

## 1.2. Compressed sensing and single-pixel cameras

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

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

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

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

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

The 2-megapixel images described above can take up a lot of disk space on the camera (or on some computer where the images are later uploaded), and also take a non-trivial amount of time (and energy) to transfer from one medium to another. So, it is common practice to get the camera to *compress* the image, from an initial large size (e.g., 2MB) to a much smaller size (e.g., 200KB, which is 10% of the size). The thing is that while the space of *all* images has 2MB worth of “degrees of freedom” or “entropy”, the space of all *interesting* images is much smaller, and can be stored using much less disk space, especially if one is willing to throw away some of the

quality of the image. (Indeed, if one generates an image at random, one will almost certainly not get an interesting image; instead, one will just get random noise looking much like the static one can get on TV screens.)

How can one compress an image? There are many ways, some of which are rather technical, but let me try to give a non-technical (and slightly inaccurate) sketch of how it is done. It is quite typical for an image to have a large featureless component—for instance, in a landscape, up to half of the picture might be taken up by a monochromatic sky background. Suppose for instance that we locate a large square, say  $100 \times 100$  pixels, which are all exactly the same colour—e.g., all white. Without compression, this square would take 10,000 bytes to store (using 8-bit grayscale); however, instead, one can simply record the dimensions and location of the square, and note a single colour with which to paint the entire square; this will require only four or five bytes in all to record, leading to a massive space saving. Now in practice, we do not get such an impressive gain in compression, because even apparently featureless regions have some small colour variation between them. So, given a featureless square, what one can do is record the *average* colour of that square, and then subtract that average off from the image, leaving a small residual error. One can then locate more squares where the average colour is significant, and subtract those off as well. If one does this a couple times, eventually the only stuff left will be very small in magnitude (intensity), and not noticeable to the human eye. So we can throw away the rest of the image and record only the size, location, and intensity of the “significant” squares of the image. We can then reverse this process later and reconstruct a slightly lower-quality replica of the original image, which uses much less space.

Now, the above algorithm is not all that effective in practice, as it does not cope well with sharp transitions from one colour to another. It turns out to be better to work not with average colours in squares, but rather with average colour *imbalances* in squares—the extent to which the intensity on (say) the right half of the square is higher on average than the intensity on the left; one can formalise this by using the (two-dimensional) *Haar wavelet system*.<sup>5</sup> These compression schemes represent the original image as a linear superposition of various “wavelets” (the analogues of the coloured squares in the preceding paragraph), store all the significant (large magnitude) wavelet coefficients, and throw away (or “threshold”) all the rest. This type of “hard wavelet coefficient thresholding” compression algorithm is not nearly as sophisticated as the ones actually used in practice (for instance in the JPEG 2000 standard) but it is somewhat illustrative of the general principles in compression.

---

<sup>5</sup>It turns out that one can work with “smoother” wavelet systems which are less susceptible to artefacts, but this is a technicality which we will not discuss here.

To summarise (and to oversimplify somewhat), the original  $1024 \times 2048$  image may have two million degrees of freedom, and in particular if one wants to express this image in terms of wavelets, then one would thus need two million different wavelets in order to reconstruct all images perfectly. However, the typical *interesting* image is<sup>6</sup> very *sparse* or *compressible* in the wavelet basis: perhaps only a hundred thousand of the wavelets already capture all the notable features of the image, with the remaining 1.9 million wavelets only contributing a very small amount of “random noise” which is largely invisible to most observers.

Now, if we (or the camera) knew in advance *which* hundred thousand of the 2 million wavelet coefficients are going to be the important ones, then the camera could just measure those coefficients and not even bother trying to measure the rest. (It is possible to measure a single coefficient by applying a suitable “filter” or “mask” to the image, and making a single intensity measurement to what comes out.) However, the camera does not know which of the coefficients are going to be the key ones, so it must instead measure all 2 million pixels, convert the image to a wavelet basis, locate the hundred thousand dominant wavelet coefficients to keep, and throw away the rest. (This is of course only a caricature of how the image compression algorithm really works, but we will use it for the sake of discussion.)

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

---

<sup>6</sup>This is not always the case: heavily *textured* images—e.g., images containing hair, fur, etc.—are not particularly compressible in the wavelet basis and pose a challenge for image compression algorithms. But that is another story.

robust with respect to, say, the loss of 10% of the sensors, or with respect to various types of noise or data corruption).

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

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

The solution to this problem is both simple and unintuitive. It is to make 300,000 measurements which are *totally unrelated* to the wavelet basis—despite all that I have said above regarding how this is the best basis in which to view and compress images. In fact, the best types of measurements to make are (pseudo)*random* measurements—generating, say, 300,000 random “mask” images and measuring the extent to which the actual image resembles each of the masks. Now, these measurements (or “correlations”) between the image and the masks are likely to be all very small, and very random. But—and this is the key point—each one of the 2 million possible wavelets which comprise the image will generate their own distinctive “signature” inside these random measurements, as they will correlate positively against some of the masks, negatively against others, and be uncorrelated with yet more masks. However (with overwhelming probability), each of the 2 million signatures will be distinct; furthermore, it turns out that arbitrary linear combinations of up to a hundred thousand of these signatures will still be distinct from each other (from a linear algebra perspective, this is because two randomly chosen 100,000-dimensional subspaces of a 300,000-dimensional ambient space will be almost certainly disjoint from each other). For this reason, it is possible *in principle* to recover the image (or at least the 100,000 most important components of the image) from these 300,000 random measurements.<sup>7</sup>

There are however two technical problems with this approach. Firstly, there is the issue of noise: an image is not perfectly the sum of 100,000

---

<sup>7</sup>To put it another way, we are constructing a linear algebra analogue of a *hash function*—a function that compresses data into a random-looking string, but which can still be used to distinguish different sets of data from each other.

wavelet coefficients, but also has small contributions from the other 1.9 million coefficients. These small contributions could conceivably disguise the contribution of the 100,000 wavelet signatures as coming from a completely unrelated set of 100,000 wavelet signatures; this is a type of “aliasing” problem. The second problem is how to use the 300,000 measurements obtained to recover the image.

Let us focus on the latter problem first. If we knew which 100,000 of the 2 million wavelets were involved, then we could use standard linear algebra<sup>8</sup> methods (Gaussian elimination, least squares, etc.) to recover the signal. However, as stated before, we do not know in advance which wavelets are involved. How can we find out? A naive least-squares approach gives horrible results which involve all 2 million coefficients and thus lead to very noisy and grainy images. One could perform a brute-force search instead, applying linear algebra once for each of the possible set of 100,000 key coefficients, but this turns out to take an insanely impractical amount of time (there are roughly  $10^{170,000}$  combinations to consider!), and in any case this type of brute-force search turns out to be NP-complete in general (it contains problems such as *subset-sum* as a special case). Fortunately, however, there are two much more feasible ways to recover the data:

- *Matching pursuit*: Locate a wavelet whose signature seems to correlate with the data collected; remove all traces of that signature from the data; and repeat until we have totally “explained” the data collected in terms of wavelet signatures.
- *Basis pursuit* (or  $l^1$  minimisation): Out of all the possible combinations of wavelets which would fit the data collected, find the one which is “sparsest” in the sense that the total sum of the magnitudes of all the coefficients is as small as possible. (It turns out that this particular minimisation tends to force most of the coefficients to vanish.) This type of minimisation can be computed in reasonable time via convex optimisation methods such as the *simplex method*.

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

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

---

<sup>8</sup>Indeed, this is one of the great advantages of linear encodings—they are much easier to invert than non-linear ones. Most hash functions are practically impossible to invert—which is an advantage in cryptography, but not in signal recovery.

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

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

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

- *Magnetic resonance imaging (MRI)*. In medicine, MRI attempts to recover an image (in this case, the water density distribution in a human body) by taking a large but finite number of measurements (basically taking a discretised Radon transform (or x-ray transform) of the body), and then reprocessing the data. Because of the large number of measurements needed, the procedure is lengthy for the patient. Compressed sensing techniques can significantly reduce the number of measurements required, leading to faster imaging (possibly even to real-time imaging, i.e., MRI videos rather than static MRI). Furthermore, one can trade off the number of measurements against the quality of the image, so that by using the same number of measurements as one traditionally does, one may be able to get much finer scales of resolution.
- *Astronomy*. Many astronomical phenomena (e.g., pulsars) have various frequency oscillation behaviours, which make them very sparse or compressible in the frequency domain. Compressed sensing techniques then allow one to measure these phenomena in the time domain (i.e., by recording telescope data) and to be able to reconstruct the original signal accurately even from incomplete and noisy data (e.g., if weather, lack of telescope time, or simply the rotation of the Earth prevents a complete time-series of data).
- *Linear coding*. Compressed sensing also gives a simple way for multiple transmitters to combine their output in an error-correcting

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

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

**1.2.2. Notes.** This article was originally posted on April 13, 2007 at [terrytao.wordpress.com/2007/04/13](http://terrytao.wordpress.com/2007/04/13)

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

[www.acm.caltech.edu/l1magic/examples.html](http://www.acm.caltech.edu/l1magic/examples.html)

A comprehensive collection of resources on compressed sensing can be found at [**BaKe2008b**].

### 1.5. Ultrafilters, non-standard analysis, and epsilon management

This article is in some ways an antithesis of Section 1.3. There, the emphasis was on taking a result in soft analysis and converting it into a hard analysis statement (making it more “quantitative” or “effective”); here we shall be focusing on the reverse procedure, in which one harnesses the power of infinitary mathematics—in particular, ultrafilters and non-standard analysis—to facilitate the proof of finitary statements.

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

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

One important step in this direction has been the development of various types of *asymptotic notation*, such as the Hardy notation of using unspecified constants  $C$ , the Landau notation of using  $O()$  and  $o()$ , or the Vinogradov notation of using symbols such as  $\ll$  or  $\lesssim$ ; each of these symbols, when properly used, absorbs one or more of the ambient quantifiers in a hard analysis statement, thus making these statements easier to read. But, as useful as these notations are, they still fall a little short of fully capturing

one's intuition regarding orders of magnitude. For instance, we tend to think of any quantity of the form  $O(1)$  as being “bounded”, and we know that bounded objects can be combined to form more bounded objects; for instance, if  $x = O(1)$  and  $y = O(1)$ , then  $x + y = O(1)$  and  $xy = O(1)$ . But if we attempt to formalise this by trying to create the set  $A := \{x \in \mathbf{R} : x = O(1)\}$  of all bounded numbers, and asserting that this set is then closed under addition and multiplication, we are speaking nonsense; the  $O()$  notation cannot be used within the *axiom schema of specification*, and so the above definition of  $A$  is meaningless.

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

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

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

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

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

These limits obey a number of laws, including

- (1) (Algebra homomorphism) If  $x_n, y_n$  are convergent sequences, and  $c$  is a real number, then  $\lim 1 = 1$ ,  $\lim cx_n = c \lim x_n$ ,  $\lim(x_n + y_n) = \lim x_n + \lim y_n$ , and  $\lim(x_n y_n) = (\lim x_n)(\lim y_n)$ . (In particular, all sequences on the left-hand side are convergent.)
- (2) (Boundedness) If  $x_n$  is a convergent sequence, then  $\inf x_n \leq \lim x_n \leq \sup x_n$ . (In particular, if  $x_n$  is non-negative, then so is  $\lim x_n$ .)
- (3) (Non-principality) If  $x_n$  and  $y_n$  are convergent sequences which differ at only finitely many values of  $n$ , then<sup>27</sup>  $\lim x_n = \lim y_n$ .
- (4) (Shift invariance) If  $x_n$  is a convergent sequence, then for any natural number  $h$  we have  $\lim x_{n+h} = \lim x_n$ .

These properties are of course very useful in computing the limits of various convergent sequences. It is natural to wonder if it is possible to generalise the notion of a limit to cover various non-convergent sequences, such as the class  $l^\infty(\mathbf{N})$  of *bounded* sequences. There are of course many ways to do this in the literature (e.g., if one considers series instead of sequences, one has Cesàro summation, zeta function regularisation, etc.), but (as observed by Euler) one has to give up at least one of the above four limit laws if one wants to evaluate the limit of sequences such as  $0, 1, 0, 1, 0, 1, \dots$

---

<sup>27</sup>Using the voting interpretation of a limit, we thus have the somewhat depressing assertion that no individual voter has any influence on the outcome of the election!

Indeed, if this sequence had a limit  $x$ , then the algebra homomorphism laws force  $x^2 = x$  and thus  $x$  is either 0 or 1; on the other hand, the algebra homomorphism laws also show us that  $1, 0, 1, 0, \dots$  has a limit  $1 - x$ , and hence by shift invariance we have  $x = 1 - x$ , which is inconsistent with the previous discussion. In the voting theory interpretation, the problem here is one of lack of consensus: half of the voters want 0 and the other half want 1, and how can one consistently and fairly elect a choice from this? Similarly, in the probabilistic interpretation, there is no concentration of measure; a randomly chosen  $x_n$  is not close to its expected value of  $1/2$ , but instead fluctuates randomly between 0 and 1.

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

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

Now suppose we consider a *Boolean sequence*  $x_n$ —one which takes only the values 0 and 1. Since  $x_n^2 = x_n$  for all  $n$ , we see from property 1 that  $(p\text{-lim } x_n)^2 = p\text{-lim } x_n$ , thus  $p\text{-lim } x_n$  must also be either 0 or 1. From a voting perspective, the  $p$ -limit is a *voting system*: a mechanism for extracting a yes-no answer out of the yes-no preferences of an infinite number of voters.

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

- (1) (Monotonicity) If  $A$  lies in  $p$ , and  $B$  contains  $A$ , then  $B$  lies in  $p$ .

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

A collection  $p$  obeying properties 1 and 2 is called a *filter*; a collection obeying 1, 2, and 3 is called an *ultrafilter*, and a collection obeying 1, 2, 3, and 4 is a *non-principal ultrafilter*.<sup>28</sup>

A property  $A(n)$  pertaining to a natural number  $n$  can be said to be *p-true* if the set  $\{n : A(n) \text{ true}\}$  lies in  $p$ , and *p-false* otherwise; for instance any tautologically true statement is also *p-true*. Using the probabilistic interpretation, these notions are analogous to those of “almost surely true” and “almost surely false” in probability theory.<sup>29</sup>

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

We have seen how the notion of a *p-limit* creates a non-principal ultrafilter  $p$ . Conversely, once one has a non-principal ultrafilter  $p$ , one can uniquely recover the *p-limit* operation. This is easiest to explain using the voting theory perspective. With the ultrafilter  $p$ , one can ask yes-no questions of an electorate, by getting each voter to answer yes or no and then seeing whether the resulting set of “yes” voters lies in  $p$ . To take a *p-limit* of a bounded sequence  $x_n$ , say in  $[0, 1]$ , what is going on is that each voter  $n$  has his or her own favourite candidate number  $x_n$  between 0 and 1, and one has to elect a real number  $x$  from all these preferences. One can do this by

---

<sup>28</sup>In contrast, a principal ultrafilter is one which is controlled by a single index  $n_0$  in the sense that  $p = \{A : n_0 \in A\}$ . In the voting theory language, this is a scenario in which  $n_0$  is a dictator; in the probability language, the random variable  $n$  is now a deterministic variable taking the values of  $n_0$ .

<sup>29</sup>Indeed, one can view  $p$  as being a probability measure on the natural numbers which always obeys a zero-one law, though one should caution that this measure is only finitely additive rather than countably additive, and so one should take some care in applying measure-theoretic technology directly to an ultrafilter.

an infinite electoral version of “Twenty Questions”: one asks all the voters whether  $x$  should be greater than  $1/2$  or not, and uses  $p$  to determine what the answer should be; then, if  $x$  is to be greater than  $1/2$ , one asks whether  $x$  should be greater than  $3/4$ , and so on and so forth. This eventually determines  $x$  uniquely; the properties 1–4 of the ultrafilter can be used to derive properties 1–3 of the  $p$ -limit.

A modification of the above argument also lets us take  $p$ -limits of any sequence in a compact metric space (or slightly more generally, in any compact Hausdorff first-countable topological space<sup>30</sup>). These  $p$ -limits then behave in the expected manner with respect to operations in those categories, such as composition with continuous functions or with direct sum. As for unbounded real-valued sequences, one can still extract a  $p$ -limit as long as one works in a suitable compactification of the reals, such as the extended real line.

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

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

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

**Proof 2.** Consider the Stone-Čech compactification  $\beta\mathbf{N}$  of the natural numbers. Since  $\mathbf{N}$  is not already compact, there exists an element  $p$  of this compactification which does not lie in  $\mathbf{N}$ . Now note that any bounded sequence  $x_n$  on the natural numbers is a bounded continuous function on  $\mathbf{N}$  (since  $\mathbf{N}$  is discrete) and thus, by definition of  $\beta\mathbf{N}$ , extends uniquely to a bounded

---

<sup>30</sup>Note however that Urysohn’s metrisation theorem implies that any compact Hausdorff first-countable space is metrisable.

continuous function on  $\beta\mathbf{N}$ ; in particular one can evaluate this function at  $p$  to obtain a real number  $x_p$ . If one then defines  $p\text{-lim } x_n := x_p$ , one easily verifies the properties 1–4 of a  $p$ -limit, which by the above discussion creates a non-principal ultrafilter (which by abuse of notation is also referred to as  $p$ ; indeed,  $\beta\mathbf{N}$  is canonically identifiable with the space of all ultrafilters).  $\square$

These proofs are short, but not particularly illuminating. A more informal, but perhaps more instructive, explanation of why non-principal ultrafilters exist can be given as follows. In the voting theory language, our task is to design a complete and consistent voting system for an infinite number of voters. In the cases where there is near-consensus, in the sense that all but finitely many of the voters vote one way or another, the decision is clear—go with the option which is preferred by the infinite voting bloc. But what if an issue splits the electorate with an infinite number of voters on each side? Then what one has to do is make an arbitrary choice—pick one side to go with and completely *disenfranchise* all the voters on the other side, so that they will have no further say in any subsequent votes. By performing this disenfranchisement, we increase the total number of issues for which our electoral system can reach a consistent decision; basically, any issue which has the consensus of all but finitely many of those voters not yet disenfranchised can now be decided upon in a consistent (though highly unfair) manner. We now continue voting until we reach another issue which splits the remaining pool of voters into two infinite groups, at which point we have to make another arbitrary choice, and disenfranchise another infinite set of voters. Very roughly speaking, if one continues this process of making arbitrary choices “ad infinitum”, then at the end of this transfinite process we eventually exhaust the (uncountable) number of issues one has to decide, and one ends up<sup>31</sup> with the non-principal ultrafilter. (If at any stage of the process one decided to disenfranchise all but finitely many of the voters, then one would quickly end up with a principal ultrafilter, i.e., a dictatorship.)

With this informal discussion, it is now rather clear why the axiom of choice (or something very much like that axiom) needs to play a role in constructing non-principal ultrafilters. However, one may wonder whether one really needs the full strength of an ultrafilter in applications; to return once again to the voting analogy, one usually does not need to vote on every

---

<sup>31</sup>One should take this informal argument with a grain of salt; it turns out that after one has made an infinite number of choices, the infinite number of disenfranchised groups, while individually having no further power to influence elections, can begin having some *collective* power, basically because property 2 of a filter only guarantees closure under finite intersections and not infinite intersections, and things begin to get rather complicated. At this point, I recommend abandoning the informal picture and returning to Zorn’s lemma.

single conceivable issue (of which there are uncountably many) in order to settle some problem; in practice, there are often only a countable or even finite number of tricky issues which one needs to put to the ultrafilter to decide upon. Because of this, many of the results in soft analysis which are proven using ultrafilters can instead be established using a “poor man’s non-standard analysis” (or “pre-infinitary analysis”) in which one simply does the “voter disenfranchisement” step mentioned above by hand. This step is more commonly referred to as the trick of “passing to a subsequence whenever necessary”, and is particularly popular in the soft analysis approach to PDE and the calculus of variations. For instance, to minimise some functional, one might begin with a minimising sequence. This sequence might not converge in any reasonable topology, but it often lies in a sequentially compact set in some weak topology (e.g., by using the sequential version of the Banach-Alaoglu theorem<sup>32</sup>), and so by passing to a subsequence one can force the sequence to converge in this topology. One can continue passing to a subsequence whenever necessary to force more and more types of convergence, and can even diagonalise using the Arzelà-Ascoli argument to achieve a countable number of convergences at once (this is of course the sequential Banach-Alaoglu theorem in disguise); in many cases, one gets such a strong convergence that one can then pass to the limit. Most of these types of arguments could also be equivalently performed by selecting an ultrafilter  $p$  at the very beginning (the precise choice of  $p$  is usually not very important), and replacing the notions of limit by  $p$ -limit throughout; roughly speaking, the ultrafilter has performed all the subsequence-selection for you in advance, and all your sequences in compact spaces will automatically converge without the need to pass to any further subsequences. (For much the same reason, ultrafilters can be used to simplify a lot of infinitary Ramsey theory, as all the pigeonholing has been done for you in advance.) On the other hand, the “by hand” approach of selecting subsequences explicitly tends to be much more constructive (for instance, it can often be performed without any appeal to the axiom of choice), and can also be more easily converted to a quantitative “hard analysis” argument (for instance, by using the finite convergence principle from Section 1.3).

As a concrete example from my own experience, in [CoKeStTaTa2008], we had a rather severe epsilon management problem in our “hard analysis” arguments, requiring *seven* (!) very different small quantities  $1 \gg \eta_0 \gg \dots \gg \eta_6 > 0$ , with each  $\eta_i$  extremely small compared with the previous one. (As a consequence of this and of our inductive argument, our eventual

---

<sup>32</sup>In other words, the closed unit ball of a Banach space is *sequentially* compact in the weak\* topology, whenever the dual space is separable.

bounds, while quantitative, were extremely large, requiring a *nine*-fold iterated Knuth arrow<sup>33</sup>!) This epsilon management also led to the paper being unusually lengthy (85 pages). Subsequently (inspired by [KeMe2006]), I learnt how the use of the above “poor man’s non-standard analysis” could conceal almost all of these epsilons (indeed, due to concentration-compactness one can soon pass to a limiting object in which most of the epsilons get sent to zero). Partly because of this, a later paper by myself, Visan, and Zhang [TaViZh2008] on a very similar topic, which adopted this softer approach, was significantly shorter (28 pages, although to be fair this paper also relies on an auxiliary 30-page paper [TaViZh2008b]), though to compensate for this it becomes much more difficult to extract any sort of quantitative bound from the argument.

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

**1.5.2. Non-standard models.** We have now thoroughly discussed non-principal ultrafilters, interpreting them as voting systems which can extract a consistent series of decisions out of a countable number of independent voters. With this we can now discuss non-standard models of a mathematical

---

<sup>33</sup>The Knuth arrow operators are defined recursively by setting  $a \uparrow b := a^b$ ,  $a \uparrow\uparrow b := a \uparrow a \dots a \uparrow a$  (with  $a$  appearing  $b$  times),  $a \uparrow\uparrow\uparrow b := a \uparrow\uparrow a \dots a \uparrow\uparrow a$ , etc.

system. There are a number of ways to build these models, but we shall stick to the most classical (and popular) construction.

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

**Definition 1.33.** Let  $X$  be any set. The *ultrapower*  ${}^*X$  of  $X$  is defined to be the collection of all sequences  $(x_n)$  with entries in  $X$ , modulo the equivalence that two sequences  $(x_n), (y_n)$  are considered equal if they agree  $p$ -almost surely (i.e., the statement  $x_n = y_n$  is  $p$ -true).

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

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

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

through the voting system  $p$  that there is any connection at all between all of the individual voters.

For similar reasons, any property that one can define on a standard object, can also be defined on a non-standard object. For instance, a non-standard integer  $m = (m_n)$  is prime iff the statement “ $m_n$  is prime” is  $p$ -true; a non-standard function  $f = (f_n)$  is continuous iff the statement “ $f_n$  is continuous” is  $p$ -true; and so forth. Basically, if you want to know anything about a non-standard object, go put your question to all the voters, and then feed the answers into the ultrafilter  $p$  to get the answer to your question. The properties 1–4 (actually, just 1–3) of the ultrafilter ensure that you will always get a consistent answer out of this.

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

Now the opposite is also true; any statement in first-order logic, avoiding  $p$ -dependent terms such as standard and non-standard, which is true for non-standard classes of objects, is automatically true for standard classes also. This follows just from applying the above principle to the *negation* of the statement one is interested in. Suppose for instance that one has somehow managed to prove the twin prime conjecture (say) for non-standard natural numbers. To see why this then implies the twin prime conjecture for standard natural numbers, we argue by contradiction. If the statement “the twin prime conjecture failed” was true for standard natural numbers, then it would also be true for non-standard natural numbers (it is instructive to work this out explicitly<sup>34</sup>), a contradiction.

That is the *transfer principle* in a nutshell; informally, everything which avoids  $p$ -dependent terminology and which is true in standard mathematics,

---

<sup>34</sup>In order to avoid some conceptual issues regarding non-standard set theory, I recommend using the following formulation of the twin prime conjecture: For every integer  $N$ , there exists a prime  $p > N$  such that  $p + 2$  is also prime.

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

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

One also sees, for instance, that  $2\omega$  is larger than the sum of  $\omega$  and any limited number, that  $\omega^2$  is larger than the product of  $\omega$  with any limited number, and so forth. It is also clear that the sum or product of any two limited numbers is limited. The number  $1/\omega$  has magnitude smaller than any positive standard real number and is thus considered to be an *infinitesimal*. Using  $p$ -limits, we quickly verify that every limited number  $x$  can be uniquely expressed as the sum of a standard<sup>35</sup> number  $\text{st}(x)$  and an infinitesimal number  $x - \text{st}(x)$ . The set of standard numbers, the set of limited numbers,

---

<sup>35</sup>The map  $x \mapsto \text{st}(\log_{\omega} x)$ , by the way, is a homomorphism from the semiring of non-standard positive reals to the tropical semiring  $(\mathbf{R}, \min, +)$ , and thus encodes the correspondence principle between ordinary rings and tropical rings.

and the set of infinitesimal numbers are all subrings of the set of all non-standard numbers. A non-zero number is infinitesimal if and only if its reciprocal is unlimited.

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

The problem here is rather subtle: a set of non-standard natural numbers is not quite the same thing as a non-standard set of natural numbers. In symbols: if  $2^X := \{A : A \subset X\}$  denotes the power set of  $X$ , then  $2^{*\mathbf{N}} \not\cong {}^*(2^{\mathbf{N}})$ . Let us look more carefully. What is a non-standard set  $A \in {}^*(2^{\mathbf{N}})$  of natural numbers? This is basically a sequence  $(A_n)$  of sets of natural numbers, one for each voter. Any given non-standard natural number  $m = (m_n)$  may belong to  $A$  or not, depending on whether the statement  $m_n \in A_n$  is  $p$ -true or not. We can collect all the non-standard numbers  $m$  which do belong in  $A$ , and call this set  $\tilde{A}$ ; this is thus an element of  $2^{*\mathbf{N}}$ . The map  $A \mapsto \tilde{A}$  from  ${}^*(2^{\mathbf{N}})$  to  $2^{*\mathbf{N}}$  turns out to be injective (why? this is the transferred axiom of extensionality), but it is not surjective; there are some sets of non-standard natural numbers which are not non-standard sets of natural numbers, and as such, the well-ordering principle, when transferred over from standard mathematics, does not apply to them. This subtlety is all rather confusing at first, but a good rule of thumb is that as long as your set (or function, or whatever) is not defined using  $p$ -dependent terminology such as “standard” or “limited”, it will be a non-standard set (or a non-standard function, etc.); otherwise it will merely<sup>36</sup> be a set of non-standard objects (or a function from one non-standard set to another, etc.).

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

---

<sup>36</sup>The situation here is similar to that with the adjective “constructive”; not every function from the constructive numbers to the constructive numbers is itself a constructive function, and so forth.

a ring). It is just that these sets are not themselves non-standard, and thus not subject to the transfer principle.

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

**Lemma 1.34.** *Let  $f : \mathbf{R} \rightarrow \mathbf{R}$  be a standard function, and let  $x, L$  be standard real numbers. We identify  $f$  with a non-standard function in the usual manner. Then the following are equivalent:*

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

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

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

There is one caveat though: Lemma 1.34 only works when  $x$  is *standard*. For instance, consider the standard function  $f(x) := x^2 \sin(1/x^3)$ , with the convention  $f(0) = 0$ . This function is everywhere differentiable, and thus

extending to non-standard numbers we have  $f(x+h) = f(x) + hf'(x) + o(|h|^2)$  for all standard  $x$  and infinitesimal  $h$ . However, the same claim is not true for arbitrary non-standard  $x$ ; consider for instance what happens if one sets  $x = -h$ .

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

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

for any non-standard sequence

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

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

**1.5.4. Big  $O()$  notation.** Big  $O()$  notation<sup>37</sup> in standard mathematics can be translated easily into the non-standard setting, as follows.

**Lemma 1.35.** *Let  $f : \mathbf{N} \rightarrow \mathbf{C}$  and  $g : \mathbf{N} \rightarrow \mathbf{R}^+$  be standard functions (which can be identified with non-standard functions in the usual manner). Then the following are equivalent.*

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

This lemma is proven similarly to Lemma 1.34; the implication of (2) from (1) is obvious from the transfer principle, while the implication of (1) from (2) is again by contradiction, converting a sequence of increasingly bad counterexamples to (1) to a counterexample to (2). Lemma 1.35 is also a special case of the “overspill principle” in non-standard analysis, which asserts that a non-standard set of numbers which contains arbitrarily large standard numbers, must also contain an unlimited non-standard number (thus the large standard numbers “spill over” to contain some non-standard numbers). The proof of the overspill principle is related to the (specious) argument discussed above in which one tried to derive a contradiction from the set of unlimited natural numbers, and is left as an exercise.

---

<sup>37</sup>In some texts, the notation  $f = O(g)$  only requires that  $|f(m)| \leq Cg(m)$  for all sufficiently large  $m$ . The non-standard counterpart to this is the claim that  $|f(m)|/g(m)$  is limited for every unlimited non-standard  $m$ .

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

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

As a specific example of this type of thing from my own experience, in [TaVu2007], we had a large parameter  $n$ , and had at some point to introduce the somewhat fuzzy notion of a “highly rational number”, by which we meant a rational number  $a/b$  whose numerator and denominator were both at most  $n^{o(n)}$  in magnitude. Such numbers looked like they were forming a field, since the sum, difference, product, or quotient of two highly rational numbers was again highly rational (but with a slightly different rate of decay

in the  $o(\cdot)$  notation). Intuitively, one should be able to do any algebraic manipulation on highly rational numbers which is legitimate for true fields (e.g., using Cramer's rule to invert a non-singular matrix) and obtain an output which is also highly rational, as long as the number of algebraic operations one uses is  $O(1)$  rather than, say,  $O(n)$ . We did not actually formalise this rigorously in our standard notation, and instead resorted to informal English sentences to describe this; but one can do everything perfectly rigorously in the non-standard setting by letting  $n$  be non-standard, and defining the field  $F$  of non-standard rationals  $a/b$  where  $a, b = O(n^{o(n)})$ ;  $F$  is genuinely a field of non-standard rationals (but not a non-standard field of rationals), and so using Cramer's rule here (but only for matrices of standard size) would be perfectly legitimate. (We did not actually write our argument in this non-standard manner, keeping everything in the usual standard hard analysis setting, but it would not have been difficult to rewrite the argument non-standardly, and there would be some modest simplifications.)

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

Now it is possible to iterate this process, by introducing a non-standard ultrafilter  ${}^*p$  on the non-standard natural numbers  ${}^*\mathbf{N}$ , and then embedding the field  ${}^*\mathbf{R}$  inside an even larger system  ${}^{**}\mathbf{R}$ , whose elements can be identified (modulo  ${}^*p$ -almost sure equivalence) with non-standard sequences  $(x_n)$  of non-standard numbers in  ${}^*\mathbf{R}$  (where  $n$  now ranges over the non-standard natural numbers  ${}^*\mathbf{N}$ ); one could view these as “doubly non-standard numbers”. This gives us some “even smaller” infinitesimals, such as the “doubly infinitesimal” number  $\eta_1$  given by the non-standard sequence  $1, \eta_0, \eta_0^2, \eta_0^3, \dots$ . This quantity is smaller than any standard or (singly) non-standard number, in particular infinitesimally smaller than any positive quantity depending (via standard or singly non-standard operations) on standard or singly non-standard constants such as 1 or  $\eta_0$ . For

instance, it is smaller than  $1/A(\lfloor 1/\eta_0 \rfloor)$ , where  $A$  is the Ackermann function,<sup>38</sup> since the sequence that defines  $\eta_1$  is indexed over the non-standard natural numbers and  $\eta_0^n$  will drop below  $1/A(\lfloor 1/\eta_0 \rfloor)$  for sufficiently large non-standard  $n$ .

One can continue in this manner, creating a triply infinitesimal quantity  $\eta_2$  which is infinitesimally smaller than anything depending on 1,  $\eta_0$ , or  $\eta_1$ , and so forth. Indeed one can iterate this construction an absurdly large number of times, though in most applications one only needs an explicitly finite number of elements from this hierarchy. Having this hierarchy of infinitesimals, each one of which is guaranteed to be infinitesimally small compared to *any* quantity formed from the preceding ones, is quite useful: it lets one avoid having to explicitly write a lot of epsilon-management phrases such as “Let  $\eta_2$  be a small number (depending on  $\eta_0$  and  $\eta_1$ ) to be chosen later” and “... assuming  $\eta_2$  was chosen sufficiently small depending on  $\eta_0$  and  $\eta_1$ ”, which are very frequent in hard analysis literature, particularly for complex arguments which involve more than one very small or very large quantity. (For instance, the paper [CoKeStTaTa2008] referred to earlier is of this type.)

**1.5.6. Conclusion.** I hope I have shown that non-standard analysis is not a totally “alien” piece of mathematics, and that it is basically only “one ultrafilter away” from standard analysis. Once one selects an ultrafilter, it is actually relatively easy to swap back and forth from the standard universe and the non-standard one (or to doubly non-standard universes, etc.). This allows one to rigorously manipulate things such as “the set of all small numbers”, or to rigorously say things like “ $\eta_1$  is smaller than anything that involves  $\eta_0$ ”, while greatly reducing epsilon management issues by automatically concealing many of the quantifiers in one’s argument. One has to take care as to which objects are standard, non-standard, sets of non-standard objects, etc., especially when transferring results between the standard and non-standard worlds, but as long as one is clearly aware of the underlying mechanism used to construct the non-standard universe and transfer back and forth (i.e., as long as one understands what an ultrafilter is), one can avoid difficulty. The main drawbacks to the use of non-standard notation (apart from the fact that it tends to scare away some of your audience) is that a certain amount of notational setup is required at the beginning, and that the bounds one obtains at the end are rather ineffective (though, of course, one can always, after painful effort, translate a non-standard argument back into a messy but quantitative standard argument if one desires).

---

<sup>38</sup>The Ackermann function can be defined as  $A(n) = n \uparrow \dots \uparrow n$ , where  $\uparrow \dots \uparrow$  is the  $n$ -fold Knuth arrow.

### 1.17. Einstein's derivation of $E = mc^2$

Einstein's equation  $E = mc^2$  describing the equivalence of mass and energy is arguably the most famous equation in physics. But his beautifully elegant *derivation* of this formula [Ei1905] from previously understood laws of physics is considerably less famous. (There is an amusing Far Side cartoon in this regard, with the punchline "squared away", which you can find on-line by searching hard enough.)

In this article I would like to present Einstein's original derivation. Actually, to be precise, in the paper mentioned above, Einstein uses the postulates of special relativity and other known laws of physics to show the following:

**Proposition 1.63** (Mass-energy equivalence). *If a body at rest emits a total energy of  $E$  while remaining at rest, then the mass of that body decreases by  $E/c^2$ .*

Assuming that bodies at rest with zero mass necessarily have zero energy, this implies the famous formula  $E = mc^2$ —but only for bodies which are at rest. For moving bodies, there is a similar formula, but one has to first decide what the correct definition of mass is for moving bodies; I will not discuss this issue here, though it can be found in any textbook on relativity.

Broadly speaking, the derivation of the above proposition proceeds via the following five steps:

- (1) Using the postulates of special relativity, determine how space and time coordinates transform under changes of reference frame (i.e., derive the Lorentz transformations).
- (2) Using 1, determine how the temporal frequency  $\nu$  (and wave number  $k$ ) of photons transform under changes of reference frame (i.e., derive the formulae for relativistic Doppler shift).
- (3) Using Planck's law  $E = h\nu$  (and de Broglie's law  $p = \hbar k$ ) and 2, determine how the energy  $E$  (and momentum  $p$ ) of photons transform under changes of reference frame.
- (4) Using the law of conservation of energy (and momentum) and 3, determine how the energy (and momentum) of bodies transform under changes of reference frame.
- (5) Comparing the results of 4 with the classical Newtonian approximations  $KE \approx \frac{1}{2}m|v|^2$  (and  $p \approx mv$ ), deduce the relativistic relationship between mass and energy for bodies at rest (and more generally, between mass, velocity, energy, and momentum for moving bodies).

Actually, as it turns out, Einstein's analysis for bodies at rest only needs to understand changes of reference frame at *infinitesimally low* velocity,  $|v| \ll c$ . However, in order to see enough relativistic effects to deduce the mass-energy equivalence, one needs to obtain formulae which are accurate to second order in  $v$  (or more precisely,  $v/c$ ), as opposed to those in Newtonian physics which are accurate to first order in  $v$  (or  $v/c$ ). Also, to understand the relationship between mass, velocity, energy, and momentum for moving bodies rather than bodies at rest, one needs to consider non-infinitesimal changes of reference frame.

**Remark 1.64.** Einstein's argument is, of course, a physical argument rather than a mathematical one. While I will use the language and formalism of pure mathematics here, it should be emphasised that I am not exactly giving a formal proof of the above proposition in the sense of modern mathematics; these arguments are instead more like the classical proofs of Euclid, in that numerous "self-evident" assumptions about space, time, velocity, etc. will be made along the way. (Indeed, there is a very strong analogy between Euclidean geometry and the Minkowskian geometry of special relativity.) One can of course make these assumptions more explicit, and this has been done in many other places, but I will avoid doing so here in order not to overly obscure Einstein's original argument.

**1.17.1. Lorentz transforms to first order.** To simplify the notation, we shall assume that the ambient spacetime  $S$  has only one spatial dimension rather than three, although the analysis here works perfectly well in three spatial dimensions (as was done in Einstein's original paper). Thus, in any inertial reference frame  $F$ , the spacetime  $S$  is parameterised by two real numbers  $(t, x)$ . Mathematically, we can describe each frame  $F$  as a bijection between  $S$  and  $\mathbf{R} \times \mathbf{R}$ . To normalise these coordinates, suppose that all reference frames agree to use a single event  $O$  in  $S$  as their origin  $(0, 0)$ ; thus

$$(1.49) \quad F(O) = (0, 0)$$

for all frames  $F$ .

Given an inertial reference frame  $F : S \rightarrow \mathbf{R} \times \mathbf{R}$ , one can generate new inertial reference frames in two different ways. One is by reflection: one takes the same frame, with the same time coordinate, but reverses the space coordinates to obtain a new frame  $\overline{F} : S \rightarrow \mathbf{R} \times \mathbf{R}$ , thus reversing the orientation of the frame. In equations, we have

$$(1.50) \quad F(E) = (t, x) \implies \overline{F}(E) = (t, -x)$$

for any spacetime event  $E$ . Another way is by replacing the observer which is stationary in  $F$  with an observer which is moving at a constant velocity  $v$  in  $F$ , to create a new inertial reference frame  $F_v : S \rightarrow \mathbf{R} \times \mathbf{R}$  with the

same orientation as  $F$ . In our analysis, we will only need to understand infinitesimally small velocities  $v$ ; there will be no need to consider observers traveling at speeds close to the speed of light.

The new frame  $F_v : S \rightarrow \mathbf{R} \times \mathbf{R}$  and the original frame  $F : S \rightarrow \mathbf{R} \times \mathbf{R}$  must be related by some transformation law

$$(1.51) \quad F_v = L_v \circ F$$

for some bijection  $L_v : \mathbf{R} \times \mathbf{R} \rightarrow \mathbf{R} \times \mathbf{R}$ . A priori, this bijection  $L_v$  could depend on the original frame  $F$  as well as on the velocity  $v$ , but the principle of relativity implies that  $L_v$  is in fact the same in all reference frames  $F$ , and so only depends on  $v$ .

It is thus of interest to determine what the bijections  $L_v : \mathbf{R} \times \mathbf{R} \rightarrow \mathbf{R} \times \mathbf{R}$  are. From our normalisation (1.49) we have

$$(1.52) \quad L_v(0, 0) = (0, 0),$$

but this is of course not enough information to fully specify  $L_v$ . To proceed further, we recall *Newton's first law*, which states that an object with no external forces applied to it moves at constant velocity, and thus traverses a straight line in spacetime as measured in any inertial reference frame.<sup>84</sup> This implies that  $L_v$  transforms straight lines to straight lines.<sup>85</sup> Combining this with (1.52), we conclude that  $L_v$  is a linear transformation.<sup>86</sup> Thus we can view  $L_v$  now as a  $2 \times 2$  matrix.

When  $v = 0$ , it is clear that  $L_v$  should be the identity matrix  $I$ . Making the plausible assumption that  $L_v$  varies smoothly with  $v$ , we thus have the Taylor expansion

$$(1.53) \quad L_v = I + L'_0 v + O(v^2)$$

for some matrix  $L'_0$  and for infinitesimally small velocities  $v$ .<sup>87</sup> Expanding everything out in coordinates, we obtain

$$(1.54) \quad L_v(t, x) = ((1 + \alpha v + O(v^2))t + (\beta v + O(v^2))x, (\gamma v + O(v^2))t + (1 + \delta v + O(v^2))x)$$

for some absolute constants  $\alpha, \beta, \gamma, \delta \in \mathbf{R}$  (not depending on  $t, x$ , or  $v$ ).

The next step, of course, is to pin down what these four constants are. We can use the reflection symmetry (1.50) to eliminate two of these constants. Indeed, if an observer is moving at velocity  $v$  in frame  $F$ , it is moving

<sup>84</sup>We are assuming here that the property of “having no external forces applied to it” is not affected by changes of inertial reference frame. For non-inertial reference frames, the situation is more complicated due to the appearance of fictitious forces.

<sup>85</sup>To be pedantic, we have only shown this for straight lines corresponding to velocities that are physically attainable, but let us ignore this minor technicality here.

<sup>86</sup>It is a cute exercise to verify this claim formally, under reasonable assumptions such as smoothness of  $L_v$ .

<sup>87</sup>Mathematically, what we are doing here is analysing the Lie group of transformations  $L_v$  via its Lie algebra.

at velocity  $-v$  in frame  $\overline{F}$ , and hence  $\overline{F}_v = \overline{F}_{-v}$ . Combining this with (1.50), (1.51), (1.54) one eventually obtains

$$(1.55) \quad \alpha = 0 \text{ and } \delta = 0.$$

Next, if a particle moves at velocity  $v$  in frame  $F$ , and more specifically moves along the worldline  $\{(t, vt) : t \in \mathbf{R}\}$ , then it will be at rest in frame  $F_v$ , and (since it passes through the universally agreed upon origin  $O$ ) must then lie on the worldline  $\{(t', 0) : t' \in \mathbf{R}\}$ . From (1.51), we conclude that

$$(1.56) \quad L_v(t, vt) \in \{(t', 0) : t' \in \mathbf{R}\} \text{ for all } t.$$

Inserting this into (1.54) (and using (1.55)) we conclude that  $\gamma = -1$ . We have thus pinned down  $L_v$  to first order almost completely:

$$(1.57) \quad L_v(t, x) = (t + \beta vx, x - vt) + O(v^2(|t| + |x|)).$$

Thus, rather remarkably, using nothing more than the principle of relativity and Newton's first law, we have almost entirely determined<sup>88</sup> the reference frame transformation laws, save for the question of determining the real number  $\beta$ . If this number vanished, we would eventually recover classical Galilean relativity. If this number was positive, we would eventually end up with the (rather unphysical) situation of Euclidean relativity, in which spacetime had a geometry isomorphic to that of the Euclidean plane. As it turns out, though, in special relativity this number is negative. This follows from the second postulate of special relativity, which asserts that the speed of light  $c$  is the same in all inertial reference frames. In equations (and because  $F_v$  has the same orientation as  $F$ ), this is asserting that

$$(1.58) \quad L_v(t, ct) \in \{(t', ct') : t' \in \mathbf{R}\} \text{ for all } t$$

and

$$(1.59) \quad L_v(t, -ct) \in \{(t', -ct') : t' \in \mathbf{R}\} \text{ for all } t.$$

Inserting either of (1.58), (1.59) into (1.57) we conclude that  $\beta = -1/c^2$ , and thus we have obtained a full description of  $L_v$  to first order:

$$(1.60) \quad L_v(t, x) = \left(t - \frac{vx}{c^2}, x - vt\right) + O(v^2(|t| + |x|)).$$

**1.17.2. Lorentz transforms to second order.** It turns out that to get the mass-energy equivalence, first-order expansion of the Lorentz transformations  $L_v$  is not sufficient; we need to expand to second order. From Taylor expansion we know that

$$(1.61) \quad L_v = I + L'_0 v + \frac{1}{2} L''_0 v^2 + O(v^3)$$

---

<sup>88</sup>In mathematical terms, what we have done is classified the one-dimensional Lie subalgebras of  $\mathfrak{g}_2(\mathbf{R})$  which are invariant under spatial reflection, and coordinatised using (1.56).

for some matrix  $L_0''$ . To compute this matrix, let us make the plausible assumption that if the frame  $F_v$  is moving at velocity  $v$  with respect to  $F$ , then  $F$  is moving at velocity  $-v$  with respect to  $F_v$ .<sup>89</sup> Applying (1.51) we conclude that  $L_{-v} \circ L_v = I$ . Inserting this into (1.61) and comparing coefficients we conclude that  $L_0'' = (L_0')^2$ . Since  $L_0'$  is determined from (1.60), we can compute everything explicitly, eventually ending up at the second order expansion

$$(1.62) \quad L_v(t, x) = \left( t - \frac{vx}{c^2} + \frac{tv^2}{2c^2}, x - vt + \frac{xv^2}{2c^2} \right) + O(v^3(|t| + |x|)).$$

One can continue in this fashion (exploiting the fact that the  $L_v$  must form a Lie group (with the Lie algebra already determined), and using (1.56) to fix the parameterisation  $v \mapsto L_v$  of that group) to eventually get the full expansion of  $L_v$ , namely

$$L_v(t, x) = \left( \frac{t - vx/c^2}{\sqrt{1 - v^2/c^2}}, \frac{x - vt}{\sqrt{1 - v^2/c^2}} \right),$$

but we will not need to do so here.

**1.17.3. Doppler shift.** The formula (1.62) is already enough to recover the relativistic Doppler shift formula (to second order in  $v$ ) for radiation moving at speed  $c$  with some wave number  $k$ . Mathematically, such radiation moving to the right in an inertial reference frame  $F$  can be modeled by the function

$$A \cos(k(x - ct) + \theta)$$

for some amplitude  $A$  and phase shift  $\theta$ . If we move to the coordinates  $(t', x') = L_v(t, x)$  provided by an inertial reference frame  $F'$ , a computation then shows that the function becomes<sup>90</sup>

$$A \cos(k_+(x' - ct') + \theta)$$

where  $k_+ = (1 - v/c + v^2/2c^2 + O(v^3))k$ . Similarly, radiation moving at speed  $c$  to the left will transform from

$$A \cos(k(x + ct) + \theta)$$

to

$$A \cos(k_-(x + ct) + \theta),$$

where  $k_- = (1 + v/c + v^2/2c^2 + O(v^3))k$ . This describes how the wave number  $k$  transforms under changes of reference frame by small velocities  $v$ .

---

<sup>89</sup>One can justify this by considering two frames receding at equal and opposite directions from a single reference frame, and using reflection symmetry to see how these two frames move with respect to each other.

<sup>90</sup>Actually, if the radiation is tensor-valued, the amplitude  $A$  might also transform in some manner, but this transformation will not be of relevance to us.

The temporal frequency  $\nu$  is linearly related to the wave number  $k$  by the formula

$$(1.63) \quad \nu = \frac{c}{2\pi}k,$$

and so this frequency transforms by the (red-shift) formula

$$(1.64) \quad \nu_+ = (1 - v/c + v^2/2c^2 + O(v^3))\nu$$

for rightward moving radiation and by the (blue-shift) formula

$$(1.65) \quad \nu_- = (1 + v/c + v^2/2c^2 + O(v^3))\nu$$

for leftward moving radiation. (As before, one can give an exact formula here, but the above asymptotic will suffice for us.)

**1.17.4. Energy and momentum of photons.** From the work of Planck, and of Einstein himself on the photoelectric effect, it was known that light could be viewed both as a form of radiation (moving at speed  $c$ ), and also made up of particles (photons). From Planck's law, each photon has an energy of  $E = h\nu$  and (from de Broglie's law) a momentum of  $p = \pm\hbar k = \pm\frac{h}{2\pi}k$ , where  $h$  is Planck's constant, and the sign depends on whether one is moving rightward or leftward. In particular, from (1.63) we have the pleasant relationship<sup>91</sup>

$$(1.66) \quad E = |p|c$$

for photons. Applying (1.64), (1.65), we see that if we view a photon in a new reference frame  $F_v$ , then the observed energy  $E$  and momentum  $p$  now become

$$(1.67) \quad E_+ = (1 - v/c + v^2/2c^2 + O(v^3))E; \quad p_+ = (1 - v/c + v^2/2c^2 + O(v^3))p$$

for rightward moving photons, and

$$(1.68) \quad E_- = (1 + v/c + v^2/2c^2 + O(v^3))E; \quad p_- = (1 + v/c + v^2/2c^2 + O(v^3))p$$

for leftward moving photons.

These two formulae (1.67), (1.68) can be unified using (1.66) into a single formula

$$(1.69) \quad (E'/c^2, p') = L_v(E/c^2, p) + O(v^3)$$

for any photon (moving either leftward or rightward) with energy  $E$  and momentum  $p$  as measured in frame  $F$ , and energy  $E'$  and momentum  $p'$  as measured in frame  $F_v$ .

---

<sup>91</sup>More generally, it turns out that for arbitrary bodies, momentum, velocity, and energy are related by the formula  $p = \frac{1}{c^2}Ev$ , though we will not derive this fact here.

**Remark 1.65.** Actually, the error term  $O(v^3)$  can be deleted entirely by working a little harder. From the linearity of  $L_v$  and the conservation of energy and momentum, it is then natural to conclude that (1.69) should also be valid not only for photons, but for any object that can exchange energy and momentum with photons. This can be used to derive the formula  $E = mc^2$  fairly quickly, but let us instead give the original argument of Einstein, which is only slightly different.

**1.17.5. Einstein's argument.** We are now ready to give Einstein's argument. Consider a body at rest in a reference frame  $F$  with some mass  $m$  and some rest energy  $E$ . (We do not yet know that  $E$  is equal to  $mc^2$ .) Now let us view this same mass in some new reference frame  $F_v$ , where  $v$  is a small velocity. From Newtonian mechanics, we know that a body of mass  $m$  moving at velocity  $v$  acquires a kinetic energy of  $\frac{1}{2}mv^2$ . Thus, assuming that Newtonian physics is valid at low velocities to top order, the net energy  $E'$  of this body as viewed in this frame  $F_v$  should be

$$(1.70) \quad E' = E + \frac{1}{2}mv^2 + O(v^3).$$

**Remark 1.66.** If we assume that the transformation law (1.69) applies for this body, one can already deduce the formula  $E = mc^2$  for this body at rest from (1.70) (and the assumption that bodies at rest have zero momentum), but let us instead give Einstein's original argument.

We return to frame  $F$ , and assume that our body emits two photons of equal energy  $\Delta E/2$ , one moving leftward and one moving rightward. By (1.66) and conservation of momentum, we see that the body remains at rest after this emission. By conservation of energy, the remaining energy in the body is  $E - \Delta E$ . Let us say that the new mass in the body is  $m - \Delta m$ . Our task is to show that  $\Delta E = \Delta mc^2$ .

To do this, we return to frame  $F_v$ . By (1.67), the rightward moving photon has energy

$$(1.71) \quad (1 - v/c + v^2/2c^2 + O(v^3)) \frac{\Delta E}{2}$$

in this frame; similarly, the leftward moving photon has energy

$$(1.72) \quad (1 + v/c + v^2/2c^2 + O(v^3)) \frac{\Delta E}{2}.$$

What about the body? By repeating the derivation of (1.69), it must have energy

$$(1.73) \quad (E - \Delta E) + \frac{1}{2}(m - \Delta m)v^2 + O(v^3).$$

By the principle of relativity, the law of conservation of energy has to hold

in the frame  $F_v$  as well as in the frame  $F$ . Thus, the energy (1.71) + (1.72) + (1.73) in frame  $F_v$  after the emission must equal the energy  $E' = (1.70)$  in frame  $F_v$  before emission. Adding everything together and comparing coefficients we obtain the desired relationship  $\Delta E = \Delta mc^2$ .

**Remark 1.67.** One might quibble that Einstein's argument only applies to emissions of energy that consist of equal and opposite pairs of photons. But one can easily generalise the argument to handle arbitrary photon emissions, especially if one takes advantage of (1.69); for instance, another well-known (and somewhat simpler) variant of the argument works by considering a photon emitted from one side of a box and absorbed on the other. More generally, any other energy emission which could potentially in the future decompose entirely into photons would also be handled by this argument, thanks to conservation of energy. Now, it is possible that other conservation laws prevent decomposition into photons; for instance, the law of conservation of charge prevents an electron (say) from decomposing entirely into photons, thus leaving open the possibility of having to add a linearly charge-dependent correction term to the formula  $E = mc^2$ . But then one can renormalise away this term by redefining the energy to subtract such a term; note that this does not affect conservation of energy, thanks to conservation of charge.

**1.17.6. Notes.** This article was originally posted on December 28, 2007 at [terrytao.wordpress.com/2007/12/28](http://terrytao.wordpress.com/2007/12/28)

Laurens Gunnarsen pointed out that Einstein's argument required the use of quantum mechanics to derive the equation  $E = mc^2$ , but that this equation can also be derived within the framework of classical mechanics by relying more heavily on the representation theory of the Lorentz group.

Thanks to Blake Stacey for corrections.