X$ depends on the parameters $A, R, N$. Typically, one has a relationship of the form

$$\|f\|_X \sim AN^s R^{d/p} \tag{5.16}$$

for some exponents $s, p$, at least in the high-frequency case when $N$ is large (in particular, from the uncertainty principle it is natural to require $N \gg 1/R$, and when dealing with inhomogeneous norms it is also natural to require $N \gg 1$). The exponent $s$ measures how sensitive the $X$ norm is to oscillation, and thus controls regularity; if $s$ is large, then oscillating functions will have large $X$ norm, and thus functions in $X$ will tend not to oscillate too much and thus be smooth. Similarly, the exponent $p$ measures how sensitive the $X$ norm is to the function $f$ spreading out to large scales; if $p$ is small, then slowly decaying functions will have large norm, so that functions in $X$ tend to decay quickly; conversely, if $p$ is large, then singular

[4]The conventions for the superscripts and subscripts here are highly variable, and vary from text to text.

functions will tend to have large norm, so that functions in $X$ will tend to not have high peaks.

Note that the exponent $s$ in (5.16) could be positive, zero, or negative, however the exponent $p$ should be non-negative, since intuitively enlarging $R$ should always lead to a larger (or at least comparable) norm. Finally, the exponent in the $A$ parameter should always be 1, since norms are by definition homogeneous. Note also that the position $x_0$ plays no role in (1); this reflects the fact that most of the popular function spaces in analysis are translation-invariant.

The type diagram in Figure 1 plots the $s, 1/p$ indices of various spaces. The black dots indicate those spaces for which the $s, 1/p$ indices are fixed; the blue dots are those spaces for which at least one of the $s, 1/p$ indices are variable (and so, depending on the value chosen for these parameters, these spaces may end up in a different location on the type diagram than the typical location indicated here).

**Remark 5.2.1.** There are some minor cheats in this diagram, for instance for the Orlicz spaces $L \log L$ and $e^L$ one has to adjust (5.15) by a logarithmic factor. Also, the norms for the Schwartz space $\mathcal{S}$ are not translation-invariant and thus not perfectly describable by this formalism. This picture should be viewed as a visual aid only, and not as a genuinely rigorous mathematical statement. This type of diagram is also known as a *de Vore-Triebel diagram*.

The type diagram can be used to clarify some of the relationships between function spaces, such as Sobolev embedding. For instance, when working with inhomogeneous spaces (which basically identifies low frequencies $N \ll 1$ with medium frequencies $N \sim 1$, so that one is effectively always in the regime $N \gg 1$), then decreasing the $s$ parameter results in decreasing the right-hand side of (5.15). Thus, one expects the function space norms to get smaller (and the function spaces to get larger) if one decreases $s$ while keeping $p$ fixed. Thus, for instance, $W^{k,p}$ should be contained in $W^{k-1,p}$, and so forth. Note however that this inclusion is not available for homogeneous function spaces such as $\dot{W}^{k,p}$, in which the frequency parameter $N$ can be either much larger than 1 or much smaller than 1.

Similarly, if one is working in a compact domain rather than in $\mathbf{R}^d$, then one has effectively capped the radius parameter $R$ to be bounded, and so we expect the function space norms to get smaller (and the function spaces to get larger) as one increases $1/p$, thus for instance $L^2$ will be contained in $L^1$. Conversely, if one is working in a discrete domain[5] such as $\mathbf{Z}^d$, then the

[5]If the domain is both compact and discrete, then it is finite, and on a finite-dimensional space all norms are equivalent.

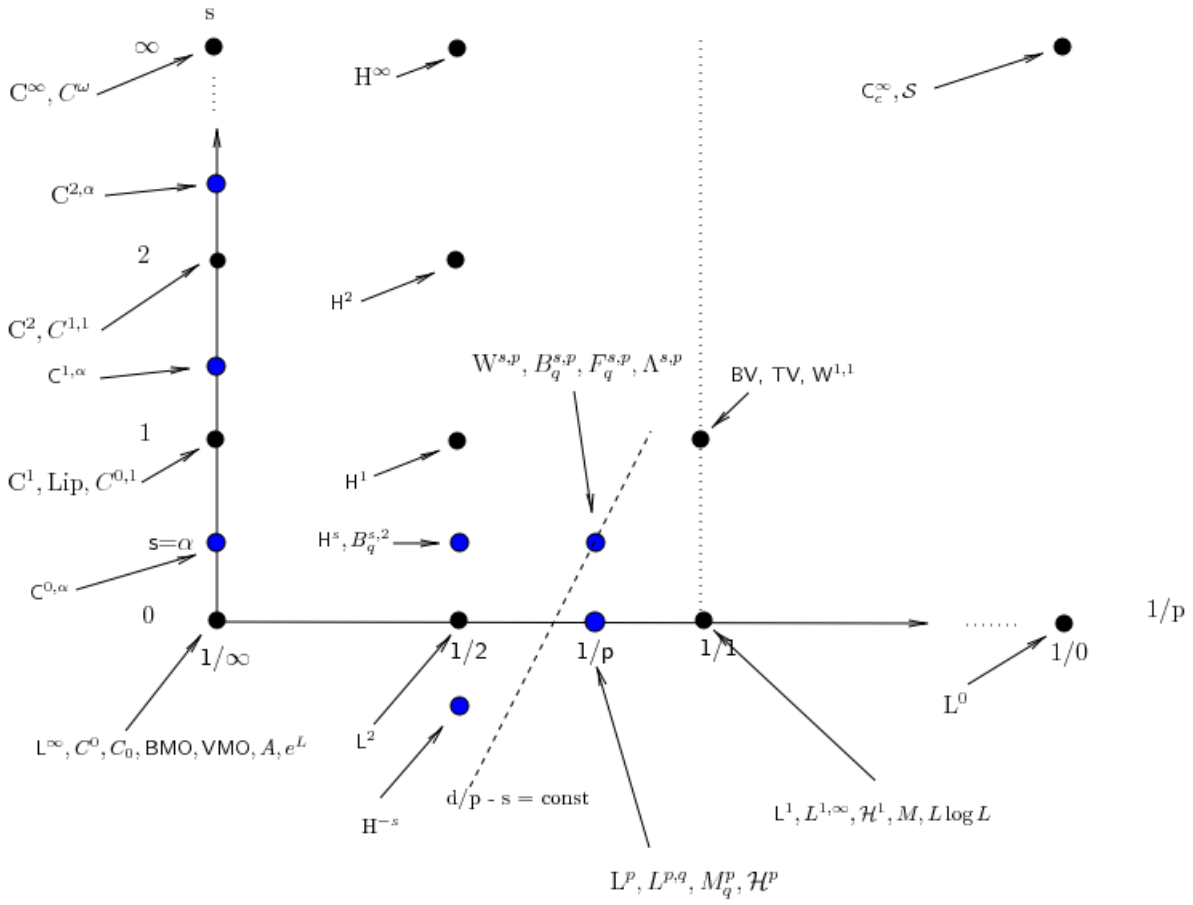


**Figure 1.** Function space type diagram.

radius parameter $R$ has now effectively been bounded from below, and the reverse should occur: the function spaces should get larger as one decreases $1/p$.

As mentioned earlier, the uncertainty principle suggests that one has the restriction $N \gg 1/R$. From this and (5.16), we expect to be able to enlarge the function space by trading in the regularity parameter $s$ for the integrability parameter $p$, keeping the dimensional quantity $d/p - s$ fixed. This is indeed how Sobolev embedding works. Note in some cases one runs out of regularity before p goes all the way to infinity (thus ending up at an $L^p$ space), while in other cases p hits infinity first. In the latter case, one can embed the Sobolev space into a Holder space such as $C^{k,\alpha}$.

On continuous domains, one can send the frequency $N$ off to infinity, keeping the amplitude $A$ and radius $R$ fixed. From this and (5.15) we see

that norms with a lower regularity $s$ can never hope to control norms with a higher regularity $s' > s$, no matter what one does with the integrability parameter. Note however that in discrete settings this obstruction disappears; when working on, say, $\mathbf{Z}^d$, then in fact one can gain as much regularity as one wishes for free, and there is no distinction between a Lebesgue space $\ell^p$ and their Sobolev counterparts $W^{k,p}$ in such a setting.

When interpolating between two spaces (using either the real or complex interpolation method), the interpolated space usually[6] has regularity and integrability exponents on the line segment between the corresponding exponents of the endpoint spaces. Typically, one can control the norm of the interpolated space by the geometric mean of the endpoint norms that is indicated by this line segment; again, this is plausible from looking at (5.16).

The space $L^2$ is self-dual. More generally, the dual of a function space $X$ will generally have type exponents that are the reflection of the original exponents around the $L^2$ origin; consider for instance the dual spaces $H^s, H^{-s}$ or $\mathcal{H}^1, \mathrm{BMO}$ in the above diagram.

Spaces whose integrability exponent $p$ is larger than 1 (i.e. which lie to the left of the dotted line) tend to be Banach spaces, while spaces whose integrability exponent is less than 1 are almost never[7] Banach spaces. The case $p = 1$ is borderline; some spaces at this level of integrability, such as $L^1$, are Banach spaces, while other spaces, such as $L^{1,\infty}$, are not.

While the regularity $s$ and integrability $p$ are usually the most important exponents in a function space (because amplitude, width, and frequency are usually the most important features of a function in analysis), they do not tell the entire story. One major reason for this is that the modulated bump functions (5.15), while an important class of test examples of functions, are by no means the only functions that one would wish to study. For instance, one could also consider sums of bump functions (5.15) at different scales. The behaviour of the function space norms on such spaces is often controlled by secondary exponents, such as the second exponent $q$ that arises in Lorentz spaces, Besov spaces, or Triebel-Lizorkin spaces. For instance, consider the function

$$f_M(x) := \sum_{m=1}^{M} 2^{-md} \phi(x/2^m), \tag{5.17}$$

where $M$ is a large integer, representing the number of distinct scales present in $f_M$. Any function space with regularity $s = 0$ and $p = 1$ should assign

[6] This can be heuristically justified from the formula (5.16) by thinking about how the real or complex interpolation methods actually work, as discussed for instance in [**Ta2010**, §1.11].

[7] This can be justified by covering a large ball into small balls and considering how (5.15) would interact with the triangle inequality in this case.

each summand $2^{-md}\phi(x/2^m)$ in (5.17) a norm of $O(1)$, so the norm of $f_M$ could be as large as $O(M)$ if one assumes the triangle inequality. This is indeed the case for the $L^1$ norm, but for the weak $L^1$ norm, i.e. the $L^{1,\infty}$ norm, $f_M$ only has size $O(1)$. More generally, for the Lorentz spaces $L^{1,q}$, $f_M$ will have a norm of about $O(M^{1/q})$. Thus we see that such secondary exponents can influence the norm of a function by an amount which is polynomial in the number of scales. In many applications, though, the number of scales is a "logarithmic" quantity and thus of lower order interest when compared against the "polynomial" exponents such as $s$ and $p$. So the fine distinctions between, say, strong $L^1$ and weak $L^1$, are only of interest in "critical" situations in which one cannot afford to lose any logarithmic factors (this is for instance the case in much of *Calderon-Zygmund theory*).

We have cheated somewhat by only working in the high frequency regime. When dealing with inhomogeneous spaces, one often has a different set of exponents for (5.15) in the low-frequency regime than in the high-frequency regime. In such cases, one sometimes has to use a more complicated type diagram to genuinely model the situation, e.g. by assigning to each space a convex set of type exponents rather than a single exponent, or perhaps having two separate type diagrams, one for the high frequency regime and one for the low frequency regime. Such diagrams can get quite complicated, and will probably not be much use to a beginner in the subject, though in the hands of an expert who knows what he or she is doing, they can still be an effective visual aid.

## 5.3. Amplitude-frequency dynamics for semilinear dispersive equations

Semilinear dispersive and wave equations, of which the *defocusing nonlinear wave equation*

$$-\partial_{tt}u + \Delta u = |u|^{p-1}u \tag{5.18}$$

is a typical example (where $p > 1$ is a fixed exponent, and $u : \mathbf{R}^{1+n} \to \mathbf{R}$ is a scalar field), can be viewed as a "tug of war" between a linear dispersive equation, in this case the *linear wave equation*

$$-\partial_{tt}u + \Delta u = 0 \tag{5.19}$$

and a nonlinear ODE, in this case the equation

$$-\partial_{tt}u = |u|^{p-1}u. \tag{5.20}$$

If the nonlinear term was not present, leaving only the dispersive equation (5.19), then as the term "dispersive" suggests, in the asymptotic limit $t \to \infty$, the solution $u(t,x)$ would spread out in space and decay in amplitude. For instance, in the model case when $d = 3$ and the initial position

$u(0,x)$ vanishes (leaving only the initial velocity $u_t(0,x)$ as non-trivial initial data), the solution[8] $u(t,x)$ for $t > 0$ is given by the formula

$$u(t,x) = \frac{1}{4\pi t} \int_{|y-x|=t} u_t(0,y)\ d\sigma$$

where $d\sigma$ is surface measure on the sphere $\{y \in \mathbf{R}^3 : |y-x| = t\}$. Thus, if the initial velocity was bounded and compactly supported, then the solution $u(t,x)$ would be bounded by $O(1/t)$ and would thus would decay uniformly to zero as $t \to \infty$. Similar phenomena occur for all dimensions greater than 1.

Conversely, if the dispersive term was not present, leaving only the ODE (5.20), then one no longer expects decay; indeed, given the conserved energy $\frac{1}{2}u_t^2 + \frac{1}{p+1}|u|^{p+1}$ for the ODE (5.20), we do not expect any decay at all (and indeed, solutions are instead periodic in time for each fixed $x$, as can easily be seen by viewing the ODE (and the energy curves) in phase space).

Depending on the relative "size" of the dispersive term $\Delta u$ and the nonlinear term $|u|^{p-1}u$, one can heuristically describe the behaviour of a solution $u$ at various positions at times as either being *dispersion dominated* (in which $|\Delta u| \gg |u|^p$), *nonlinearity dominated* (in which $|u|^p \gg |\Delta u|$), or *contested* (in which $|\Delta u|$, $|u|^p$ are comparable in size). Very roughly speaking, when one is in the dispersion dominated regime, then *perturbation theory* becomes effective, and one can often show that the solution to the nonlinear equation indeed behaves like the solution to the linear counterpart, in particular exhibiting decay as $t \to \infty$. In principle, perturbation theory is also available in the nonlinearity dominated regime (in which the dispersion is now viewed as the perturbation, and the nonlinearity as the main term), but in practice this is often difficult to apply (due to the nonlinearity of the approximating equation and the large number of derivatives present in the perturbative term), and so one has to fall back on non-perturbative tools, such as conservation laws and monotonicity formulae. The contested regime is the most interesting, and gives rise to intermediate types of behaviour that are not present in the purely dispersive or purely nonlinear equations, such as solitary wave solutions (solitons) or solutions that blow up in finite time.

In order to analyse how solutions behave in each of these regimes rigorously, one usually works with a variety of function spaces (such as Lebesgue spaces $L^p$ and Sobolev spaces $H^s$). As such, one generally needs to first establish a number of function space estimates (e.g. Sobolev inequalities,

[8] To avoid technical issues, let us restrict attention in this section to classical (smooth) solutions.

Hölder-type inequalities, Strichartz estimates, etc.) in order to study these equations at the formal level.

Unfortunately, this emphasis on function spaces and their estimates can obscure the underlying physical intuition behind the dynamics of these equations, and the field of analysis of PDE sometimes acquires a reputation for being unduly technical as a consequence. However, as noted in Section 5.2, one can view function space norms as a way to formalise the intuitive notions of the "height" (amplitude) and "width" (wavelength) of a function (wave).

It turns out that one can similarly analyse the behaviour of nonlinear dispersive equations on a similar heuristic level, as that of understanding the dynamics as the amplitude $A(t)$ and wavelength $1/N(t)$ (or frequency $N(t)$) of a wave. Below the fold I give some examples of this heuristic; for sake of concreteness I restrict attention to the nonlinear wave equation (5.18), though one can of course extend this heuristic to many other models also. Rigorous analogues of the arguments here can be found in several places, such as the [**ShSt1998**] or [**Ta2006b**].

**5.3.1. Bump functions.** To initiate the heuristic analysis, we make the assumption that any given time $t$, the wave $u(t,x)$ "resembles" (or is "dominated" by) a bump function

$$u(t,x) \approx A(t)e^{i\theta(t)}\phi(N(t)(x-x(t))) \tag{5.21}$$

of some amplitude $A(t) > 0$, some phase $\theta(t) \in \mathbf{R}$, some frequency $N(t) > 0$, and some position $x(t) \in \mathbf{R}^d$, where $\phi$ is a bump function. We will leave the terms "resembles" and "dominated" deliberately vague; $u(t,x)$ might not consist entirely of this bump function, but could instead be a superposition of multiple components, with this bump function being the "strongest" of these components in some sense. It is of course possible for a solution to concentrate its mass and energy in a different configuration than a bump function; but experience has shown that the most nonlinear behaviour tends to occur when such a concentration occurs, and so this ansatz is expected to capture the "worst-case" behaviour of the solution[9]. In particular, this type of bump function dominance is often seen when the solution exhibits soliton or near-soliton like behaviour, and often occurs shortly prior to blowup (especially for equations with critical nonlinearity). There are a variety of tools to formalise these sorts of intuitions, such as concentration-compactness and the induction-on-energy method, but we will not focus on these tools here.

[9]Basically, if a wave splits its energy into too many distinct components, then the nonlinear effects of each component become quite weak, even when superimposed back together again.

**Remark 5.3.1.** One can also refine the above ansatz in a number of ways, for instance by also introducing a frequency modulation $e^{ix\cdot\xi(t)}$, which is particularly important in models such as the mass-critical NLS which admit a frequency modulation symmetry, but for simplicity we will not consider this more complicated situation here.

For this analysis, we shall ignore the role of the phase $\phi(t)$ and position $x(t)$, focusing instead on the amplitude $A(t)$ and frequency $N(t)$. This collapses the infinite numbers of degrees of freedom for the wave $u(t)$ down to just two degrees of freedom. Of course, there is a significant amount of information lost when performing this collapse - in particular, the exact PDE (5.18) will no longer retain its deterministic form when projected to these two coordinates - but one can still discern non-trivial features of the original dynamics from this two-dimensional viewpoint.

With the ansatz (5.21), the solution $u(t,x)$ has magnitude comparable $A(t)$ to a ball of radius roughly $1/N(t)$. As a consequence, the nonlinearity $|u|^{p-1}u$ will have magnitude about $A(t)^p$ on this ball. Meanwhile, the dispersive term $\Delta u$ would be expected to have magnitude about $A(t)N(t)^2$ (using a crude "rise-over-run" interpretation of the derivative, or else just computing the Laplacian of (5.21) explicitly). We thus expect *dispersion dominant* behaviour when $A(t)N(t)^2 \gg A(t)^p$, or in other words when

$$A(t) \ll N(t)^{2/(p-1)}, \tag{5.22}$$

*nonlinearity dominant* behaviour when $A(t)N(t)^2 \ll A(t)^p$, or in other words when

$$A(t) \gg N(t)^{2/(p-1)}, \tag{5.23}$$

and *contested* behaviour when $A(t)N(t)^2$ is comparable to $A(t)^p$, or in other words when

$$A(t) \sim N(t)^{2/(p-1)}. \tag{5.24}$$

The evolution of the parameters $A(t), N(t)$ is partly constained by a variety of conservation laws and monotonicity formulae. Consider for instance the energy conservation law, which asserts that the *energy*

$$E = \int_{\mathbf{R}^d} \frac{1}{2}|u_t|^2 + \frac{1}{2}|\nabla u|^2 + \frac{1}{p+1}|u|^{p+1}\ dx$$

is conserved in time. Inserting the ansatz (5.21) into the right-hand side, we obtain the heuristic bound[10]

$$A_t(t)^2 N(t)^{-d} + A(t)^2 N(t)^{2-d} + A(t)^{p+1} N(t)^{-d} \ll E.$$

[10]We only write an upper bound here for the left-hand side, and not a lower bound, to allow for the possibility that most of the energy of the wave is not invested in the bump function (5.21), but is instead dispersed elsewhere.

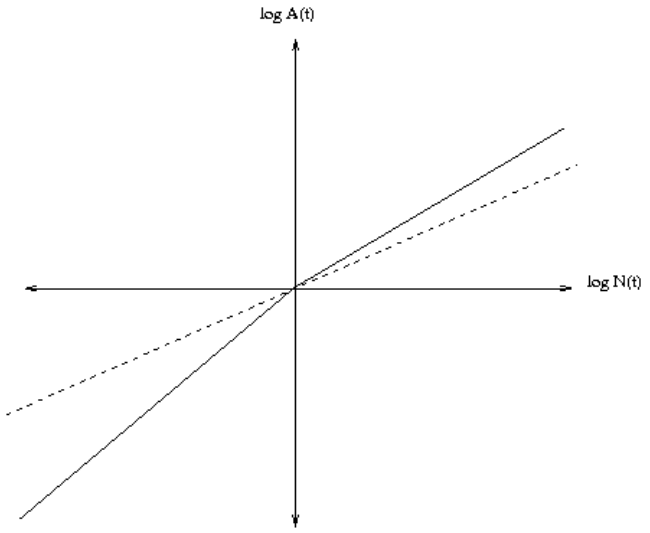


**Figure 2.** Supercritical amplitude-frequency dynamics.

This gives us the *a priori* bounds

$$A(t) \ll E^{1/2} N(t)^{(d-2)/2}, E^{1/(p+1)} N(t)^{d/(p+1)} \tag{5.25}$$

and

$$A_t(t) \ll E^{1/2} N(t)^{d/2}. \tag{5.26}$$

The bounds (5.25) can be viewed as describing a sort of "energy surface" that the parameters $A(t)$, $N(t)$ can vary in.

It is instructive to see how these bounds interact with the criteria (5.22), (5.23), (5.24), for various choices of dimension $d$ and exponent $p$. Let us first see what happens in a *supercritical* setting, such as $d = 3$ and $p = 7$, with bounded energy $E = O(1)$. In this case, the energy conservation law gives the bounds

$$A(t) \ll \min(N(t)^{1/2}, N(t)^{3/8}).$$

Meanwhile, the threshold between dispersive behaviour and nonlinear behaviour is when $A(t) \sim N(t)^{1/3}$. We can illustrate this by performing a log-log plot between $\log N(t)$ and $\log A(t)$; see Figure 2.

The region below the dotted line corresponds to dispersion-dominated behaviour, and the region above corresponds to nonlinearity-dominated behaviour. The region below the solid line corresponds to the possible values of amplitude and frequency that are permitted by energy conservation.

This diagram illustrates that for low frequencies $N(t) \ll 1$, the energy constraint ensures dispersive behaviour; but for high frequencies $N(t) \gg 1$, one can venture increasingly far into the nonlinearity dominated regime

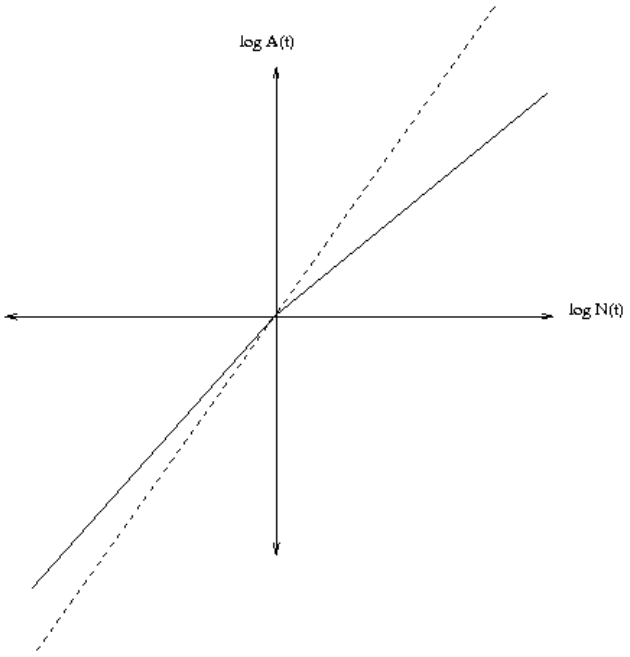


**Figure 3.** Subcritical amplitude-frequency dynamics.

while still being consistent with energy conservation. In particular, energy conservation does not prevent a scenario in which the frequency and amplitude both increase to infinity in finite time, while staying inside the nonlinearity dominated regime. And indeed, global regularity for this supercritical equation is a notoriously hard open problem, analogous in many ways to the even more famous global regularity problem for Navier-Stokes (see [**Ta2008**, §2.4] for further discussion).

In contrast, let us consider a *subcritical* setting, such as $d = 3$ and $p = 3$, again with bounded energy $E = O(1)$. Now, the energy conservation law gives the bounds

$$A(t) \ll \min(N(t)^{1/2}, N(t)^{3/4})$$

while the threshold between dispersive behaviour and nonlinear behaviour is when $A(t) \sim N(t)$. The log-log plot is now illustrated in Figure 3.

We now see that for high frequencies $N(t) \gg 1$, the energy constraint ensures dispersive behaviour; but conversely, for low frequencies $N(t) \ll 1$, one can have highly nonlinear behaviour. On the other hand, low frequencies cannot exhibit finite time blowup (as can be inferred from (5.26)); however, other non-dispersive scenarios exist, such as a soliton-type solution in which $N(t)$ is low-frequency but essentially constant in time, or a self-similar decay in which $N(t)$ and $A(t)$ go slowly to 0 as $t \to \infty$, while staying out of the

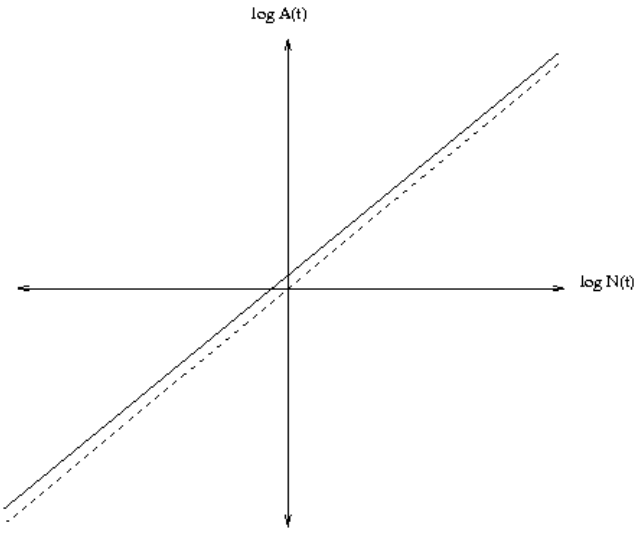


**Figure 4.** Critical amplitude-frequency dynamics.

dispersion-dominated regime. Again, this is reflected in the known theory for this equation for large (but finite energy) data: global regularity is known (there is no blowup), but it is unknown whether the solution disperses like a linear solution in the limit $t \to \infty$.

Finally, we look at a critical setting, in which $d = 3$ and $p = 5$. Here, energy conservation gives the bounds

$$A(t) \ll \min(E^{1/2}, E^{1/6}) N(t)^{1/2}$$

and the threshold between dispersive and nonlinear behaviour is $A(t) \sim N(t)^{1/2}$. Thus, when the energy $E$ is small, one expects only dispersive behaviour; and when the energy is large, then both dispersive and contested behaviour (at both high and low frequencies) are consistent with energy conservation. The large energy case is depicted in Figure 4, with the solid line slightly above the dashed line; in the small energy case, the positions of the two lines are reversed.

Now, it turns out that for small energy, one indeed has global regularity and scattering (for both the defocusing nonlinearity $+|u|^4 u$ and the focusing nonlinearity $-|u|^4 u$), which is consistent with the above heuristics. For large energy, blowup can occur in the focusing case, but in the defocusing case what happens is that the solution can linger in the contested region for a finite period of time, but eventually the nonlinearity "concedes" to the dispersion, and the solution enters the dispersion-dominated regime and scatters. This cannot be deduced solely from energy conservation, but requires some additional inputs.

First, let us assume for contradiction that one never enters the dispersion-dominated regime, but instead remains in the contested regime $A(t) \sim N(t)^{1/2}$ throughout. Then from (5.26) we see that for any time $t_0$, the quantities $A(t)$ and $N(t)$ will not change in magnitude much in the time interval $\{t : t = t_0 + O(1/N(t_0))\}$. This means that one can subdivide time into intervals $I$, with $N(t)$ comparable to $|I|^{-1}$ on this time interval, and $A(t)$ comparable to $|I|^{-1/2}$. The asymptotic behaviour of the solution is then encoded in the combinatorial structure of these intervals. For instance, a soliton-like solution would correspond to a string of intervals $I$, all of roughly the same size, while a finite time blowup would correspond to a shrinking sequence of intervals converging to the blowup time, whereas a slow decay at infinity would be represented by a sequence of intervals of increasing length going off to infinity[11].

One can also take a spacetime view instead of a temporal view, and view the solution as a string of spacetime "bubbles", each of which has some lifespan $I$ and amplitude comparable to $|I|^{-1/2}$, and lives on a spacetime cube of sidelength comparable to $|I|$. If the number of bubbles is finite, then the nonlinearity eventually concedes and one has dispersive behaviour; if instead the number of bubbles is infinite, then one has contested behaviour that can lead either to finite time blowup or infinite time blowup (where blowup is defined here as failure of dispersion to dominate asymptotically, rather than formation of a singularity). While "number of bubbles" is not a precise quantity, in the rigorous theory of the critical NLW, one uses more quantitative expressions, such as the $L^8$ norm

$$\int_{I \times \mathbf{R}^3} |u(t,x)|^8 \, dxdt$$

or variants such as $\|u\|^4_{L_t^4 L_x^{12}(I \times \mathbf{R}^3)}$, as proxies for this concept. Note that each bubble contributes an amount comparable to unity to each of the above expressions. This may help explain why obtaining bounds for these types of norms is so key to establishing global regularity and scattering for this equation.

The next ingredient is the *Morawetz inequality*

$$\int_{\mathbf{R}} \int_{\mathbf{R}^3} \frac{|u(t,x)|^6}{|x|} \, dxdt \ll E$$

which can be established by an integration by parts argument. This inequality is easiest to exploit in the model case of spherical symmetry. To be consistent with the ansatz (5.21), we must have $x(t) = 0$ in this case. We

[11] See [**Ta2006b**, Chapter 5] for various depictions of these bubble evolutions.

then have

$$\int_{\mathbf{R}^3} \frac{|u(t,x)|^6}{|x|}\ dx \gg N(t)$$

and so

$$\int_{\mathbf{R}} N(t)\ dt \ll E.$$

Each bubble of contested dynamics contributes roughly a unit amount to the integral on the left, and so the Morawetz inequality bounds the total number of bubbles and thus is a mechanism for forcing dispersive behaviour asymptotically.

In the non-spherically symmetric case, the position $x(t)$ of the bubble can vary. However, finite speed of propagation heuristics indicate that this position[12] cannot move faster than the speed of light, which is normalised to be 1, thus $x(t') - x(t) = O(|t' - t|)$. The Morawetz inequality then instead gives bounds such as

$$\int_{\mathbf{R}} \min(N(t), \frac{1}{|x(t)|})\ dt \ll E.$$

In the model case when $N(t)$ stay sroughly bounded, $x(t)$ can only grow at most linearly in $t$, and the logarithmic divergence of the integral $\int \frac{1}{t}\ dt$ at infinity then again forces the number of bubbles to be finite (but this time the bound is exponential in the energy, rather than polynomial; see [**Na1999**], [**Ta2006**] for further discussion).

One can extend this heuristic analysis to explain why the global regularity results for the energy-critical equation can extend very slightly to the supercritical regime, and in particular (in the spherically symmetric case) to the logarithmically supercritical equation

$$-\partial_{tt} u + \Delta u = |u|^4 u \log(2 + |u|^2)$$

as was done in [**Ta2007**], [**Ro2009**]. This equation behaves more or less identically to the critical NLW for low frequencies $N(t) \ll 1$, but exhibits slightly different behaviour for high frequencies $N(t) \gg 1$. In this regime, the dividing line between dispersive and nonlinear behaviour is now $A(t) \sim N(t)^{1/2} \log^{-1/4} N(t)$. Meanwhile, the energy bounds (assuming bounded energy) now give

$$A(t) \ll N(t)^{1/2} \log^{-1/6} N(t)$$

[12]The situation is more complicated if one generalises the ansatz (5.21) to allow for the solution $u$ to consist of a superposition of several bump functions at several different places for each point in time. However, for the critical equation and in the contested regime, each bump function absorbs an amount of energy bounded from below, and so there can only be a bounded number of such bumps existing at any given time; as such, one should morally be able to decompose the evolution into independent "particle-like" components, each of which obeys finite speed of propagation.

so that there is now a logarithmically wide window of opportunity for nonlinear behaviour at high frequencies.

The energy bounds also give

$$A_t(t) \ll N(t)^{3/2}$$

from (5.26), but we can do a little bit better if we invoke the heuristic of *equipartition of energy*, which states that the kinetic portion $\int_{\mathbf{R}^3} \frac{1}{2}|u_t|^2\ dx$ of the energy is roughly in balance with the potential portion $\int_{\mathbf{R}^3} \frac{1}{2}|\nabla u|^2 + V(u)\ dx$ (where $V(x)$ is the antiderivative of $x^5 \log(2+x^2)$). There are several ways to make this heuristic precise; one is to start with the identity

$$\partial_t \int_{\mathbf{R}^3} u u_t(x) = \int_{\mathbf{R}^3} |u_t|^2 - |\nabla u|^2 - |u|^6 \log(2+|u|^2)\ dx$$

which suggests (together with the fundamental theorem of calculus) that the right-hand side should average out to zero after integration in time. Using this heuristic, one is soon led to the slight improvement

$$A_t(t) \ll A(t)N(t) + A(t)^3 \log^{1/2} N(t)$$

of the previous bound.

The contested regions of the evolution then break up into bubbles in spacetime, each of which has some length and lifespan comparable to $1/N$, and amplitude $A$ comparable to $N^{1/2}\log^{-1/4} N$ (in the high-frequency case $N \gg 1$).

In contrast, the Morawetz inequality for this equation asserts that

$$\int_{\mathbf{R}} \int_{\mathbf{R}^3} \frac{|u(t,x)|^6 \log(2+|u(t,x)|^2)}{|x|}\ dxdt \ll 1. \tag{5.27}$$

In the spherically symmetric case, a bubble of length $1/N$ and amplitude $A$ with $N \gg 1$ contributes about $A^6 N^3 (\log A) \gg \log^{-1/2} N$ to the integral in (5.27). This quantity goes to zero as $N \to \infty$, but very slowly; in particular, as $N$ increases to infinity along dyadic scales $N = 2^k$, the sum $\log^{-1/2} N$ is divergent, which explains why the nonlinearity cannot sustain an infinite chain of such bubbles[13].

## 5.4. The Euler-Arnold equation

A (smooth) *Riemannian manifold* is a smooth manifold $M$ without boundary, equipped with a Riemannian metric[14] g, which assigns a length $|v|_{\mathrm{g}(x)} \in$

[13] It also suggests that perhaps the logarithmic supercriticality is not quite the right threshold here, indeed, this threshold has recently been improved upon by Hsi-Wei Shih (private communication).

[14] We use Roman font for $g$ here, as we will need to use $g$ to denote group elements later in this post.

$\mathbf{R}^+$ to every tangent vector $v \in T_x M$ at a point $x \in M$, and more generally assigns an inner product

$$\langle v, w \rangle_{\mathrm{g}(x)} \in \mathbf{R}$$

to every pair of tangent vectors $v, w \in T_x M$ at a point $x \in M$. This inner product is assumed to symmetric, positive definite, and smoothly varying in $x$, and the length is then given in terms of the inner product by the formula

$$|v|^2_{\mathrm{g}(x)} := \langle v, v \rangle_{\mathrm{g}(x)}.$$

In coordinates (and also using *abstract index notation*), the metric g can be viewed as an invertible symmetric rank $(0,2)$ tensor $\mathrm{g}_{ij}(x)$, with

$$\langle v, w \rangle_{\mathrm{g}(x)} = \mathrm{g}_{ij}(x) v^i w^j.$$

One can also view the Riemannian metric as providing a (self-adjoint) identification between the *tangent bundle* $TM$ of the manifold and the *cotangent bundle* $T^*M$; indeed, every tangent vector $v \in T_x M$ is then identified with the cotangent vector $\iota_{TM \to T^*M}(v) \in T^*_x M$, defined by the formula

$$\iota_{TM \to T^*M}(v)(w) := \langle v, w \rangle_{\mathrm{g}(x)}.$$

In coordinates, $\iota_{TM \to T^*M}(v)_i = \mathrm{g}_{ij}\, v^j$.

A fundamental dynamical system on the tangent bundle (or equivalently, the cotangent bundle, using the above identification) of a Riemannian manifold is that of *geodesic flow*. Recall that geodesics are smooth curves $\gamma : [a,b] \to M$ that minimise the length

$$|\gamma| := \int_a^b |\gamma'(t)|_{\mathrm{g}(\gamma(t))}\, dt.$$

There is some degeneracy in this definition, because one can reparameterise the curve $\gamma$ without affecting the length. In order to fix this degeneracy (and also because the square of the speed is a more tractable quantity analytically than the speed itself), it is better if one replaces the length with the *energy*

$$E(\gamma) := \frac{1}{2} \int_a^b |\gamma'(t)|^2_{\mathrm{g}(\gamma(t))}\, dt.$$

Minimising the energy of a parameterised curve $\gamma$ turns out to be the same as minimising the length, together with an additional requirement that the speed $|\gamma'(t)|_{\mathrm{g}(\gamma(t))}$ stay constant in time. Minimisers (and more generally, critical points) of the energy functional (holding the endpoints fixed) are known as *geodesic flows*. From a physical perspective, geodesic flow governs the motion of a particle that is subject to no external forces and thus moves freely, save for the constraint that it must always lie on the manifold $M$.

One can also view geodesic flows as a dynamical system on the tangent bundle (with the state at any time $t$ given by the position $\gamma(t) \in M$ and the velocity $\gamma'(t) \in T_{\gamma(t)} M$) or on the cotangent bundle (with the state

then given by the position $\gamma(t) \in M$ and the *momentum* $\iota_{TM \to T^*M}(\gamma'(t)) \in T^*_{\gamma(t)}M$). With the latter perspective (sometimes referred to as *cogeodesic flow*), geodesic flow becomes a Hamiltonian flow, with Hamiltonian $H : T^*M \to \mathbf{R}$ given as

$$H(x,p) := \frac{1}{2}\langle p, p\rangle_{\mathrm{g}(x)^{-1}} = \frac{1}{2}\,\mathrm{g}^{ij}(x)p_i p_j$$

where $\langle , \rangle_{\mathrm{g}(x)^{-1}} : T^*_x M \times T^*_x M \to \mathbf{R}$ is the inverse inner product to $\langle , \rangle_{\mathrm{g}(x)} : T_x M \times T_x M \to \mathbf{R}$, which can be defined for instance by the formula

$$\langle p_1, p_2\rangle_{\mathrm{g}(x)^{-1}} = \langle \iota^{-1}_{TM \to T^*M}(p_1), \iota^{-1}_{TM \to T^*M}(p_2)\rangle_{\mathrm{g}(x)}.$$

In coordinates, geodesic flow is given by Hamilton's equations of motion

$$\frac{d}{dt}x^i = \mathrm{g}^{ij}\,p_j; \quad \frac{d}{dt}p_i = -\frac{1}{2}(\partial_i\,\mathrm{g}^{jk}(x))p_j p_k.$$

In terms of the velocity $v^i := \frac{d}{dt}x^i = \mathrm{g}^{ij}\,p_j$, we can rewrite these equations as the geodesic equation

$$\frac{d}{dt}v^i = -\Gamma^i_{jk}v^j v^k$$

where

$$\Gamma^i_{jk} = \frac{1}{2}\,\mathrm{g}^{im}(\partial_k\,\mathrm{g}_{mj} + \partial_j\,\mathrm{g}_{mk} - \partial_m\,\mathrm{g}_{jk})$$

are the *Christoffel symbols*; using *the Levi-Civita connection* $\nabla$, this can be written more succinctly as

$$(\gamma^*\nabla)_t v = 0.$$

If the manifold $M$ is an embedded submanifold of a larger Euclidean space $R^n$, with the metric g on $M$ being induced from the standard metric on $\mathbf{R}^n$, then the geodesic flow equation can be rewritten in the equivalent form

$$\gamma''(t) \perp T_{\gamma(t)}M,$$

where $\gamma$ is now viewed as taking values in $\mathbf{R}^n$, and $T_{\gamma(t)}M$ is similarly viewed as a subspace of $\mathbf{R}^n$. This is intuitively obvious from the geometric interpretation of geodesics: if the curvature of a curve $\gamma$ contains components that are transverse to the manifold rather than normal to it, then it is geometrically clear that one should be able to shorten the curve by shifting it along the indicated transverse direction. It is an instructive exercise to rigorously formulate the above intuitive argument. This fact also conforms well with one's physical intuition of geodesic flow as the motion of a free particle constrained to be in $M$; the normal quantity $\gamma''(t)$ then corresponds to the *centripetal force* necessary to keep the particle lying in $M$ (otherwise it would fly off along a tangent line to $M$, as per Newton's first law). The precise value of the normal vector $\gamma''(t)$ can be computed via the *second fundamental form* as $\gamma''(t) = \Pi_{\gamma(t)}(\gamma'(t), \gamma'(t))$, but we will not need this formula here.

In a beautiful paper from 1966, Arnold [**Ar1966**] observed that many basic equations in physics, including the Euler equations of motion of a rigid body, and also (by what is *a priori* a remarkable coincidence) the Euler equations of fluid dynamics of an inviscid incompressible fluid, can be viewed (formally, at least) as geodesic flows on a (finite or infinite dimensional) Riemannian manifold. And not just any Riemannian manifold: the manifold is a Lie group (or, to be truly pedantic, a *torsor* of that group), equipped with a right-invariant (or left-invariant, depending on one's conventions) metric. In the context of rigid bodies, the Lie group is the group $SE(3) = \mathbf{R}^3 \rtimes SO(3)$ of rigid motions; in the context of incompressible fluids, it is the group $\mathrm{Sdiff}(\mathbf{R}^3)$ of measure-preserving diffeomorphisms. The right-invariance makes the Hamiltonian mechanics of geodesic flow in this context (where it is sometimes known as the *Euler-Arnold equation* or the *Euler-Poisson equation*) quite special; it becomes (formally, at least) completely integrable, and also indicates (in principle, at least) a way to reformulate these equations in a Lax pair formulation. And indeed, many further completely integrable equations, such as the Korteweg-de Vries equation, have since been reinterpreted as Euler-Arnold flows.

From a physical perspective, this all fits well with the interpretation of geodesic flow as the free motion of a system subject only to a physical constraint, such as rigidity or incompressibility. (I do not know, though, of a similarly intuitive explanation as to why the Korteweg de Vries equation is a geodesic flow.)

One consequence of being a completely integrable system is that one has a large number of conserved quantities. In the case of the Euler equations of motion of a rigid body, the conserved quantities are the linear and angular momentum (as observed in an external reference frame, rather than the frame of the object). In the case of the two-dimensional Euler equations, the conserved quantities are the pointwise values of the vorticity (as viewed in *Lagrangian coordinates, rather than Eulerian coordinates*). In higher dimensions, the conserved quantity is now the (Hodge star of) the vorticity, again viewed in Lagrangian coordinates. The vorticity itself then evolves by the *vorticity equation*, and is subject to *vortex stretching* as the diffeomorphism between the initial and final state becomes increasingly sheared.

The elegant Euler-Arnold formalism is reasonably well-known in some circles (particularly in Lagrangian and symplectic dynamics, where it can be viewed as a special case of the *Euler-Poincaré formalism* or *Lie-Poisson formalism* respectively), but not in others; I for instance was only vaguely aware of it until recently, and I think that even in fluid mechanics this

perspective to the subject is not always emphasised[15]. I therefore have chosen to describe some of the conclusions of Arnold's original paper here.

In order to avoid technical issues, I will work formally, ignoring questions of regularity or integrability, and pretending that infinite-dimensional manifolds behave in exactly the same way as their finite-dimensional counterparts. In the finite-dimensional setting, it is not difficult to make all of the formal discussion below rigorous; but the situation in infinite dimensions is substantially more delicate[16]. However, I do not want to discuss these analytic issues here; see [**EbMa1970**] for a treatment of these topics.

**5.4.1. Geodesic flow using a right-invariant metric.** Let $G$ be a Lie group. From a physical perspective, one should think of a group element $g$ as describing the relationship between a fixed reference observer $O$, and a moving object $A = gO$; mathematically, one could think of $A$ and $O$ as belonging to a (left)[17] *torsor* $M$ of $G$. For instance, in the case of rigid motions, $O$ would be the reference state of a rigid body, $A$ would be the current state, and $g = A/O$ would be the element of the rigid motion group $\mathrm{SE}(3) = \mathbf{R}^3 \rtimes \mathrm{SO}(3)$ that moves $O$ to $A$; $M$ is then the *configuration space* of the rigid body. Similarly, in the case of incompressible fluids, $O$ would be a reference state of the fluid (e.g. the initial state at time $t = 0$), $A$ would be the current state, and $g \in \mathrm{SDiff}(\mathbf{R}^3)$ would be the measure-preserving diffeomorphism required to map the location each particle of the fluid at $O$ to the corresponding location of the same particle at $A$. Again, $M$ would be the configuration space of the fluid.

Once one fixes the reference observer $O$, one can set up a bijection between the torsor $M$ and the group $G$; but one can also adopt a "coordinate-free" perspective in which the observer $O$ is not present, in which case one should keep $M$ and $G$ distinct. Strictly speaking, the geodesic flow we will introduce will be on $M$ rather than on $G$, but for some minor notational reasons it is convenient to fix a reference observer $O$ in order to identify the two objects.

Let $\mathfrak{g} := T_{\mathrm{id}}G$ denote the Lie algebra of $G$, i.e. the tangent space of $G$ at the identity. This Lie algebra can be identified with the tangent space $T_A M$ of a state $A$ in $M$ in two different ways: an intrinsic one that does not use a reference observer $O$, and an extrinsic one which does rely on this observer. Specifically, we have

---

[15] For a more modern treatment of these topics, see the [**ArKh1998**] or [**MaRa1999**].

[16] Indeed, it is a notorious open problem whether the Euler equations for incompressible fluids even forms a global continuous flow in a reasonable topology in the first place!

[17] One could also work with right torsors; this would require a number of sign conventions below to be altered.

(1) (Intrinsic identification) If $V \in T_M A$ is a tangent vector to $A$, we let $V/A \in \mathfrak{g}$ be the associated element of the Lie algebra defined infinitesimally as

$$A + \varepsilon V = (1 + \varepsilon V/A)A$$

modulo higher order terms for infinitesimal $\varepsilon$, or more traditionally by requiring $\gamma'(0)/\gamma(0) = g'(0)$ whenever $\gamma : \mathbf{R} \to M$, $g : \mathbf{R} \to G$ are smooth curves such that $\gamma(t) = g(t)\gamma(0)$. Conversely, if $X \in \mathfrak{g}$, we let $XA \in T_M A$ be the tangent vector at $A$ defined infinitesimally as

$$A + \varepsilon XA = (1 + \varepsilon X)A$$

modulo higher order terms for infinitesimal $\varepsilon$, or more traditionally by requiring $\gamma'(0) = g'(0)\gamma(0)$ whenever $\gamma : \mathbf{R} \to M$, $g : \mathbf{R} \to G$ are smooth curves such that $\gamma(t) = g(t)\gamma(0)$ (so $g(0) = id$). Clearly, these two operations invert each other.

(2) (Extrinsic identification) If $V \in T_M A$ is a tangent vector to $A$, and $A = gO$ for some fixed reference $O$, we let $g^{-1}V/O \in \mathfrak{g}$ be the element of the Lie algebra defined infinitesimally as

$$A + \varepsilon V = g((1 + \varepsilon g^{-1}V/O)O)$$

or more traditionally by requiring $g^{-1}\gamma'(0)/O = h'(0)$ whenever $\gamma : \mathbf{R} \to M$, $h : \mathbf{R} \to G$ are such that $\gamma(t) = gh(t)O$ and $h(0) = id$.

The distinction between intrinsic and extrinsic identifications is closely related to the distinction between *active* and *passive* transformations: $V/A$ denotes the direction in which $A$ must move in order to effect a change of $V$ in the apparent position of $A$ relative to any observer $O$, whereas $g^{-1}V/O$ is the (inverse of the) direction in which the *reference* $O$ would move to effect the same change in the apparent position. The two quantities are related to each other by conjugation:

$$g^{-1}V/O = g^{-1}(V/A)g; \quad V/A = g(g^{-1}V/O)g^{-1}$$

where we define conjugation $X \mapsto gXg^{-1}$ of a Lie algebra element $X$ by a Lie group element $g$ in the usual manner.

If $A(t) \in M$ is the state of a rigid body at time $t$, then $A'(t)/A(t)$ is the linear and angular velocity of $A(t)$ as measured in $A(t)$'s current spatial reference frame, while if $A(t) = g(t)O$, then $g(t)^{-1}A'(t)/O$ is the linear and angular velocity of $A(t)$ as measured in the frame of $O$. Similarly, if $A(t) \in M$ is the state of an incompressible fluid at time $t$, then $A'(t)/A(t)$ is the velocity field $u(t)$ in *Eulerian coordinates*, while $g(t)^{-1}A'(t)/O$ is the velocity field $u \circ g(t)$ in *Lagrangian coordinates*.

The left action of $G$ $g : A \mapsto gA$ on the torsor $M$ induces a corresponding action $g : V \to gV$ on the tangent bundle $TM$. Indeed, this action was implicitly present in the notation $g^{-1}V/O$ used earlier.

Now suppose we choose a non-degenerate inner product $\langle , \rangle_{\mathfrak{g}}$ on the Lie algebra $\mathfrak{g}$. We do *not* assume any symmetries or invariances of this inner product with respect to the group structure, such as conjugation invariance (in particular, this inner product will usually *not* be the Cartan-Killing form). At any rate, once we select an inner product, we can construct a right-invariant Riemannian metric g on $M$ by the formula

$$\langle V, W \rangle_{\mathrm{g}(A)} := \langle V/A, W/A \rangle_{\mathfrak{g}}. \tag{5.28}$$

Because we do not require the inner product to be conjugation invariant, this metric will usually not be bi-invariant, instead being merely right-invariant.

The quantity $H(V) := \frac{1}{2}\langle V, V \rangle_{\mathrm{g}(A)}$ is the *Hamiltonian* associated to this metric. For rigid bodies, this Hamiltonian is the total kinetic energy of the body, which is the sum of the kinetic energy $\frac{1}{2}m|v|^2$ of the centre of mass, plus the rotational kinetic energy $\frac{1}{2}I(\omega, \omega)$ which is determined by the *moments of inertia* $I$. For incompressible fluids, the Hamiltonian is (up to a normalising constant) the energy $\frac{1}{2}\int_{\mathbf{R}^3} |u|^2 = \frac{1}{2}\int_{\mathbf{R}^3} |u \circ A|^2$ of the fluid, which can be computed either in Eulerian coordinates or in Lagrangian coordinates (there are no Jacobian factors here thanks to incompressibility).

Another important object in the Euler-Arnold formalism is the bilinear form $B : \mathfrak{g} \times \mathfrak{g} \to \mathfrak{g}$ associated to the inner product $\langle , \rangle$, defined via the Lie bracket and duality using the formula[18]

$$\langle [X, Y], Z \rangle = \langle B(Z, Y), X \rangle, \tag{5.29}$$

thus $B$ is a partial adjoint of the Lie bracket operator. Note that this form need not be symmetric. The importance of this form comes from the fact that it describes the geodesic flow:

**Theorem 5.4.1** (Euler-Arnold equation)**.** *Let $\gamma : \mathbf{R} \to M$ be a geodesic flow on $M$ using the right-invariant metric* g *defined above, and let $X(t) := \gamma'(t)/\gamma(t) \in \mathfrak{g}$ be the intrinsic velocity vector. Then $X$ obeys the equation*

$$\frac{d}{dt}X(t) = B(X(t), X(t)). \tag{5.30}$$

The Euler-Arnold equation is also known as the *Euler-Poincaré equation*; see for instance [**CeMaPeRa2003**] for further discussion.

**Proof.** For notational reasons, we will prove this in the model case when $G$ is a matrix group (so that we can place $G$, $M$, and $\mathfrak{g}$ in a common vector space, or more precisely a common matrix space); the general case is similar

[18] The conventions here differ slightly from those in Arnold's paper.

but requires more abstract notation. We consider a variation $\gamma(s,t)$ of the original curve $\gamma(t) = \gamma(s,t)$, and consider the first variation of the energy

$$\partial_s \frac{1}{2} \int_a^b \langle \partial_t \gamma(s,t), \partial_t \gamma(s,t) \rangle_{g(\gamma(s,t))} \, dt$$

which we write using (5.28) as

$$\partial_s \frac{1}{2} \int_a^b \langle \gamma_t \gamma^{-1}, \gamma_t \gamma^{-1} \rangle \, dt.$$

We move the derivative inside and use symmetry to write this as

$$\int_a^b \langle \partial_s(\gamma_t \gamma^{-1}), \gamma_t \gamma^{-1} \rangle \, dt$$

or

$$\int_a^b \langle \partial_s(\gamma_t \gamma^{-1}), X \rangle \, dt$$

We expand

$$\partial_s(\gamma_t \gamma^{-1}) = \gamma_{ts} \gamma^{-1} - \gamma_t \gamma^{-1} \gamma_s \gamma^{-1}.$$

Similarly

$$\partial_t(\gamma_s \gamma^{-1}) = \gamma_{ts} \gamma^{-1} - \gamma_s \gamma^{-1} \gamma_t \gamma^{-1}$$

and thus

$$\partial_s(\gamma_t \gamma^{-1}) = \partial_t(\gamma_s \gamma^{-1}) + [\gamma_s \gamma^{-1}, X].$$

Inserting this into the first variation and integrating by parts, we obtain

$$\int_a^b \langle [\gamma_s \gamma^{-1}, X], X \rangle - \langle \gamma_s \gamma^{-1}, \partial_t X \rangle \, dt;$$

using (5.29), this is

$$\int_a^b \langle \gamma_s \gamma^{-1}, B(X,X) \rangle - \langle \gamma_s \gamma^{-1}, \partial_t X \rangle \, dt$$

and so the first variation vanishes for arbitrary choices of perturbation $\gamma_s$ precisely when $\partial_t X = B(X,X)$, as required. □

It is instructive to verify that the Hamiltonian $H = \frac{1}{2}\langle X, X \rangle$ is preserved by this equation, as it should be. In the case of rigid motions, (5.30) is essentially *Euler's equations of motion*.

The right-invariance of the Riemannian manifold implies that the geodesic flow is similarly right-invariant. And this is reflected by the fact that the Euler-Arnold equation (5.30) does not involve the position $\gamma(t)$. This position of course evolves by the equation

$$\frac{d}{dt}\gamma = X\gamma \tag{5.31}$$

which is just the definition of $X$.

Note that while the velocity $X$ influences the evolution of the position $\gamma$, the position $\gamma$ does not influence the evolution of the velocity $X$. This is of course a manifestation of the right-invariance of the problem. This reduction of the flow is known as *Euler-Poincaré reduction*, and is essentially a basic example of both *Lagrangian reduction* and *symplectic reduction*, in which the symmetries of a Lagrangian or Hamiltonian evolution are used to reduce the dimension of the dynamics while preserving the Lagrangian or Hamiltonian structure.

We can rephrase the Euler equation in a *Lax pair* formulation by introducing the *Cartan-Killing form*

$$(X, Y) := \mathrm{tr}(\mathrm{ad}(X)\,\mathrm{ad}(Y)).$$

Like $\langle , \rangle$, the Cartan-Killing form $(,)$ is a symmetric bilinear form on the Lie algebra $\mathfrak{g}$. If we assume that the group $G$ is *semisimple* (and finite-dimensional), then this form will be non-degenerate. It obeys the identity

$$([X, Y], Z) = -(Y, [X, Z]), \tag{5.32}$$

thus adjoints are automatically skew-adjoint in the Cartan-Killing form.

If the Cartan-Killing form is non-degenerate, it can be used to express the inner product $\langle , \rangle$ via a formula of the form

$$\langle X, Y \rangle := (X, \Lambda^{-1} Y) \tag{5.33}$$

where $\Lambda : \mathfrak{g} \to \mathfrak{g}$ is an invertible linear transformation which is self-adjoint with respect to both $\langle , \rangle$ and $(,)$. We then define the *intrinsic momentum* $\mathbf{M}(t)$ of the Euler-Arnold flow $\gamma(t)$ by the formula

$$\mathbf{M}(t) := \Lambda^{-1} X(t).$$

From (5.30), we see that $\mathbf{M}$ evolves by the equation

$$\frac{d}{dt}\mathbf{M} := \Lambda^{-1} B(\Lambda\mathbf{M}, \Lambda\mathbf{M}).$$

But observe from (5.33), (5.29), (5.33), (5.32) that

$$\begin{aligned}(\Lambda^{-1} B(\Lambda\mathbf{M}, \Lambda\mathbf{M}), Y) &= \langle B(\Lambda\mathbf{M}, \Lambda\mathbf{M}), Y\rangle \\ &= \langle [Y, \Lambda\mathbf{M}], \Lambda\mathbf{M}\rangle \\ &= ([Y, \Lambda\mathbf{M}], \mathbf{M}) \\ &= ([\Lambda\mathbf{M}, \mathbf{M}], Y)\end{aligned}$$

for any test vector $Y \in \mathfrak{g}$, which by nondegeneracy implies that

$$\Lambda^{-1} B(\Lambda\mathbf{M}, \Lambda\mathbf{M}) = [\Lambda\mathbf{M}, \mathbf{M}]$$

leading to the Lax pair form

$$\frac{d}{dt}\mathbf{M} = [\Lambda\mathbf{M}, \mathbf{M}]$$

of the Euler-Arnold equation, known as the[19] *Lie-Poisson equation*. In particular, the spectrum of $\mathbf{M}$ is invariant, or equivalently $\mathbf{M}$ evolves along a single *coadjoint orbit* in $\mathfrak{g} \equiv \mathfrak{g}^*$.

By *Noether's theorem*, the right-invariance of the geodesic flow should create a conserved quantity (or *moment map*); as the right-invariance is an action of the group $G$, the conserved quantity should take place in the adjoint $\mathfrak{g}^*$. If we write $\gamma(t) = g(t)O$ for some fixed observer $O$, then this conserved quantity can be computed as the *extrinsic momentum*

$$P : Y \mapsto \langle X, gYg^{-1} \rangle, \tag{5.34}$$

thus $\omega$ is the 1-form associated to $X$, pulled back to extrinsic coordinates. Indeed, from (5.31) one has

$$\partial_t g = Xg$$

and thus

$$\partial_t g^{-1} = -g^{-1}X$$

and hence for any test vector $Y$

$$\begin{aligned} \partial_t P(Y) &= \langle X_t, gYg^{-1} \rangle + \langle X, [X, gYg^{-1}] \rangle \\ &= \langle B(X,X), gYg^{-1} \rangle - \langle B(X,X), gYg^{-1} \rangle \\ &= 0 \end{aligned}$$

thanks to (5.29), and the claim follows. Using the Cartan-Killing form, the extrinsic momentum can also be identified with $g^{-1}\mathbf{M}g$, thus linking the extrinsic and intrinsic momenta to each other.

**5.4.2. Incompressible fluids.** Now consider an incompressible fluid in $\mathbf{R}^3$, whose initial state is $O$ and whose state at any time $t$ is given as $\gamma(t)$. One can express $\gamma(t) = g(t)O$, where $g(t) \in \mathrm{Sdiff}(\mathbf{R}^3)$ is the diffeomorphism from $\mathbf{R}^3$ to itself that maps the location of each particle at $O$ to the location of the same particle at $\gamma(t)$. As the fluid is assumed incompressible, the diffeomorphism must be measure-preserving (and orientation preserving); we denote the group of such special diffeomorphisms as $\mathrm{Sdiff}(\mathbf{R}^3)$.

The Lie algebra to the group $\mathrm{Diff}(\mathbf{R}^3)$ of all diffeomorphisms, is the space of all (smooth) vector fields $X : \mathbf{R}^3 \to \mathbf{R}^3$. The Lie algebra of the subgroup $\mathrm{Sdiff}(\mathbf{R}^3)$ of measure-preserving diffeomorphisms is the space of all *divergence-free* vector fields; indeed, this is one of the primary motivations of introducing the concept of *divergence* of a vector field. We give both Lie algebras the usual $L^2$ inner product:

$$\langle u, v \rangle := \int_{\mathbf{R}^3} u \cdot v.$$

[19] The sign conventions here are the opposite of those in Arnold's paper, ultimately because I am assuming right-invariance instead of left-invariance.

The Lie bracket on $\mathrm{Sdiff}(\mathbf{R}^3)$ or $\mathrm{Diff}(\mathbf{R}^3)$ is the same as the usual Lie bracket of vector fields.

Let $u(t) := \gamma'(t)/\gamma(t) = g' \circ g^{-1}$ be the intrinsic velocity vector; then this is a divergence-free vector field, which physically represents the velocity field in Eulerian coordinates. The extrinsic velocity vector $g(t)^{-1}u(t)g(t) = u \circ g(t)$ is then the velocity field in Lagrangian coordinates; it is also divergence-free.

If there were no constraint of incompressibility (i.e. if one were working in $\mathrm{Diff}(\mathbf{R}^3)$ rather than $\mathrm{Sdiff}(\mathbf{R}^3)$), then the metric is flat, and the geodesic equation of motion is simply given by Newton's first law

$$\frac{d^2}{dt^2} g(t) = 0$$

or in terms of the intrinsic velocity field $u$,

$$\partial_t u(t) + (u \cdot \nabla)u = 0.$$

Once we restrict to incompressible fluids, this becomes

$$\frac{d^2}{dt^2} g(t) \perp T_{g(t)} \mathrm{Sdiff}(\mathbf{R}^3)$$

or, in terms of the intrinsic velocity field,

$$\partial_t u(t) + (u \cdot \nabla)u \perp \text{ divergence free fields}$$

or equivalently (by Hodge theory)

$$\partial_t u(t) + (u \cdot \nabla)u = \nabla p$$

for some $p$; this is precisely the *Euler equations of incompressible fluids*. This equation can also be deduced from (5.30), after first calculating using (5.29) and the formula for Lie bracket of vector fields that $B(X, Y)$ is the divergence-free component of $X \neg dY$; we omit the details, which are in [**Ar1966**].

Let us now compute the extrinsic momentum $P$, which is conserved by the Euler equations. Given any divergence-free vector field $v$ (in Lagrangian coordinates), we see from (5.34) that $P$ is given by the formula

$$P(v) := \int_{\mathbf{R}^3} u \cdot (g_* v),$$

thus the form $P(v)$ is computed by pushing $v$ over to Eulerian coordinates to get $g_* v := (Dg \circ g^{-1})(v \circ g^{-1})$ and then taking the inner product with $u$. Let us check that this is indeed conserved. Since

$$u_t = -(u \cdot \nabla)u + \nabla p$$

and

$$\partial_t (g_* v) = -\mathcal{L}_u (g_* v),$$

where $\mathcal{L}$ denotes the Lie derivative along the vector field $u$, we see that

$$\partial_t P(v) = \int_{\mathbf{R}^3} (-(u \cdot \nabla)u + \nabla p) \cdot w - u \cdot \mathcal{L}_u w,$$

where $w := g_* v$ is $v$ in Eulerian coordinates. The $\nabla p$ term vanishes by integration by parts, since $v$ (and hence $w$) is divergence-free. The Lie derivative is computed by the formula

$$\mathcal{L}_u w = (u \cdot \nabla)w - (w \cdot \nabla)u.$$

As $u \cdot (w \cdot \nabla)u = \frac{1}{2}(w \cdot \nabla)|u|^2$ is a total derivative (recall here that $w$ is divergence-free), this term vanishes. The other two terms combine to form a total derivative $-(u \cdot \nabla)(u \cdot w)$, which also vanishes, and so the claim follows.

The external momentum is closely related to the *vorticity* $\omega := \operatorname{curl} u$. This is because a divergence-free vector field $v$ can (in principle, at least) be written as the divergence $v = \operatorname{div} \alpha$ of a 2-vector field $\alpha$. As divergence is diffeomorphism invariant, it commutes with pushforward:

$$g_*(\operatorname{div} \alpha) = \operatorname{div}(g_* \alpha)$$

and thus

$$\begin{aligned} P(\operatorname{div} \alpha) &= \int_{\mathbf{R}^3} u \cdot \operatorname{div}(g_* \alpha) \\ &= \int_{\mathbf{R}^3} \omega \cdot g_* \alpha \\ &= \int_{\mathbf{R}^3} (*\omega) \wedge g_* \alpha \end{aligned}$$

where $*$ is the *Hodge star*. We can pull this back to Lagrangian coordinates to obtain

$$P(\operatorname{div} \alpha) = \int_{\mathbf{R}^3} g_*^{-1}(*\omega) \wedge \alpha.$$

As $\alpha$ was an arbitrary 2-form, we thus see that the pullback $g_*^{-1}(*\omega)$ of the Hodge star of the vorticity in Lagrangian coordinates is preserved by the flow, or equivalently that $*\omega$ is transported by the velocity field $u$. In the two-dimensional case, this is well known ($*\omega$ is a scalar in this case); in higher dimensions, this is fact is implicit in the *vorticity equation*

$$\partial_t \omega_{ij} + u_k \partial_k \omega_{ij} + \omega_{ik} \partial_k u_j = 0$$

which can be rewritten as

$$\partial_t(*\omega) + \mathcal{L}_u(*\omega) = 0.$$

In principle, the Euler-Arnold formalism allows one to write the Euler equations for incompressible fluids into a Lax pair form. To properly carry this out by the machinery above, though, would require calculating the

Cartan-Killing form for the infinite-dimensional Lie group Sdiff($\mathbf{R}^3$), which looked quite tricky to me, and I was not able to complete the calculation. However, a Lax pair formulation for this system is known [**FrVi1990**], and it is likely that that formulation is essentially equivalent to the Lax pair that one could construct from the Euler-Arnold formalism. In the simpler two-dimensional case, it was observed in [**Li2001**] that the vorticity equation can also be recast into a slightly different Lax pair form. While this formalism does allow for some of the inverse scattering machinery to be brought to bear on the initial value problem for the Euler equations, it does not as yet seem that this machinery can be successfully used for the global regularity problem.

It would, of course, also be very interesting to see what aspects of this formalism carry over to the Navier-Stokes equation. The first naive guess would be to add a friction term, but this seems to basically correspond to adding a damping factor of $-cu$ (rather than a viscosity factor of $\nu\Delta u$) to the Euler equations and ends up being rather uninteresting (it basically slows down the time variable but otherwise does not affect the dynamics). More generally, it would be of interest to see how the Hamiltonian formalism can be generalised to incorporate dissipation or viscosity.

**5.4.3. Notes.** Thanks to Jerry Marsden for comments.

*Chapter 6*

# Miscellaneous

## 6.1. Multiplicity of perspective

Bill Thurston's *On proof and progress in mathematics* [**Th1994**] has many nice observations about the nature and practice of modern mathematics. One of them is that for any fundamental concept in mathematics, there is usually no "best" way to define or think about that concept, but instead there is often a family of interrelated and overlapping, but distinct, perspectives on that concept, each of which conveying its own useful intuition and generalisations; often, the combination of all of these perspectives is far greater than the sum of the parts. Thurston illustrates this with the concept of *differentiation*, to which he lists seven basic perspectives and one more advanced perspective, and hints at dozens more.

But even the most basic of mathematical concepts admit this multiplicity of interpretation and perspective. Consider for instance the operation of *addition*, that takes two numbers $x$ and $y$ and forms their sum $x+y$. There are many such ways to interpret this operation:

(1) (Disjoint union) $x+y$ is the "size"[1] of the disjoint union $X \uplus Y$ of an object $X$ of size $x$, and an object $Y$ of size $y$.

(2) (Concatenation) $x+y$ is the size of the object formed by concatenating an object $X$ of size $x$ with an object $Y$ of size $y$ (or by appending $Y$ to $X$).

(3) (Iteration) $x+y$ is formed from $x$ by incrementing it $y$ times.

(4) (Superposition) $x+y$ is the "strength" of the superposition of a force (or field, intensity, etc.) of strength $x$ with a force of strength $y$.

(5) (Translation action) $x+y$ is the translation of $x$ by $y$.

(6) (Translation representation) $x+y$ is the amount of translation or displacement incurred by composing a translation by $x$ with a translation by $y$.

(7) (Algebraic) $+$ is a binary operation on numbers that give it the structure of an additive group (or monoid), with 0 being the additive identity and 1 being the generator of the natural numbers or integers.

(8) (Logical) $+$, when combined with the other basic arithmetic operations, are a family of structures on numbers that obey a set of axioms such as the *Peano axioms*.

(9) (Algorithmic) $x+y$ is the output of the long addition algorithm that takes $x$ and $y$ as input.

[1] Size is, of course, another concept with many different interpretations: cardinality, volume, mass, length, measure, etc.

(10) etc.

These perspectives are all closely related to each other; this is why we are willing to give them all the common name of "addition", and the common symbol of $+$. Nevertheless there are some slight differences between each perspective. For instance, addition of cardinals is based on the disjoint union perspective, while addition of ordinals is based on the concatenation perspective. This distinction is more or less invisible at the finite level, but becomes apparent once one considers infinite cardinals or ordinals: for instance, in cardinal arithmetic, $\aleph_0 = 1 + \aleph_0 = \aleph_0 + 1 = \aleph_0 + \aleph_0$, whereas in ordinal arithmetic, $\omega = 1 + \omega < \omega + 1 < \omega + \omega$.

Transitioning from one perspective to another is often a necessary first conceptual step when the time comes to generalise the concept. As a child, addition of natural numbers is usually taught initially by using the disjoint union or iteration perspective, but to generalise to addition of integers, one must first switch to a superposition or translation perspective; similar conceptual shifts are needed when one then turns to addition of rationals, real numbers, complex numbers, residue classes, functions, matrices, elements of abstract additive groups, nonstandard number systems, etc. Eventually, one internalises all of the perspectives (and their inter-relationships) simultaneously, and then becomes comfortable with the addition concept in a very broad set of contexts; but it can be more of a struggle to do so when one has grasped only a subset of the possible ways of thinking about addition.

In many situations, the various perspectives of a concept are either completely equivalent to each other, or close enough to equivalent that one can safely "abuse notation" by identifying them together. But occasionally, one of the equivalences breaks down, and then it becomes useful to maintain a careful distinction between two perspectives that are almost, but not quite, compatible. Consider for instance the following ways of interpreting the operation of exponentiation $x^y$ of two numbers $x$, $y$:

- (Combinatorial) $x^y$ is the number of ways to make $y$ independent choices, each of which chooses from $x$ alternatives.
- (Set theoretic) $x^y$ is the size of the space of functions from a set $Y$ of size $y$ to a set $X$ of size $x$.
- (Geometric) $x^y$ is the volume (or measure) of a $y$-dimensional cube (or hypercube) whose sidelength is $x$.
- (Iteration) $x^y$ is the operation of starting at 1 and then multiplying by $x$ $y$ times.
- (Homomorphism) $y \to x^y$ is the continuous homomorphism from the domain of $y$ (with the additive group structure) to the range of $x^y$ (with the multiplicative structure) that maps 1 to $x$.

- (Algebraic) $\wedge$ is the operation that obeys the laws of exponentiation in algebra.
- (Log-exponential) $x^y$ is[2] $\exp(y \log x)$.
- (Complex-analytic) Complex exponentiation is the analytic continuation of real exponentiation.
- (Computational) $x^y$ is whatever my calculator or computer outputs when it is asked to evaluate $x^y$.
- etc.

Again, these interpretations are usually compatible with each other, but there are some key exceptions. For instance, the quantity $0^0$ would be equal to zero using some of these interpretations, equal to one in others, and undefined in yet others. The quantity $4^{1/2}$ would be equal to 2 in some interpretations, be undefined in others, and be equal to the multivalued expression $\pm 2$ (or to depend on a choice of branch) in yet further interpretations. And quantities such as $i^i$ are sufficiently problematic that it is usually best to try to avoid exponentiation of one arbitrary complex number by another arbitrary complex number unless one knows exactly what one is doing. In such situations, it is best not to think about a single, one-size-fits-all notion of a concept such as exponentiation, but instead be aware of the context one is in (e.g. is one raising a complex number to an integer power? A positive real to a complex power? A complex number to a fractional power? etc.) and to know which interpretations are most natural for that context, as this will help protect against making errors when manipulating expressions involving exponentiation.

It is also quite instructive to build one's own list of interpretations for various basic concepts, analogously to those above (or in [**Th1994**]). Some good examples of concepts to try this on include "multiplication", "integration", "function", "measure", "solution", "space", "size", "distance", "curvature", "number", "convergence", "probability" or "smoothness". For the concept of a "group", see [**Ta2010b**, §1.14] and Section 2.3.

## 6.2. Memorisation vs. derivation

Mathematics is infamous for the large number of basic formulae and results that one has to learn in the subject. But, in contrast to some other subjects with a comparable amount of foundational material to memorise, one can at least *deduce* (or at least formally derive) some of these formulae from others, thus reducing the amount of memory needed to cover the basics.

[2]This raises the question of how to interpret exp and log, and again there are multiple perspectives for each....

Consider for instance the *quotient rule*

$$(f/g)' = \frac{f'g - fg'}{g^2} \tag{6.1}$$

in differential calculus; for this formal discussion we ignore issues about lack of differentiability or division by zero. The quotient rule can be deduced from the simpler (and more fundamental) *product rule*

$$(fg)' = f'g + fg'$$

in a few lines. Indeed, if we set $h$ to be the quotient of $f$ and $g$,

$$h := \frac{f}{g},$$

then we can rewrite this division equation as a product equation,

$$f = gh.$$

Differentiating both sides using the product rule, we get

$$f' = g'h + gh';$$

solving for h', we obtain

$$h' = \frac{f' - g'h}{g}$$

which upon substituting $h = f/g$ gives the quotient rule (6.1).

The above derivation was only formally correct, but one can make it rigorous by invoking the *implicit function theorem* to verify that the quotient $h$ is indeed continuously differentiable whenever $f$, $g$ are, and when $g$ is bounded away from zero; we omit the details.

Now, one may argue that it would have been easier simply to memorise the quotient rule than to memorise the derivation. But the derivation is far more general and is ultimately of greater value when one moves on to more advanced mathematical tasks. For instance, suppose that one reaches a point where one has a time-dependent matrix-valued function $A(t)$, and one wants to compute the derivative $(A(t)^{-1})'$ of the inverse of this matrix. (Again, we assume for now that the matrix is invertible and smoothly dependent on time, to avoid technicalities.) If $A$ was scalar, one could use the quotient rule (6.1) (or the chain rule) to obtain

$$(A(t)^{-1})' = -A'(t)/A(t)^2 \tag{6.2}$$

but this answer is not quite correct in the matrix setting. To get the right answer, one can use the above method of converting a division problem into a multiplicative one. Indeed, writing $B = B(t)$ for the inverse of $A$,

$$B = A^{-1},$$

we convert to a multiplicative equation[3]

$$AB = I$$

and differentiate (using the product rule for matrices, which is identical to its scalar counterpart, with an identical proof - another sign that the product rule is more fundamental than the quotient rule) to obtain

$$A'B + AB' = 0$$

(the identity matrix $I$ of course being constant in time). Carefully solving for $B'$ (keeping in mind that matrix multiplication is not commutative) we obtain

$$B' = -A^{-1}A'B$$

and so on substituting $B(t) = A(t)^{-1}$ we see that the correct version of (6.2) in the matrix case is

$$(A(t)^{-1})' = -A(t)^{-1}A'(t)A(t)^{-1}. \tag{6.3}$$

Note that in the scalar case this collapses back to (6.2) because multiplication is now commutative; but in the non-commutative setting we need to use (6.3) instead.

The above discussion shows that remembering the *method* is a more flexible and general practice than simply memorising the *result*. An even more general practice is to remember the underlying *principle*: expressions involving multiplication are usually easier to manipulate than expressions involving division. Yet more general still is to remember the broader *strategy*: transform and simplify one's expressions *first*, before performing something complicated and delicate.

**6.2.1. Reconstruction via dimensional analysis or coordinate invariance.** If one forgot the rule for differentiating a matrix inverse, and only vaguely remembered that the answer was something like (6.2) or (6.3), then another way to figure out the right answer is to use a kind of "dimensional analysis" (or more precisely, a coordinate-free perspective). We view $A(t)$ not as an $n \times n$ matrix, but instead as an invertible linear transformation from one $n$-dimensional vector space $V$ to another $n$-dimensional vector space W; crucially, we do not require $V$, $W$ to be exactly the same space, instead being merely isomorphic to each other. Then $A'(t)$ is also a transformation from $V$ to $W$, whereas $A(t)^{-1}$ and $(A(t)^{-1})'$ are transformations from $W$ to $V$. One then sees that (6.3) is basically the only plausible generalisation of the scalar equation (6.2) which is *dimensionally consistent* or *coordinate-invariant* in the sense that it does not rely on any artificial

[3]One could also use $BA = I$; this will ultimately lead to the same answer.

identification between $V$ and $W$ (or between $V, W$ and $\mathbf{R}^n$); for instance the candidates

$$(A'(t))^{-1} = -A'(t)A(t)^{-2}$$

or

$$(A'(t))^{-1} = -A(t)^{-2}A'(t)$$

fail this test. This illustrates one of the advantages of a coordinate-independent way of thinking; by purging coordinate-dependent concepts from one's mathematical framework, one eliminates a number of incorrect formulae, sometimes to the extent that the correct formula that one wants is almost the only[4] possible choice that matches various easy special cases and passes some obvious consistency checks.

Of course, in some cases it is more advantageous to be able to perform calculations easily, even at the risk of introducing incorrect formulae. In such cases, an explicit coordinate-dependent viewpoint can be useful. Ideally, one should be able to work comfortably with or without coordinates, and translate between the two whenever one becomes more convenient than the other.

## 6.3. Coordinates

Mathematicians like to describe geometric spaces in terms of numbers. In particular, they often describe a space $X$ (or a portion thereof) via a *coordinate system* $C : X \to \mathbf{R}^n$, which takes any point $p$ in the space $X$ (or a portion thereof) and returns a set of coordinates $C(p) = (x_1(p), \ldots, x_n(p))$ for that point. One can think of this coordinate system as a system of $n$ *coordinate functions* $x_1(), \ldots, x_n()$, each of which maps points in space to a single number, each of which partially describes the location of that point.

For instance, in the Cartesian coordinate system of the plane, every point $p$ has an $x$-coordinate $x(p)$ and an $y$-coordinate $y(p)$, so the coordinate functions are $x()$ and $y()$. If instead we use polar coordinates, every point $p$ now has a radial coordinate $r(p)$ (the distance to the origin) and an angular coordinate $\theta(p)$. On Earth, we have[5] latitude() and longitude() as coordinate functions (and also altitude(), if one is allowed to leave the surface of the earth).

Units of measurement also give rise to coordinate functions. For instance, consider the yard as a unit of length. It gives rise to the yard coordinate function yards(), that takes a line segment in physical space and

[4]This, for instance, is how Einstein was famously led to his equations for gravitation in general relativity.

[5]For the purposes of this discussion I will ignore the issue of coordinate singularities, for instance the issue of defining the angular coordinate at the origin, or longitude at the North or South poles.

returns its length in yards. For instance, if $L$ is a line segment that is 10 yards long, then $\text{yards}(L) = 10$.

Coordinate systems convert points in space to systems of numbers. One can[6] *invert* this system, and create an *inverse coordinate system* that converts a set of numbers $(x_1, \ldots, x_n)$ to a point $C^{-1}(x_1, \ldots, x_n)$ in space. For instance, given a latitude between 90S and 90N, and a longitude between 180E and 180W, one can locate a point on the Earth. Given a positive real number $x$, one can create a line segment that is $x$ yards long, and so forth.

When the coordinate system (and hence, its inverse) is sufficiently linear in nature, one can often describe an inverse coordinate system $C^{-1}$ in terms of a *basis* $v_1, \ldots, v_n$ for the space; the inverse $C^{-1}(x_1, \ldots, x_n)$ of an $n$-tuple of real numbers $x_1, \ldots, x_n$ is then formed by combining $x_1$ copies of $v_1$ with $x_2$ copies of $v_2$, and so forth up to $x_n$ copies of $v_n$. In equations,

$$C^{-1}(x_1, \ldots, x_n) = x_1 v_1 + \ldots + x_n v_n.$$

This basis $v_1, \ldots, v_n$ are then *dual* to the coordinate functions $x_1(), \ldots, x_n()$; indeed, the latter is often referred to as the *dual basis* to $v_1, \ldots, v_n$.

For instance, the dual to the yards() coordinate function is the unit yardstick; to make a line segment $L$ that is 10 yards long (i.e. $\text{yards}(L) = 10$), one simply takes the unit yardstick and dilates it by a factor of 10. The dual to the $x()$ and $y()$ Cartesian coordinates are the standard basis vectors $(1, 0)$ and $(0, 1)$. The latitude() and longitude() functions are nonlinear, but a dual basis can still be prescribed in terms of operations to perform, rather than specific vectors or line segments; the dual to the latitude function is the operation of moving one degree in the north or south direction (depending on one's sign conventions), and similarly the dual to the longitude function is the operation of moving one degree in the east or west direction[7].

One of the quirks of duality is that basis vectors often act in the *opposite* way to the coordinate functions that they are dual to. For instance, the unit yardstick is three times as long as the unit footstick:

$$1 \text{ yard} = 3 \text{ feet}.$$

But, dually, the yards() coordinate is only *one-third* of the feet() coordinate:

---

[6] In some cases, this set of numbers needs to be within an acceptable range before one can invert.

[7] But there is no "one degree north yardstick" or "one degree east basis vector", except in an infinitesimal sense at each location on earth. In mathematics, we can formalise these concepts using the notion of a *tangent space*.

$$\text{yards}(L) = \text{feet}(L)/3.$$

For instance, a line segment $L$ which is 30 feet long, is only 10 yards long. The larger the unit of length, the smaller the coordinate function becomes; it is quite literally an inverse relationship.

Sometimes, this inverse relationship can cause confusion if the distinction between bases and coordinate functions have not been carefully maintained. For instance, in the Cartesian plane, the set of points $(x, y)$ in which $x = 0$ is, confusingly, the *y-axis,* whereas the set of points where $y = 0$ is the *x-axis*. The problem here is that when we say something like "$x = 0$", we are using the $x()$ coordinate, but when we say something like "$x$-axis", we are thinking of an object generated by the dual basis vector $(1, 0)$ to that $x()$ coordinate.

A similar question: is latitude a "north-south" concept or an "east-west" concept? To change the latitude, one moves in a north-south direction, but all the lines of constant latitude, such as the equator, are oriented in an east-west direction.

There is a particularly confusing failure to distinguish between bases and coordinate functions in the terminology of several variable calculus. For instance, consider the partial derivative

$$\frac{\partial f}{\partial x}(x, y, z)$$

of a three-dimensional function in the $x$ direction. It appears that we are somehow differentiating with respect to the $x()$ coordinate, but this is not correct; we are instead differentiating in the direction of the basis vector $(1, 0, 0)$ that is dual to that coordinate.

This may seem like a trivial semantic distinction, but it becomes important when there are multiple coordinate systems in play. Consider for instance the study of an ideal gas $G$ that obeys the *ideal gas law*

$$pV = nRT$$

linking[8] the pressure $p$, the volume $V$, and the temperature $T$. We can view these three quantities $p()$, $V()$, and $T()$ as coordinate functions describing the state of the gas $G$. But because of the ideal gas law, we don't need all three of these quantities to specify the state; just two of them will suffice. For instance, we can use pressure-volume coordinates $(p(), V())$, volume-temperature coordinates $(V(), T())$, or pressure-temperature coordinates $(p(), T())$.

[8]We treat $n$ and $R$ as constants for this discussion.

Now suppose we are measuring some statistic $F(G)$ of the gas $G$ (e.g. its density, its colour, its entropy, etc.), and want to "differentiate $F$ with respect to temperature" to create a rate of change

$$\frac{\partial F}{\partial T}(G).$$

Unfortunately, this expression is not well-defined, because it is sensitive to exactly what coordinate system one is using. If, for instance, one is using volume-temperature coordinates, the above partial derivative describes how $F$ varies with respect to temperature *while holding volume fixed*; this is the basis operation dual to temperature in the volume-temperature coordinate system. If instead one is using pressure-temperature coordinates, the above partial derivative describes how $F$ varies with respect to temperature *while holding pressure fixed* (i.e. allowing the gas to expand or contract as the temperature allows, rather than being constrained to a fixed container); this is the basis operation dual to temperature in the pressure-temperature coordinate system. The two operations can be quite different. For instance, the gas density has a zero derivative with respect to temperature in volume-temperature coordinates, but a negative derivative in pressure-temperature coordinates.

To resolve this issue, chemists often subscript the partial derivative by the other coordinates in the system to emphasise that they are being held fixed. For instance

$$\left(\frac{\partial F}{\partial T}\right)_p (G)$$

would be the rate of change with respect to temperature, holding pressure fixed, whilst

$$\left(\frac{\partial F}{\partial T}\right)_V (G)$$

would be the rate of change with respect to temperature, holding volume fixed. Mathematicians generally avoid this sort of notation, instead using notation such as $X \cdot \nabla F$, $D_v F$, or $dF(X)$ that emphasises the role of basis vectors (or vector fields) instead of coordinate systems.

One point that the ideal gas example illustrates is that the dual basis vector to a coordinate function does not depend only on that coordinate function, but also on the other coordinates in the system; the operation "change the temperature" is not well defined[9] until one specifies what other coordinates are being held fixed.

In the one-dimensional example of the yards() coordinate and the unit yardstick, we saw that a change of basis (e.g. changing yards to feet) affects

[9]The failure to realise this fact is a basic fallacy in economics known as the *ceteris paribus fallacy*.

the coordinate system in an inverse manner to the basis vectors. The same inverse relationship continues to hold in higher dimensional coordinate systems, but is less intuitive because now one must invert matrices or linear transformations instead of numbers in order to quantify the relationship. For instance, to convert the basis operation "increase temperature, while keeping pressure constant" into volume-temperature coordinates, one would have to take a suitable combination of "increase temperature, while keeping volume constant" and "increase volume, while keeping temperature constant"; on the other hand, the other basis operation of pressure-temperature coordinates, namely "increase pressure, while keeping temperature constant" becomes simply "decrease volume, while keeping temperature constant" in volume-temperature coordinates.

The eccentricities of matrix multiplication and matrix inversion can lead to some unintuitive consequences in higher dimensions. For instance, in special relativity, one has the phenomenon of *length contraction*: if one observer $A$ is travelling at a constant velocity with respect to another observer $B$, and $A$ is carrying a rod of length $L$ at rest in $A$'s frame of reference, then that rod will be contracted to be less than $L$ in $B$'s frame. So it would appear that $B$'s unit yardstick is longer than $A$'s unit yardstick. But by symmetry, a rod of length $L$ at rest in $B$'s frame will appear to be shorter than $L$ in $A$'s frame, so that $A$'s unit yardstick appears to be longer than $B$'s unit yardstick. These two facts would contradict each other in a one-dimensional setting, but are compatible with each other in higher dimensions. The reason is due to the difference in the time coordinate functions of $A$ and $B$. The more correct description is that $B$'s unit yardstick is a lengthening of $A$'s unit yardstick *combined with* a time displacement in $A$'s frame of reference; meanwhile, $A$'s unit yardstick is a lengthening of $B$'s unit yardstick *combined with* a time displacement in $B$'s frame of reference. The entangling of time and space given by special relativity ends up causing these two lengthening effects (or contraction effects, when viewed from the opposite perspective) to cancel each other out.

## 6.4. Spatial scales

As a crude heuristic, one can describe the complexity of any given spatial domain $X$ by three parameters[10]:

- The largest (or coarsest) scale $R$ that appears in the domain (this is usually comparable to the *diameter* of $X$, and is infinite if $X$ is unbounded);

[10]If the space X is not isotropic, then the situation is more complicated; for instance, if $X$ is an eccentric rectangle $X = [-A, A]x[-B, B]$, then there are two largest scales $A$, $B$ rather than one. But for this heuristic discussion we shall restrict attention to isotropic settings.

- The smallest (or finest) scale $r$ that appears in the domain (this is only non-zero for discrete (or discretised) spaces $X$); and
- The dimension $d$ of the domain (e.g. Hausdorff or Minkowski dimension).

Thus, for instance, if $X$ is the unit cube $[0,1]^3$ or the 3-torus $(\mathbf{R}/\mathbf{Z})^3$, then $R \sim 1$, $r = 0$, and $d = 3$. Or if $X$ is the lattice $\mathbf{Z}^2$, then $R = \infty$, $r \sim 1$, and $d = 2$. For a cyclic group $\mathbf{Z}/N\mathbf{Z}$, one can either adopt a "discrete" perspective and take $R \sim N$, $r \sim 1$, $d = 1$, or else one can take a "compact" perspective (identifying $\mathbf{Z}/N\mathbf{Z}$ with the $N^{\text{th}}$ roots of unity in $\mathbf{R}/\mathbf{Z}$) and take $R \sim 1$, $r \sim 1/N$, $d = 1$.

As a rule of thumb, a set $X$ is bounded if $R$ is bounded; it is discrete if $r$ is bounded away from zero; and it is (pre-)compact (in a strong topology) if $R$ is bounded *and* $d$ is bounded. (In the weak topology, the hypothesis that $d$ is bounded is not needed.)

The cardinality of $X$ is roughly $(R/r)^d$. Thus, for instance, we have a heuristic explanation as to why spaces which are simultaneously discrete and compact are finite.

If $X$ is a finite abelian group, then the Pontryagin dual of $X$ has the same dimension $d$, but has inverted scales; thus the coarsest scale is now $1/r$ and the finest scale is $1/R$. Thus we see a heuristic explanation as to why discrete groups have compact duals and vice versa (note that the Pontryagin topology is the weak topology in infinite-dimensional settings), and also why the Pontryagin dual has the same cardinality as the original group. We also see an explanation of why the Pontryagin dual of $\mathbf{Z}/N\mathbf{Z}$ with the discrete Haar measure (counting measure) is $\mathbf{Z}/N\mathbf{Z}$ with the compact Haar measure (normalised counting measure), and vice versa.

We thus see that there are three basic ways a space can fail to be finite; by having arbitrarily large scales, arbitrarily small scales, or arbitrarily many dimensions. To approximate an infinite space by a finite one, one has to finitise all three aspects, by some combination of truncation (to finitise the largest scale), discretisation (to finitise the smallest scale), or dimension reduction.

The above heuristic can also be used to classify the different ways in which sequential compactness can fail in a function space in the strong topology (i.e. a sequence $f_n$ of functions fails to have any convergent subsequence):

(1) **Escape to norm infinity**. The norm of elements of the sequence is unbounded.

(2) **Escape to spatial infinity**. The location of elements of the sequence is unbounded.

(3) **Escape to frequency infinity**. The frequency of elements of the sequence is unbounded (or equivalently, the physical scale is not bounded from below).

(4) **Escape to dimensional infinity**. The location of elements in the sequence cannot be captured inside a finite-dimensional space. (This last escape is only relevant for function spaces on infinite-dimensional domains.)

On the other hand, once one closes off all four avenues of escape to infinity, one usually recovers compactness (when the domain is locally compact). Several results and principles in analysis make this more precise: Arzela-Ascoli theorem, Rellich compactness theorem, dominated convergence theorem, concentration compactness, uncertainty principle, etc.

## 6.5. Averaging

Given $n$ quantities $a_1, \ldots, a_n$, how does one define the average value of these quantities? There are, unfortunately, multiple answers to this question. One can take a simple unweighted average

$$\frac{a_1 + \ldots + a_n}{n}.$$

Or, one could assign each quantity $a_i$ a different weight $w_i$, and end up with a weighted average

$$\frac{a_1 w_1 + \ldots + a_n w_n}{w_1 + \ldots + w_n}.$$

And then there are any number of other useful averages (e.g. the median, the mode, the root mean square, the geometric mean, etc.) which are slightly different from the simple or weighted averages, and can be more relevant in some cases.

Many "paradoxes" in statistics, as well as many discrepancies between official statistics and subjective experience, arise because of the distinction between these various averages. For instance, consider the question of what the population density of the United States is. If one does a simple average, dividing the population of the US by the area of the US, one gets a density of about 300 people per square mile, which if the population was spread uniformly, would suggest that each person is about 100 yards from the nearest neighbour.

Of course, this does not conform to actual experience. It is true that if one selects a random square mile patch of land from the US at random, it will contain about 300 people in it on the average. However, not all such patches are equally inhabited by humans. If one wants to know what what density the average *human* in the US sees, rather than the average square

mile patch of land, one has to weight each square mile by its population before taking an average. If one does so, the human-weighted population density now increases to about 1400 people per square mile - a significantly different statistic.

Similarly, the average (unweighted) traffic density on a freeway tends to be much smaller than the density that one personally experiences on that freeway, not because the statistics are fudged or that one is unlucky, but because one is much more likely to be on the freeway during high-traffic times than during low-traffic times, by definition. The more accurate way to model the subjective impression of traffic density is to weight each unit of time by the traffic flow before taking an average.

Another example: the average family size in the US is about 3.1, but the family that you are currently in is likely to be larger than this, since a large family is more likely to contain you than a small family. Again, to reflect subjective experience, one should weight each family by its size before averaging.

Yet another example of how weights distort averages is the *friendship paradox*. This paradox asserts (somewhat depressingly) that your friends are usually likely to be more popular than you are, where "more popular" is defined as "having more friends". The reason for this is that more popular people are, by definition, more likely to show up in your circle of friends than the average person, and so drag up the average. One can see this phenomenon empirically in modern social networks - visit the pages of some of your friends at random, and count their number of friends against yours. An extreme instance of this phenomenon occurs when one considers celebrities on social networks that have enormous numbers of followers; by definition, a large number of people on these networks will be following at least one celebrity, but very few of them will be celebrities themselves.

When combining averages of small sub-populations together to form an average of the combined population, one needs to weight each sub-average by the sub-population size in order to not distort the final average. If the sub-populations being averaged over vary, this can then lead to *Simpson's paradox* : when averaging two quantities $X$ and $Y$, it is possible for the average value of $X$ to exceed that of $Y$ on each sub-population, while being less than that of $Y$ for the whole population, if the size of the sub-populations that $X$ is averaging over is different from those that $Y$ is averaging over. A textbook example of this occured with graduate student application success rates of men and women to UC Berkeley in the 1970s: it turns out that in most departments, women had a slightly higher success rate in their applications than men, but in the university as a whole, women had a lower success rate. The ultimate reason for this was that women tended to apply to

more competitive departments, which lowered their overall average success rate.

In pure mathematics, one can sometimes exploit the difference between unweighted and weighted averages to one's advantage. One example of this is the *Balog-Szemerédi-Gowers lemma* in graph theory (see e.g. [**TaVu2006**]), which has a number of very useful applications to additive combinatorics (it plays a role, for instance, in recent work on expanders, on Szemerédi's theorem, and on the inverse conjecture for the Gowers norm). One can state it informally in terms of friendship graphs, as with the friendship paradox. Consider a population of people, many of whom are friends with each other (mathematically, this is a *dense* graph). Even if there are lots of friends in this population, it is still possible that the graph is not highly connected; the population may be segregated into cliques or other subdivisions, with few connections between these subgroups. However, (one formulation of) the Balog-Szemeredi-Gowers lemma asserts that one can always find a fairly large subgroup in which almost everybody in that subgroup either knows each other, or at least have many friends in common (i.e. are highly connected with one degree of separation).

The idea of the proof is actually very simple: what one does is one picks a popular person at random, and looks at that person's circle of friends. Every pair of people in that circle is already guaranteed to have at least one common friend, namely the original person; but they are now very likely to have a lot of other common friends as well. The reason is that a pair of people who have very few friends in common would have been quite unlikely to have arisen in the circle of friends of the randomly selected popular person in the first case.

## 6.6. What colour is the sun?

Sometimes, children ask good questions that adults would overlook as being too "obvious", without further analysis. Today, my seven-year old son was arguing with a friend over the colour of the sun. The friend took the position that the sun was yellow, whereas my son said it was white, based on direct observation (not a terribly good idea in this case, but never mind that). To my own surprise, I was not able to immediately adjudicate the issue, but had to go do a bit of research first.

To answer the question properly, one has to distinguish between two notions of colour: *physical colour*, which is the power intensity function that describes the intensity of each wavelength of light (or *spectral colour*) in the visible spectrum, and the *perceived colour*, which is the colour that one

actually sees, based on the three[11] different color receptors (the red, green, and blue cones) in the retina. The physical notion of colour is essentially an infinite-dimensional one, with each colour of the visible spectrum being capable of occurring at an independent intensity. But perceived colour is essentially a three-dimensional (RGB) concept, which arises from projecting the physical power intensity function against the absorption spectra of the three types of cones in the retina.

Sunlight has a physical power intensity function that has a peak at the yellow spectral colour but is also large at other spectral colours, most notably at green (which is part of the reason, incidentally, for the rare phenomenon of *green flashes* at sunsets). So in this sense, at least, the sun is yellow[12] (and astronomers indeed classify the sun as a yellow star).

On the other hand, sunlight contains significant intensity at all other colours of the visible spectrum (which is why the rainbows, which are ultimately powered by sunlight, contain all of these colours), leading to all three cones of the retina being saturated and leaving a white colour (though if the light was dimmed enough to avoid saturation, the colour would in fact be slightly pink (!), though this depends to some extent on one's choices of normalisations, which includes a choice of gamma correction and choice of white point).

When the sun is setting, there is more atmospheric scattering due to the low angle of the sunlight. The atmosphere tends to scatter away the blue wavelengths (this is why the sky is blue in the presence of sunlight) and this gives the sun a more yellow or orange colour at sunset. At sunset, the sunlight makes all other objects appear yellowish or orangeish too; whereas at full daylight, there is little colour distortion caused by the sun, which is another indication of its whiteness[13].

Indeed, it seems that one's impression of the colour of the sun is driven more by cultural considerations than by direct observation. In eastern countries, and most notably in Japan, the sun is usually portrayed as being red (most prominently in the Japanese flag), whereas in the West it is usually portrayed as yellow or orange. To further confuse the issue, most astronomical images of the Sun are depicted in false colour (as they usually focus on frequencies other than those favoured by human retinal cones), and the

[11] Partial or complete colour-blindness may occur when one or more of the receptors is not working normally.

[12] Though, just to make things more confusing, if one views the sun from space, without the filtering effect of the atmosphere, the peak intensity is actually at blue-green wavelengths rather than at yellow.

[13] The whiteness of a full moon at night is yet another piece of evidence in this direction, given that the light of a full moon is ultimately coming from reflected sunlight.

images that match one's cultural perception of the Sun tend to be the ones that are retained in popular consciousness.

So the answer to the innocuous question "what colour is the sun?" is in fact remarkably complicated: white, yellow, green, blue, orange, pink, and red are all defensible answers from one viewpoint or another.

## 6.7. Zeno's paradoxes and induction

Zeno of Elea (490BCE?-430BCE?) was arguably the first person to make non-trivial contributions to the field of mathematics now known as real analysis, through his famous paradoxes. His first two paradoxes - the paradox of Achilles and the Tortoise, and the dichotomy paradox - can be viewed as the first rigorous demonstration that the discrete notion of infinity (which we would nowadays call infinite cardinality) is distinct from the continuous notion of infinity (which we would nowadays call unboundedness), in that a set can have the former property without the latter.

One can contrast the discrete and continuous notions of infinity by comparing discrete and continuous induction methods with each other. Observe that if a set of times $T$ is non-empty, and has the property that given every time $t$ in $T$, there exists another $t'$ in $T$ that is larger than $t$, then $T$ is necessarily infinite in the discrete sense (i.e. it has infinite cardinality). However, as the Achilles and the Tortoise paradox demonstrates, $T$ does not need to be infinite in the continuous sense (i.e. it can be bounded). However, suppose we add an additional property to $T$, namely that the limit of any convergent sequence of times in $T$ also lies in $T$ (or equivalently, that $T$ is closed). Then it is not hard to show that $T$ is now infinite in the continuous sense as well (i.e. it is unbounded); this observation forms the basis of the *continuity method* in partial differential equations, which is a continuous analogue of the discrete principle of mathematical induction. So to deal with continuous infinities, one needs the ability to jump to the limit[14].

In this viewpoint, the first two of Zeno's paradoxes make the important point that real analysis cannot be reduced to a branch of discrete mathematics, but requires additional tools in order to deal with the continuum.

The third of Zeno's famous three paradoxes, the paradox of the arrow, also has a very nice interpretation from the perspective of modern analysis. This paradox demonstrates that the future state of a system (in this case, an arrow), cannot be determined solely from its initial position; the initial velocity must also be specified. Viewed from the perspective of the modern theory of differential equations, this observation tells us that the equations of motion must be at least second-order in time, rather than first-order.

[14] *Transfinite induction* operates in a very similar fashion.

As such, one can view Zeno's arrow paradox as a very early precursor of Newton's famous law $F = ma$.

## 6.8. Jevons' paradox

*Jevons' paradox* is the counterintuitive phenomenon that an increase in efficiency when using a resource may in fact lead to *increased* consumption of that resource. For instance the introduction of energy-efficient lighting may increase the net energy cost of lighting, because it increases the incentive to use more lights.

A simple numerical example can illustrate this principle. Suppose one has to decide whether to use one light bulb or two light bulbs to light a room. Ignoring energy costs (and the initial cost of purchasing the bulbs), let's say that lighting a room with one light bulb will provide \$10/month of utility to the room owner, whereas lighting with two light bulbs will provide \$15/month of utility. (Like most goods, the utility from lighting tends to obey a law of diminishing returns.)

Let us first suppose that the energy cost of a light bulb is \$6/month. Then the net utility per month becomes \$4 for one light bulb and \$3 for two light bulbs, so the rational choice would be to use one light bulb, for a net energy cost of \$6/month.

Now suppose that, thanks to advances in energy efficiency, the energy cost of a light bulb drops to \$4/month. Then the net utility becomes \$6/month for one light bulb and \$7/month for two light bulbs; so it is now rational to switch to two light bulbs. But by doing so, the net energy cost jumps up to \$8/month.

So is a gain in energy efficiency good for the environment in this case? It depends on how one measures it. In the first scenario, there was less energy used (the equivalent of \$6/month), but also there was less net utility obtained (\$4/month in this case). In the second scenario, more energy was used (\$8/month). but more net utility was obtained as a consequence (\$7/month). As a consequence of energy efficiency gains, the energy cost per capita increased (from \$6/month to \$8/month); but the energy cost per unit of utility decreased (from $6/4 = 1.5$ to $8/7 \approx 1.14$).

By use of government policies, such as taxation, one can lessen the environmental impact, but at the cost of total utility. For instance, suppose we are in the energy efficient scenario (\$4/month per light bulb), but to encourage conservation, the government imposes an additional \$2/month tax for each light bulb use, thus raising the effective energy price back up to \$6/month. (Such taxes are known as *Pigovian taxes*.) As a consequence, it becomes rational for the room owner to just use one light bulb again. In

this scenario, the energy cost[15] is now down to \$4/month, but the utility to the room owner has dropped to \$4/month (and the utility to the government is \$2/month, for a net utility of \$6/month to the community as a whole). This "taxed energy-efficient" scenario is better than the "free-market energy-efficient" scenario in some metrics (energy costs are lower, both per capita and per unit of private utility, or of community utility) but not in others (the absolute utility to the private citizen, and to the community as a whole, is lower). Note though that this scenario is unconditionally better than the "free market energy-inefficient" scenario (which has effectively the same outcome, but without the \$2/month utility to the government, and with reduced impact on the environment).

Similar (though not quite identical) tradeoffs occur for other government policies, such as quotas or subsidies. So there is no unequivocal answer to the question of whether a government policy is beneficial or not; it depends on what exactly one wishes to optimise.

The free-market energy-efficient scenario maximises private and communal short-term utility, but the taxed energy-efficient scenario maximises private and communal long-term utility due to the higher energy efficiency. To illustrate this, let us now also assume that energy resources are finite. For sake of making the numbers nice and round, let us assume that there are only \$1200 units of energy resources per capita available for lighting; after this resource is depleted, there is no further energy available[16].

In the "free-market energy-inefficient" scenario, each citizen would obtain a utility of \$4/month for a period of \$1200/\$6 = 200 months = 16.67 years, for a total long-term utility of \$800.

In the "free-market energy-efficient" scenario, each citizen would obtain a utility of \$7/month for a period of \$1200/\$8 = 150 months = 12.5 years, for a total long-term utility of \$1050.

In the "taxed energy-efficient" scenario, each citizen obtains a utility of \$4/month and the government obtains a utility of \$2/month for a period of \$1200/\$4 = 300 months = 25 years, for a total long-term utility of \$1200 for the private citizen and an additional \$600 for the government.

[15] Actually, this is not quite the full story, because some portion of the \$2/month of taxation income will ultimately be spent on energy expenditures again, but this is a secondary effect which we will ignore here.

[16] To oversimplify, we assume that energy prices remain constant throughout this process; in practice, energy costs would rise as the pool of available resources shrinks, but this would lead to a much more complicated analysis, so we omit this important aspect of energy prices here.

Thus we see[17] that without taxation, increases in energy efficiency increase utility markedly in both the short term and long term, but significantly reduces the time for which the energy resources last. However, with taxation, increases in energy efficiency significantly increases both long-term utility (both for the private citizens, and for the community) and the time for which the resources last, but at the cost of short-term utility.

The above analysis has implicitly assumed *short-term rationality*: each citizen acts to maximise his or her short-term (month-to-month) utility. Let us now return to the free-market energy-efficient scenario, in which we assume that the citizens, being environmentally conscious, are long-term rationalists rather than short-term rationalists; they seek to optimise their cumulative utility over time[18], rather than their short-term monthly utility. If there was only one citizen accessing the energy resource, then the rational strategy would now be to conserve (i.e. to use one light bulb), as this yields \$6/month utility over 300 months for a net long-term utility of \$1800, rather than \$7/month over 150 months for a net long-term utility of \$1050. This outcome is similar to the taxed energy-efficient scenario, except that all of the utility gained accrues to the private citizen, rather than being partially absorbed by the government.

However, if there are enough private citizens sharing the same resource, then the "tragedy of the commons" effect kicks in. Suppose for instance that there are 100 citizens sharing the same energy resource, which is worth $\$1200 \times 100 = \$120,000$ units of energy. If all the citizens conserve, then the resource lasts for $\$120,000/\$400 = 300$ months and everyone obtains \$1800 long-term utility. But then if one of the citizens "defects" by using two light bulbs, driving up the net monthly energy cost from \$400 to \$404, then the resource now only lasts for $\$120,000/\$404 \sim 297$ months; the defecting citizen now gains approximately $\$7 \times 297 = \$2079$ utility, while the remaining conserving citizens' utility drops from \$1800 to $\$6 \times 297 = \$1782$. Thus we see that it is in each citizen's long-term interest (and not merely short-term interest) to defect; and indeed if one continues this process one can see that one ends up in the situation in which all citizens defect. Thus we see that the tragedy of the commons effectively replaces long-term incentives with short-term ones, and the effects of voluntary conservation are not equivalent to the compulsory effects caused by government policy.

[17] Of course, these effects are dependent to some extent on the choice of numbers selected for this toy model; but the outcomes are fairly typical, as can be seen by experimenting with other numbers or other models.

[18] For simplicity we ignore any discounting due to inflation or the time value of money in this analysis.

## 6.9. Bayesian probability

In classical logic, one can represent one's information about a system as a set of possible states that the system could be in, based on the information at hand. With each new measurement of the system, some possibilities could be eliminated, leading to an updated *posterior* set of information that is an improvement over the *prior* set of information. A good example of this type of updating occurs when solving a Sudoku puzzle; each new cell value that one learns about constrains the possible values of the remaining cells. Other examples can be found in the classic detective stories of Arthur Conan Doyle featuring Sherlock Holmes. *Proof by contradiction* can also be viewed as an instance of this type of deduction.

A modern refinement of classical deduction is that of *Bayesian probability*. Here, one's information about a system is not merely represented as a *set* of possible states, but by a *probability distribution* on the space of all states, indicating one's current beliefs on the likelihood of each particular state actually being the true state. Each new measurement of the system then updates a prior probability distribution to a posterior probability distribution, using *Bayes' formula*

$$\mathbf{P}(A|B) = \frac{\mathbf{P}(B|A)\mathbf{P}(A)}{\mathbf{P}(B)}. \tag{6.4}$$

Bayesian probability is widely used in statistics, in machine learning, and in the sciences.

To relate Bayesian probability to classical deduction, recall that every probability distribution has a *support*, which (in the case when the space of states is discrete) is the set of all states that occur with non-zero probability. When performing a Bayesian update on a discrete space, any state which is inconsistent with the new piece of information will have its posterior probability set to zero, and thus be removed from the support. Thus we see that whilst the probability distribution evolves by Bayesian updating, the support evolves by classical deductive logic. Thus one can view classical logic as the qualitative projection of Bayesian probability, or equivalently, one can view Bayesian probability as a quantitative refinement of classical logic.

Alternatively, one can view Bayesian probability as a special case of classical logic by taking a *frequentist interpretation*. In this interpretation, one views the actual universe (or at least the actual system) as just one of a large number of possible universes (or systems). In each of these universes, the system is in one of the possible states; the probability assigned to each state is then the proportion of the possible universes in which that state is attained. Each new measurement eliminates some fraction of the universes

in a given state, depending on how likely or unlikely that state was to actually produce that measurement; the surviving universes then have a new posterior probability distribution, which is related to the prior distribution by Bayes' formula.

It is instructive to interpret Sherlock Holmes' famous quote, "When you have eliminated all which is impossible, then whatever remains, however improbable, must be the truth," from a Bayesian viewpoint. The statement is technically correct; however, when performing this type of elimination to an (a priori) improbable conclusion, the denominator in Bayes' formula is extremely small, and so the deduction is unstable if it later turns out that some of the possibilities thought to have been completely eliminated, were in fact only incompletely eliminated[19]

**6.9.1. Implication.** We now compare classical and Bayesian logic with regard to the concept of *implication*.

In classical logic, we have the notion of *material implication*: given two statements $A$ and $B$, one can form the statement "$A$ implies $B$", which is the assertion that $B$ is true whenever $A$ is true.

In Bayesian probability, the analogous notion is that of *conditional probability*: given two events $A$ and $B$, one can form the conditional probability $\mathbf{P}(B|A)$, which measures the likelihood that $B$ is true given that $A$ is true.

If $\mathbf{P}(B|A) = 1$, then this is essentially equivalent (outside of an event of probability zero) to the assertion that $A$ implies $B$. At the other extreme, if $\mathbf{P}(B|A) = 0$, this is essentially equivalent to the assertion that $A$ implies not-$B$. If $\mathbf{P}(B|A)$ is instead strictly between 0 and 1, then $A$ implies $B$ some of the time, and not-$B$ at other times.

In classical logic, if one knows that $A$ implies $B$, one cannot then deduce that $B$ implies $A$. However, in Bayesian probability, if one knows that the presence of $A$ elevates the likelihood that $B$ is true, then an observation of $B$ will conversely elevate the prior probability that A is true, thanks to Bayes' formula (6.4):

$$(\mathbf{P}(B|A) > \mathbf{P}(B)) \implies (\mathbf{P}(A|B) > \mathbf{P}(A)).$$

On the other hand, $\mathbf{P}(B|A) = 1$ does *not* imply $\mathbf{P}(A|B) = 1$, which corresponds to the inability to take converses in classical logic.

This may help explain why taking converses is an intuitive operation to those who have not yet been thoroughly exposed to classical logic. It is also instructive to understand why this disparity between the two types of deduction is not in conflict with the previously mentioned links between the

[19]Compare with the mantra "extraordinary claims require extraordinary evidence", which can be viewed as the Bayesian counterpoint to Holmes' classical remark.

two. This disparity is roughly analogous to the disparity between worst-case analysis and average-case analysis; see Section 6.10.

A similar disparity occurs with taking contrapositives. In classical logic, "$A$ implies $B$" is completely equivalent to "not-$B$ implies not-$A$". However, in Bayesian probability, the conditional probabilities $\mathbf{P}(A|B)$ and $\mathbf{P}(\neg B|\neg A)$ can be completely different. A classic example is that of the two statements "All swans are white", and "All non-white things are non-swans". Classically, these two statements are logically equivalent: one is true if and only if the other is true. However, from a probabilistic viewpoint, the two statements are very different. For instance, it is easy to conceive of a situation in which half (say) of all swans are black, whereas the overwhelming majority of non-white things are not swans. Thus, if $x$ is a randomly selected object[20], the probabilities

$$\mathbf{P}(x \text{ is white}|x \text{ is a swan})$$

and

$$\mathbf{P}(x \text{ is a non-swan}|x \text{ is not white})$$

can be completely different.

This can shed light on the *problem of induction* in the philosophy of science. Intuitively, if one wished to test the hypothesis "all swans are white", then every observation of a white swan should help confirm this hypothesis. However, the hypothesis is logically equivalent to "all non-white things are non-swans", and it is intuitively clear that an observation of a non-white non-swan does very little to confirm the hypothesis.

This distinction can be clarified by comparing the hypothesis

$$H = \text{"all swans are white"}$$

with a *null hypothesis*, such as

$$H_0 = \text{"Half of all swans are white, and the other half black"}.$$

If $A$ is the event that a randomly selected swan is white, then $P(A|H) = 1$ and $P(A|H_0) = 1/2$, so by Bayes' formula, an observation of $A$ doubles[21] the probability of $H$ occurring relative to $H_0$:

$$\frac{\mathbf{P}(H|A)}{\mathbf{P}(H_0|A)} = \frac{\mathbf{P}(A|H)}{\mathbf{P}(A|H_0)} \frac{\mathbf{P}(H)}{\mathbf{P}(H_0)} = 2\frac{\mathbf{P}(H)}{\mathbf{P}(H_0)}.$$

On the other hand, if $B$ is the event that a randomly selected non-white object is not a swan, then $\mathbf{P}(B|H) = 1$ and $\mathbf{P}(B|H_0)$ is extremely close to 1 (since the number of non-white non-swans massively outnumber

[20] Let us ignore for now the important issue of how to define "randomly selected object".

[21] Note that this is only under the assumption that the swan really was chosen uniformly at random. If there are biases in the selection procedure, e.g. if the swans were only selected from Europe rather than from both Europe and Australia, then the above analysis is inaccurate.

the number of swans, regardless of how many swans are non-white). So the relative likelihood of $H$ compared with the null hypothesis $H_0$ barely budges with an observation of $B$. (Similarly with just about any other choice of null hypothesis.)

This illustrates a more general point: in order to properly determine whether a piece of evidence $A$ truly supports a given hypothesis $H$, it is not enough to determine how likely $A$ would have occurred under that hypothesis (i.e. to compute $\mathbf{P}(A|H)$), but one also has to determine how likely A would have occurred under rival hypotheses (i.e. to compute $\mathbf{P}(A|H_0)$ for various competing hypotheses $H_0$). It is the *ratio* $\mathbf{P}(A|H)/\mathbf{P}(A|H_0)$ between the two that determines the strength of the evidence: a strong piece of evidence needs to be plausible under hypothesis H, while *simultaneously* being implausible under rival hypotheses.

**6.9.2. Deduction and confirmation.** The most basic deduction in classical reasoning is that of *modus ponens*: if one knows $A$, and one knows that $A$ implies $B$, then one can deduce $B$. The Bayesian analogue of this is the inequality

$$\mathbf{P}(B) \geq \mathbf{P}(B \wedge A) = \mathbf{P}(B|A)\mathbf{P}(A).$$

In particular, if $\mathbf{P}(A) = 1$, and $\mathbf{P}(B|A) = 1$, then $\mathbf{P}(B) = 1$.

More generally, one has the inequality

$$\mathbf{P}(C|A) >= \mathbf{P}(C|B)\mathbf{P}(B|A),$$

which generalises the classical fact that given "$A$ implies $B$" and "$B$ implies $C$", one can deduce "$A$ implies $C$".

In classical logic, one has the *principle of mathematical induction*, which asserts that if $A_1$ is true, and if $A_n$ implies $A_{n+1}$ for all $n = 1, 2, \ldots$, then $A_n$ is true for all $n$. The Bayesian analogue of this is the inequality

$$\mathbf{P}(A_n) \geq \mathbf{P}(A_n|A_{n-1})\mathbf{P}(A_{n-1}|A_{n-2}) \ldots \mathbf{P}(A_2|A_1)\mathbf{P}(A_1).$$

In particular, if all the probability factors on the right-hand side are equal to 1, then the left-hand side is equal to 1 also. But observe that if the probability factors on the right-hand side are strictly less than 1, then this inequality becomes increasingly weak as $n$ goes to infinity. For instance, if we only know that $\mathbf{P}(A_{i+1}|A_i) \geq 0.99$ for all $i$ (informally, we are only "99% confident" in each inductive step), then even if we have complete confidence in the base case $A_1$ (i.e. $P(A_1) = 1$), we can only obtain the bound

$$P(A_n) \geq (0.99)^n,$$

which is a bound that converges exponentially to zero as $n \to infinity$. Thus we see that induction can only be safely applied if one is working in

a "mathematical" mode of thinking (see Section 6.10), in which all implications are known to be true[22] with 100% confidence rather than merely 99% confidence.

We thus see that a chain of inductive reasoning can become increasingly shaky in the Bayesian world. However, one can buttress such a chain by using *independent confirmations*. Suppose for instance one wants to compute some physical quantity $X$. We can take a measurement $X_1$ of $X$, but suppose that this measurement is only 90% reliable, in the sense that $\mathbf{P}(X_1 = a|X = a) \geq 0.9$ for any value $a$ of the actual quantity $X$. Then we only have a 90% confidence that $X$ will equal $X_1$: $P(X = X_1) \geq 0.9$.

But suppose we take two independent measurements $X_1, X_2$ of the same measurement $X$; thus, if $X = a$, then the events $X_1 = a$ and $X_2 = a$ each occur with an independent probability of at least 0.9. Then we see that for any fixed value $a$ of $X$, the probability that $X_1 = X_2 = X$ is at least $0.9 \times 0.9 = 0.81$, while the probability that $X_1 = X_2 \neq X$ is at most $0.1 \times 0.1 = 0.01$. Computing the conditional probabilities, we see that if $X_1$ and $X_2$ agree, then the confidence that this value is equal to X now increases to $\frac{0.81}{0.82} \approx 99\%$:

$$\mathbf{P}(X = X_1 = X_2|X_1 = X_2) \geq \frac{81}{82}.$$

Thus we see that one can use repeated independent trials to boost an unreliable measurement into an increasingly reliable measurement. This basic idea is developed much further in the theory of *confidence intervals* in statistics. Note however that it is crucial that the different trials really are independent; if there is a *systematic* error that affects all the trials in the same way, then one may not get nearly as much of a boost in reliability from increasing the number of trials[23].

Nevertheless, having many independent confirmations of a deductive chain of reasoning

$$A_1 \implies A_2 \implies \ldots \implies A_n$$

can greatly increase the confidence[24] that the final conclusion $A_n$ is indeed correct. For instance, if one wants to be convinced of the validity of a lengthy

[22] Actually, one can allow a small amount of error in each implication of an induction, provided that one constrains the length of the induction to be much less than the reciprocal of that error.

[23] A further caveat: the confidence expressed by these calculations is only valid *before* one actually takes the measurements $X_1$, $X_2$. Once one knows the values of these measurements, the posterior probability distribution of $X$ changes as per Bayes' formula, in a manner that depends on one's prior distribution on $X$. In particular, if $X_1$ and $X_2$ both equal $a$ for some value of $a$ which one believes is very unlikely that $X$ should equal, then one's posterior probability that $X = a$ will be larger than one's prior probability, but would still be small.

[24] The fact that chains of reasoning can degrade the final confidence in the conclusion, whilst independent confirmations can buttress such confidence, is somewhat analogous to the fact that resistances add together when placed in series, but decrease when placed in parallel.

mathematical proof, the existence of independent proofs of key steps of the argument will help build confidence. Even heuristic proofs of such steps, while insufficient to be truly convincing in and of themselves, can be very valuable in confirming a more rigorous proof in the event that one of the steps in that latter proof turns out to contain some minor flaws or gaps.

Interestingly, the method of *proof by contradiction*, which seems so similar to that of taking contrapositives, is much more stable in the Bayesian sense. Classically, this method starts with the hypotheses "$A$ implies $B$" and "not $B$", and deduces "not $A$". The Bayesian analogue of this is the inequality

$$\mathbf{P}(\neg A) \geq 1 - \frac{1 - \mathbf{P}(\neg B)}{\mathbf{P}(B|A)}$$

which is easily verified; in particular, if $\mathbf{P}(\neg B)$ and $\mathbf{P}(B|A)$ are both equal to 1, then $\mathbf{P}(\neg A)$ is also equal to 1. Furthermore, if $\mathbf{P}(\neg B)$ and $bfP(B|A)$ are close to 1, then $\mathbf{P}(\neg A)$ is close to 1. For instance, if the former probabilities are at least 90%, then the latter probability is at least 88%.

Thus we see that different rules of reasoning in classical logic have quite different stability properties once one introduces some Bayesian uncertainty: contrapositives are unstable, proofs by contradiction are stable, and inductions are only stable if the length of the induction is short, or if one buttresses the chain of inductions by independent confirmations. On the other hand, we also saw that taking converses, while illegal in classical logic, has some partial justification in Bayesian probability. So the relationship between classical reasoning and Bayesian reasoning is in fact rather subtle.

**Remark 6.9.1.** Bayesian probability can be generalised further; for instance, quantum mechanics (with the Copenhagen interpretation) can be viewed as a noncommutative generalisation of Bayesian probability, though the connection to classical logic is then lost when one is dealing with observables that do not commute. But this is another story...

## 6.10. Best, worst, and average-case analysis

One can broadly divide the outcomes of any given action into three[25] categories: the worst-case scenario, the average-case scenario, and the best-case scenario. When trying to decide on what action to take, one has to set upper bounds and lower bounds for what scenarios are reasonable to consider. One can roughly categorise different modes of thinking by the upper bounds and lower bounds one chooses to set:

[25] In principle, there should in fact be a spectrum of intermediate scenarios between these three, but thanks to the phenomenon of *concentration of measure*, these three scenarios tend to dominate in any given analysis.

- **Mathematical and pedantic thinkers** tend to set the best-case scenario as the upper bound and the worst-case scenario as the lower bound. (This is the basis for various standard jokes involving stereotypical mathematicians.)
- **Practical and scientific thinkers** tend to set the average-case scenario as both the upper bound and the lower bound. (This helps explain why arguments that are convincing to a physicist might not be as convincing to a mathematician, and vice versa.)
- **Conservative and pessimistic thinkers** tend to set the average-case scenario as the upper bound, and the worst-case scenario as the lower bound. (Here, "conservative" is with respect to risk tolerance, not political orientation, although there is arguably some correlation between the two uses of the term.)
- **Risk-taking and optimistic thinkers** tend to set the average-case scenario as the lower bound, and the best-case scenario as the upper bound.
- **Wishful and idealistic thinkers** tend to set the best-case scenario as both the lower bound and the upper bound.
- **Paranoid and cynical thinkers** tend to set the worst-case scenario as both the lower bound and the upper bound.
- **Absolutist and ideological thinkers** tend to consider *only* the worst-case and best-case scenarios, and ignore the average-case scenario. (As such, this mode of thinking is vulnerable to *false dichotomies* and *slippery slope* arguments.)
- **Reckless and impulsive thinkers** tend consider *none*[26] of the scenarios when making a decision.

Each of these modes of thinking can lead to different decisions from the same set of data. Whether one of these modes is more "correct" than another is however a complicated question; the answer depends to a large extent to one's tolerance for downside risk and one's desire for upside risk. This in turn can depend heavily on the context; for instance, a 10% failure rate might be completely acceptable for, say, cooking a meal, but not for flying a plane. A proper approach to risk management is not locked into any one of the above modes of thinking, but instead evaluates all three scenarios and balances them against one's appetite for downside and upside risk for the situation at hand.

It is also worth noting that the best-case and worst-case scenarios can be very sensitive to one's prior assumptions or models, whilst the average-case

[26] In particular, "reckless" is quite distinct from (calculated) "risk-taking"; cf. Han Solo's quote "Never tell me the odds!".

scenario tends to be much more robust in this regard. This is of particular concern for aggressive risk-management strategies which are very sensitive to the precise likelihood of these extreme scenarios.

These modes of thinking often come up at different stages of solving a mathematical problem. For instance, at the very initial stages of trying to solve a mathematical problem, in which one is brainstorming (either by oneself, or with collaborators) to find a halfway plausible strategy to attack the problem, one often works in a wishful or "best-case scenario" mode of thinking, in which error terms magically cancel themselves out or are otherwise negligible, in which functions have all the regularity, invertibility, etc. needed to justify one's computations, that various obstructions are somehow trivial, and that facts that are true in toy model cases can be automatically extrapolated to cover the general case. This mode of thinking can lead to arguments that are highly unconvincing, but they are still useful for discovering possible strategies to then investigate further.

After a best-case analysis has located a possible strategy, the next step is then often to heuristically validate the strategy using an *average case analysis*, in which error terms and obstructions are not magically assumed to be trivial, but are instead somehow "typical" in size, and rare "pathological" cases are assumed to not be present. For instance, statements that are known to be true "almost everywhere" or "with high probability" might be assumed to actually be true "everywhere" or "surely" for the purposes of such an analysis. It is also common at this stage to assume that results that have been proven in the literature for a slightly different (but still very similar) setting than the current one, are also applicable here. If this average case analysis works out, then the unconvincingly justified strategy would now be upgraded to having the status of a heuristically plausible argument.

Then, to stress-test this argument, one often then takes a "devil's advocate" position and considers a worst-case scenario, in which error terms and other obstructions conspire to cause the maximum damage possible to the argument, and the objects one is manipulating resemble those given as textbook counterexamples to various plausible statements that were going to be used in one's argument. Sometimes these scenarios indicate that the argument is in fact unworkable, or beyond the reach of current methods to make rigorous; in other cases, they may actually lead to a formal counterexample to the problem at hand. In yet other cases, when trying to actually build a consistent worst-case scenario, one discovers a key feature of the problem that blocks all the really bad things from happening simultaneously, which then turns out to be the key to resolving the full problem.

It is usually only after all these scenarios are understood that one takes a properly rigorous and mathematical approach to the problem, considering

all possible scenarios together (possibly dividing into cases as necessary) and making sure that they are all properly dealt with.

The other modes of thinking mentioned previously are also used from time to time in the problem-solving process. For instance, one quick and dirty way to predict (with reasonable accuracy) whether a given statement is true is to check all of the extreme cases (thus adopting "absolutist thinking"). If a statement is true in both the best-case scenario and the worst-case scenario, then it is reasonably likely to be true in all intermediate scenarios as well (particularly if the statement has a fair amount of "convexity" in it). Such heuristics can provide useful shortcuts in process of finding a rigorous solution to a problem (though of course they should not be used in the actual solution itself, unless one can rigorously justify such "convexity" claims, e.g. using interpolation theorems).

When solving a complex problem, it is often beneficial to adopt different modes of thinking to different aspects of a problem, for instance taking a best-case scenario for all but one component of the problem, which one then treats with a worst-case analysis. This "turns off" all but one of the difficulties of the problem, thus focusing attention on questions such as "Assuming everything else works out, can we really get around Obstruction $X$?". After subjecting each of the different aspects in turn to such a focused worst-case analysis, one can often piece together the correct strategy for dealing with all the difficulties simultaneously.

## 6.11. Duality

In linear algebra (and in many other fields of mathematics), one can use *duality* to convert an input into an output, or vice versa. For example, a *covector* is an (linear) operation that takes a vector as input and returns a scalar as output. It arises for instance when differentiating a scalar function $f$ of many variables. One can either view the derivative of $f$ at a point $x$ as an operation which takes a vector $v$ as input and returns a scalar $D_v f(x)$ (the *directional derivative*) as output, or one can simply view the derivative of $f$ at $x$ as a covector[27] $df(x)$.

Dually, one can view a vector as an operation that takes a covector as input, and returns a scalar as output.

Similarly, a linear transformation $T$ from one vector space $V$ to another $W$ can be viewed as a map from a $V$-vector to a $W$-vector, or from a $V$-vector and a $W$-covector to a scalar, or from a $W$-covector to a $V$-covector (this map is of course just the *adjoint* of $T$).

[27] Letting $x$ vary, one obtains a covector field $df$, more commonly known as a 1-*form*.

Or: a bilinear form $B$ on $V$ and $W$ can be viewed as a map from $V$-vectors and $W$-vectors to scalars, or from $V$-vectors to $W$-covectors, or $W$-vectors to $V$-covectors. For instance, the Euclidean metric is a (non-degenerate, symmetric) bilinear form that identifies vectors with covectors, and in particular can identify the derivative $df(x)$ of a function at a point, which is a covector, with a vector $\nabla f(x)$, known as the *gradient* of $f$ at $x$.

These manipulations can seem rather abstract, but they can also be applied to objects and operations in everyday life. For instance, if one views money as the scalars, then goods and markets can be viewed as dual to each other (in the way that vectors and covectors are dual to each other): a market takes a good as input and returns an amount of money as output (the price of that good in that market). Dually, one can view a good as an operation that takes a market as input and returns an amount of money as output; this might be how a comparison shopper would view a good.

An industrial process that converts raw goods into finished goods (the analogue of a linear transformation) can also be viewed as a means of taking a raw good and a finished goods market as input and returning an amount of money as output (the money that the finished goods market will offer for the processed good); this might be how the owner of a factory would view this process. Or, one could view the process as an operation that takes a finished goods market as input and returns a raw goods market as output (this is the adjoint of the process); this might be how a (micro-)economist would view the process.

Here is a somewhat different (and less linear) example. Suppose one has a set of girls and a set of boys, with each girl assigning a rating to each boy indicating that boy's desirability to that girl. (We make no symmetry assumptions here about the converse desirability.) Here, the ratings play the role of the scalars. Desirability is then an analogue of a bilinear form; it takes a girl and a boy as input and returns a rating as output. Alternatively, it can be viewed as an operation that takes a girl as input and returns a ranking of the boys as output, with the boys being ranked by their desirability to that girl. As a third alternative, it can be viewed as an operation that takes a boy as input and returns a ranking of the girls as output (the girls are ranked now by how much they desire that particular boy). Notice how different the desirability operation will look from the girls' point of view than from the boys' point of view, though the two are of course adjoint to each other.

These different perspectives of a single operation are mathematically equivalent to each other (at least in the cases when the operations are linear and the underlying spaces are finite-dimensional, or at least reflexive), but from a conceptual point of view they can be radically different, as they emphasise different features of that operation.

## 6.12. Open and closed conditions

When defining the concept of a mathematical space or structure (e.g. a group, a vector space, a Hilbert space, etc.), one needs to list a certain number of axioms or conditions that one wants the space to satisfy. Broadly speaking, one can divide these conditions into three classes:

(1) **Closed conditions**. These are conditions that generally involve an $=$, $\leq$, or $\geq$ sign or the universal quantifier, and thus codify such things as algebraic structure, non-negativity, non-strict monotonicity, semi-definiteness, etc. As the name suggests, such conditions tend to be closed with respect to limits and morphisms.

(2) **Open conditions**. These are conditions that generally involve a $\neq$, $>$, or $<$ sign or the existential quantifier, and thus codify such things as non-degeneracy, finiteness, injectivity, surjectivity, invertibility, positivity, strict monotonicity, definiteness, genericity, etc. These conditions tend to be stable with respect to perturbations.

(3) **Hybrid conditions**. These are conditions that involve too many quantifiers and relations of both types to be either open or closed. Conditions that codify topological, smooth, or metric structure (e.g. continuity, compactness, completeness, connectedness, regularity) tend to be of this type (this is the notorious "epsilon-delta" business), as are conditions that involve subobjects (e.g. the property of a group being simple, or a representation being irreducible). These conditions tend to have fewer closure and stability properties than the first two (e.g. they may only be closed or stable in sufficiently strong topologies); but there are sometimes some deep and powerful *rigidity theorems* that give more closure and stability here than one might naively expect.

Ideally, one wants to have one's concept of a mathematical structure be both closed under limits, and also stable with respect to perturbations, but it is rare that one can do both at once. Indeed, many *induction arguments* or *continuity arguments* exploit the lack of nontrivial structures that are both closed or stable; see Section 1.7.

In many cases, one often has to have two classes for a single concept: a larger class of "weak" spaces that only have the closed conditions (and so are closed under limits) but could possibly be degenerate or singular in a number of ways, and a smaller class of "strong" spaces inside that have the open and hybrid conditions also. A typical example: the class of Hilbert spaces is contained inside the larger class of pre-Hilbert spaces. Another example: the class of smooth functions is contained inside the larger class of distributions.

As a general rule, algebra tends to favour closed and hybrid conditions, whereas analysis tends to favour open and hybrid conditions. Thus, in the more algebraic part of mathematics, one usually includes degenerate elements in a class (e.g. the empty set is a set; a line is a curve; a square or line segment is a rectangle; the zero morphism is a morphism; etc.), while in the more analytic parts of mathematics, one often excludes them (Hilbert spaces are strictly positive-definite; topologies are usually Hausdorff (or at least $T_0$); traces are usually faithful; etc.).

To put it more succinctly: algebra is the mathematics of the "equals" sign, of identity, and of the "main term"; analysis is the mathematics of the "less than" sign, of magnitude, and of the "error term". Algebra prizes structure, symmetry, and exact formulae; analysis prizes smoothness, stability, and estimates. It is because of these complementary foci that either subfield of mathematics looks strange when viewed through the lens of the other.

# Bibliography


[Ar1966] V. Arnold, *Sur la géométrie différentielle des groupes de Lie de dimension infinie et ses applications à l'hydrodynamique des fluides parfaits*, Ann. Inst. Fourier (Grenoble) **16** 1966 fasc. 1, 319-361.

[ArKh1998] V. Arnold, B. Khesin, Topological methods in hydrodynamics. (English summary) Applied Mathematical Sciences, 125. Springer-Verlag, New York, 1998.

[AuTa2010] T. Austin, T. Tao, *On the testability and repair of hereditary hypergraph properties*, Random Structures and Algorithms **36** (2010), 373–463.

[AvGeTo2010] J. Avigad, P. Gerhardy, H. Towsner, *Local stability of ergodic averages*, Trans. Amer. Math. Soc. **362** (2010), no. 1, 261-288.

[BeCaTa2006] J. Bennett, A. Carbery, T. Tao, *On the multilinear restriction and Kakeya conjectures*, Acta Math. 196 (2006), no. 2, 261-302.

[BeHoMcPa2000] V. Bergelson, B. Host, R. McCutcheon, F. Parreau, *Aspects of uniformity in recurrence*. Colloq. Math. 84/85 (2000), part 2, 549–576.

[BeTaZi2010] V. Bergelson, T. Tao, T. Ziegler, *An inverse theorem for the uniformity seminorms associated with the action of $F_p$*, Geom. Funct. Anal. **19** (2010), no. 6, 1539-1596.

[Bo1991] J. Bourgain, *Besicovitch type maximal operators and applications to Fourier analysis*, Geom. Funct. Anal. 1 (1991), no. 2, 147-187.

[Bo1995] J. Bourgain, *Estimates for cone multipliers*, Geometric aspects of functional analysis (Israel, 19921994), 4160, Oper. Theory Adv. Appl., 77, Birkhuser, Basel, 1995.

[BoGu2010] J. Bourgain, L. Guth, *Bounds on oscillatory integral operators based on multilinear estimates*, `arXiv:1012.3760`

[BoSm1975] J. Bona, R. Smith, *The initial-value problem for the Korteweg-de Vries equation*, Philos. Trans. Roy. Soc. London Ser. A **278** (1975), no. 1287, 555-601.

[CeMaPeRa2003] H. Cendra, J. Marsden, S. Pekarsky, T. Ratiu, *Variational principles for Lie-Poisson and Hamilton-Poincaré equations*, Dedicated to Vladimir Igorevich Arnold on the occasion of his 65th birthday. Mosc. Math. J. **3** (2003), no. 3, 833867, 1197-1198.

[ChSzTr1992] F. Chung, E. Szemerédi, W. T. Trotter, *The number of different distances determined by a set of points in the Euclidean plane*, Discrete Comput. Geom. **7** (1992), no. 1, 1-11.

[CoMi1997] T. Colding, W. Minicozzi, II. *Harmonic functions on manifolds*, Ann. Math. **146** (1997), 725–747.

[Co1982] A. Córdoba, *Geometric Fourier analysis*, Ann. Inst. Fourier (Grenoble) 32 (1982), no. 3, vii, 215-226.

[Co1985] A. Córdoba, *Restriction lemmas, spherical summation, maximal functions, square functions and all that*, Recent progress in Fourier analysis (El Escorial, 1983), 5764, North-Holland Math. Stud., 111, North-Holland, Amsterdam, 1985,

[De1912] M. Dehn, *Transformation der Kurven auf zweiseitigen Flächen*, Math. Ann. 72 (1912), no. 3, 413-421.

[Dv2009] Z. Dvir, *On the size of Kakeya sets in finite fields*, J. Amer. Math. Soc. **22** (2009), no. 4, 1093-1097.

[EbMa1970] D. Ebin, J. Marsden, *Groups of diffeomorphisms and the motion of an incompressible fluid*, Ann. of Math. **92** (1970) 102-163.

[ElSz2007] G. Elek, B. Szegedy, *Limits of Hypergraphs, Removal and Regularity Lemmas. A Non-standard Approach*, `arXiv:0705.2179`

[ElSh2010] G. Elekes, M. Sharir, *Incidences in three dimensions and distinct distances in the plane*, Computational geometry (SCG'10), 413422, ACM, New York, 2010.

[Er1946] P. Erdős, *On sets of distances of n points*, Amer. Math. Monthly **53** (1946), 248–250.

[FrVi1990] S. Friedlander, M. Vishik, *Lax pair formulation for the Euler equation*, Phys. Lett. A **148** (1990), no. 6-7, 313-319.

[Ge1936] G. Gentzen, *Die Widerspruchfreiheit der reinen Zahlentheorie*, Mathematische Annalen **112** (1936), 493-565.

[Go2008] W. T. Gowers, *How can one equivalent statement be stronger than another?*, `gowers.wordpress.com/2008/12/28`

[Go2010] W. T. Gowers, *Decompositions, approximate structure, transference, and the Hahn-Banach theorem*, Bull. Lond. Math. Soc. **42** (2010), no. 4, 573-606.

[Gr2005] B. Green, *Roth's theorem in the primes*, Ann. of Math. (2) **161** (2005), no. 3, 1609-1636.

[GrTa2008] B. Green, T. Tao, *The primes contain arbitrarily long arithmetic progressions*, Annals of Math. **167** (2008), 481–547.

[Gr1981] M. Gromov, *Groups of polynomial growth and expanding maps*, Inst. Hautes Études Sci. Publ. Math. No. **53** (1981), 53–73.

[Gu2010] L. Guth, *The endpoint case of the Bennett-Carbery-Tao multilinear Kakeya conjecture*, Acta Math. **205** (2010), no. 2, 263-286.

[GuKa2010] L. Guth, N. Katz, *Algebraic methods in discrete analogs of the Kakeya problem*, Adv. Math. **225** (2010), no. 5, 2828-2839.

[GuKa2010b] L. Guth, N.H. Katz, *On the Erdős distinct distance problem in the plane*, arXiv:1011.4105v1 [math.CO]

[HeJeKoSt2009] S. Herrmann, A. Jensen, M. Joswig, B. Sturmfels, *How to draw tropical planes*, Electron. J. Combin. 16 (2009), no. 2, Special volume in honor of Anders Bjorner, Research Paper 6, 26 pp.

[KaTa2004] N. Katz, G. Tardos, *A new entropy inequality for the Erdos distance problem*, Towards a theory of geometric graphs, 119126, Contemp. Math., 342, Amer. Math. Soc., Providence, RI, 2004.

[KiVa2008] R. Killip, M. Visan, *Nonlinear Schrödinger Equations at Critical Regularity*, `www.math.ucla.edu/∼visan/ClayLectureNotes.pdf`

[Kl2010] B. Kleiner, *A new proof of Gromov's theorem on groups of polynomial growth*, J. Amer. Math. Soc. **23** (2010), no. 3, 815-829.

[KoSc1997] N. Korevaar, R. Schoen, *Global existence theorems for harmonic maps to non-locally compact spaces*, Comm. Anal. Geom. **5** (1997), no. 2, 333-387.

[LaPi2011] M. Larsen, R. Pink, *Finite Subgroups of Algebraic Groups*, preprint.

[Li2001] Y. Li, *A Lax pair for the two dimensional Euler equation*, J. Math. Phys. **42** (2001), no. 8, 3552-3553.

[Lo1955] J. Łos, Quelques remarques, théorèmes et problèmes sur les classes définissables d'algèbres, in: Mathematical Interpretation of Formal Systems, North-Holland, Amsterdam, 1955, 98–113.

[MaRa1999] J. Marsden, T. Ratiu, Introduction to mechanics and symmetry, A basic exposition of classical mechanical systems. Second edition. Texts in Applied Mathematics, 17. Springer-Verlag, New York, 1999.

[Mi1968] J. Milnor, *Growth of finitely generated solvable groups*, J. Diff. Geom. **2** (1968), 447–449.

[Mo1995] N. Mok, *Harmonic forms with values in locally compact Hilbert bundles*, in Proceedings of the Conference in Honor of Jean-Pierre Kahane (Orsay 1993), Special Issue, 1995, pp. 433–454.

[Na1999] K. Nakanishi, *Scattering theory for the nonlinear Klein-Gordon equation with Sobolev critical power*, Internat. Math. Res. Notices 1999, no. 1, 31-60.

[ReTrTuVa2008] O. Reingold, L. Trevisan, M. Tulsiani, S. Vadhan, *New Proofs of the Green-Tao-Ziegler Dense Model Theorem: An Exposition*, preprint. `arXiv:0806.0381`

[Ro1953] K.F. Roth, *On certain sets of integers*, J. London Math. Soc. **28** (1953), 245-252.

[Ro2009] T. Roy, *Global existence of smooth solutions of a 3D log-log energy-supercritical wave equation*, Anal. PDE **2** (2009), no. 3, 261-280.

[RuSz1978] I. Ruzsa, E. Szemerédi, *Triple systems with no six points carrying three triangles*, Colloq. Math. Soc. J. Bolyai **18** (1978), 939–945.

[Sa1915] G. Salmon, A Treatise on the Analytic Geometry of Three Dimensions, Vol. 2, 5th edition Hodges, Figgis And Co. Ltd. (1915).

[ScTi2002] I. Schindler, K. Tintarev, *An abstract version of the concentration compactness principle*, Rev. Mat. Complut. 15 (2002), no. 2, 417-436.

[ShTa2010] Y. Shalom, T. Tao, *A finitary version of Gromov's polynomial growth theorem*, Geom. Funct. Anal. **20** (2010), no. 6, 1502-1547.

[ShSt1998] J. Shatah, M. Struwe, Geometric wave equations. Courant Lecture Notes in Mathematics, 2. New York University, Courant Institute of Mathematical Sciences, New York; American Mathematical Society, Providence, RI, 1998.

[Se1954] A. Seidenberg, *A new decision method for elementary algebra*, Ann. of Math. (2), **60** (1954), 365-374.

[St1979] E. M. Stein, Some problems in harmonic analysis. Harmonic analysis in Euclidean spaces (Proc. Sympos. Pure Math., Williams Coll., Williamstown, Mass., 1978), Part 1, pp. 320, Proc. Sympos. Pure Math., XXXV, Part, Amer. Math. Soc., Providence, R.I., 1979.

[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 "Problemés Combinatoires et Théorie des Graphes, Proc. Colloque Inter. CNRS," (Bermond, Fournier, Las Vergnas, Sotteau, eds.), CNRS Paris, 1978, 399–401.

[SzTr1873] E. Szemerédi, W. T. Trotter Jr., *Extremal problems in discrete geometry*, Combinatorica **3** (1983), 381–392.

[Ta1951] A. Tarski, *A decision method for elementary algebra and geometry*, University of California Press, Berkeley and Los Angeles, Calif., 1951.

[Ta1999] T. Tao, *The Bochner-Riesz conjecture implies the restriction conjecture*, Duke Math. J. **96** (1999), no. 2, 363-375.

[Ta2003] T. Tao, *A sharp bilinear restrictions estimate for paraboloids*, Geom. Funct. Anal. 13 (2003), no. 6, 1359-1384.

[Ta2003b] T. Tao, *Recent progress on the Restriction conjecture*, `arXiv:math.CA/0311181`

[Ta2004] T. Tao, *A remark on Goldston-Yıldırım correlation estimates*, unpublished. `www.math.ucla.edu/~tao/preprints/Expository/gy-corr.dvi`

[Ta2006] T. Tao, *Spacetime bounds for the energy-critical nonlinear wave equation in three spatial dimensions*, Dyn. Partial Differ. Equ. **3** (2006), no. 2, 93-110.

[Ta2006b] T. Tao, Nonlinear dispersive equations: local and global analysis, CBMS regional conference series in mathematics, 2006.

[Ta2006c] T. Tao, Analysis Vols. I, II, Hindustan Book Agency, 2006.

[Ta2007] T. Tao, *Global regularity for a logarithmically supercritical defocusing nonlinear wave equation for spherically symmetric data*, J. Hyperbolic Differ. Equ. 4 (2007), no. 2, 259-265.

[Ta2007b] T. Tao, *A correspondence principle between (hyper)graph theory and probability theory, and the (hyper)graph removal lemma*, J. d.Analyse Mathematique **103** (2007), 1–45.

[Ta2008] T. Tao, Structure and Randomness: pages from year one of a mathematical blog, American Mathematical Society, Providence RI, 2008.

[Ta2008b] T. Tao, *Norm convergence of multiple ergodic averages for commuting transformations*, Ergodic Theory and Dynamical Systems **28** (2008), 657–688.

[Ta2009] T. Tao, Poincaré's Legacies: pages from year two of a mathematical blog, Vol. I, American Mathematical Society, Providence RI, 2009.

[Ta2009b] T. Tao, Poincaré's Legacies: pages from year two of a mathematical blog, Vol. II, American Mathematical Society, Providence RI, 2009.

[Ta2010] T. Tao, An epsilon of room, Vol. I, American Mathematical Society, Providence RI, 2010.

[Ta2010b] T. Tao, An epsilon of room, Vol. II, American Mathematical Society, Providence RI, 2010.

[Ta2011] T. Tao, An introduction to measure theory, American Mathematical Society, Providence RI, 2011.

[Ta2011b] T. Tao, Higher order Fourier analysis, American Mathematical Society, Providence RI, 2011.

[Ta2011c] T. Tao, Topics in random matrix theory, American Mathematical Society, Providence RI, 2011.

[TaVaVe1998] T. Tao, A. Vargas, L. Vega, *A bilinear approach to the restriction and Kakeya conjectures*, J. Amer. Math. Soc. **11** (1998), no. 4, 967-1000.

[TaVu2006] T. Tao, V. Vu, Additive combinatorics. Cambridge Studies in Advanced Mathematics, 105. Cambridge University Press, Cambridge, 2006.

[TaZi2010] T. Tao, T. Ziegler, *The inverse conjecture for the Gowers norm over finite fields via the correspondence principle*, Anal. PDE **3** (2010), no. 1, 1-20.

[TaZi2011] T. Tao, T. Ziegler, *The inverse conjecture for the Gowers norm over finite fields in low characteristic*, preprint.

[Th1994] W. Thurston, *On proof and progress in mathematics*, Bull. Amer. Math. Soc. (N.S.) **30** (1994), no. 2, 161-177.

[Ti1972] J. Tits, *Free subgroups in linear groups*, J. Algebra **20** (1972), 250–270.

[To1975] P. Tomas, *A restriction theorem for the Fourier transform*, Bull. Amer. Math. Soc. **81** (1975), 477-478.

[Tz2006] N. Tzvetkov, *Ill-posedness issues for nonlinear dispersive equations*, Lectures on nonlinear dispersive equations, 63-103, GAKUTO Internat. Ser. Math. Sci. Appl., 27, Gakkotosho, Tokyo, 2006.

[Va1959] P. Varnavides, *On certain sets of positive density*, J. London Math. Soc. **39** (1959), 358–360.

[Wo1968] J. Wolf, *Growth of finitely generated solvable groups and curvature of Riemannian manifolds*, J. Diff. Geom. **2** (1968), 421–446.

[Wo1995] T. Wolff, *An improved bound for Kakeya type maximal functions*, Rev. Mat. Iberoamericana **11** (1995), no. 3, 651-674.

[Wo2001] T. Wolff, *A sharp bilinear cone restriction estimate*, Ann. of Math. (2) 153 (2001), no. 3, 661-698.

# Index