Some Early Analytic Number Theory

Carl D. Offner

Contents

1 Introduction

In these notes, we are going to touch upon three topics:

1. Euler's proof that the number of primes is infinite.
2. How Euler found the sum $\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$
3. Dirichlet's evaluation of the probability that two positive integers chosen at random are relatively prime.

There is some relationship between all three of these topics, oddly enough. As is already evident, the first two results are due to Leonhard Euler. Euler, who lived from 1707 to 1783, was the greatest mathematician of the 18^{th} century, and the most prolific mathematician of all time. His collected works in mathematics and physics take up more than 60 bound volumes.

Euler's mathematical intuition and skill were formidable. The ideas he came up with that we are going to write about here are still bearing fruit.

The third of these topics is a result due to Dirichlet in 1849, which uses Euler's result in the second topic. We do not follow his original reasoning however. Instead, we first give a plausibility argument for his result, based on some simple ideas of mathematical probability. Then we give a rigorous proof, based on a simple version of the Lebesgue bounded convergence theorem.

These results were astonishing when they were first published. They are examples of how seemingly distinct branches of mathematics come unexpectedly together. In particular, one might ask: How is it that the number π, which arose first in geometry, can have anything at all to do with the sequence of prime numbers?

2 Euler

2.1 The geometric series and the harmonic series

The *geometric series*

$$1 + x + x^2 + x^3 + \cdots + x^n + \cdots$$

converges for $|x| < 1$. This is easy to see: Let S_n denote the sum of the first n terms:

$$S_n = 1 + x + x^2 + \cdots + x^{n-1}$$

Then we have

$$\begin{aligned} S_n &= 1 + x + x_2 + \cdots + x^{n-1} \\ xS_n &= \quad x + x^2 + \cdots + x^{n-1} + x^n \end{aligned}$$

and so upon subtracting we get

$$S_n - xS_n = 1 - x^n$$

and so

$$S_n = \frac{1 - x^n}{1 - x}$$

If $|x| < 1$, then $x^n \to 0$ as $n \to \infty$, so

$$\lim_{n\to\infty} S_n = \frac{1}{1 - x}$$

and this is what is meant when we write it as an infinite sum:

$$\sum_{n=0}^{\infty} x_n = \frac{1}{1 - x} \qquad \text{for all } x \text{ such that } |x| < 1$$

The geometric series is a very rapidly converging series. This is because the individual terms go to zero exponentially. (In fact, a better name for the geometric series would be the exponential series.) For instance, if $x = 1/5$, the series is

$$1 + \frac{1}{5} + \frac{1}{25} + \frac{1}{125} + \frac{1}{625} + \frac{1}{3125} + \frac{1}{15625} + \cdots$$

Each term is $1/5$ the size of the previous one, and it is evident that the convergence is quite fast—if one did not know that the sum was going to be

$$\sum_{n=0}^{\infty} \left(\frac{1}{5}\right)^n = \frac{1}{1 - \frac{1}{5}} = 5/4 = 1.25$$

then one could simply compute the partial sums

n	S_n
0	1
1	1.2...
2	1.24...
3	1.248...
4	1.2496...
5	1.24992...
6	1.249984...

Now suppose that

$$a_1 + a_2 + a_3 + \cdots = \sum_{n=1}^{\infty} a_n$$

is *any* convergent sequence. (Note that here we are letting the index variable n run from 1 to ∞, rather than from 0 to ∞. This is just a matter of convenience. Sometimes it is more convenient to do it one way, and sometimes the other, depending on what series we are dealing with.)

Certainly the individual terms a_n have to converge to 0—otherwise the series could not possibly converge.

On the other hand, there are series in which the individual terms tend to 0 but the series as a whole does not converge. (That is, the partial sums S_n do not converge.) The most famous example is the *harmonic series*

$$\sum_{n=1}^{\infty} \frac{1}{n} = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \cdots$$

The usual way to see that this series does not converge is to group all the terms (except for the first two) into blocks, each block having twice the number of terms as the previous one:

$$1 + \frac{1}{2} + \underbrace{\frac{1}{3} + \frac{1}{4}}_{\frac{1}{2}} + \underbrace{\frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8}}_{\frac{1}{2}} + \underbrace{\frac{1}{9} + \frac{1}{10} + \frac{1}{11} + \frac{1}{12} + \frac{1}{13} + \frac{1}{14} + \frac{1}{15} + \frac{1}{16}}_{\frac{1}{2}} + \cdots$$

The way it works is this: the two terms in the first bracket are both $\geq \frac{1}{4}$, so their sum is $\geq \frac{1}{4} + \frac{1}{4} = \frac{2}{4} = \frac{1}{2}$, which is what we have written underneath. Similarly the four terms in the next bracket are each $\geq \frac{1}{8}$, so their sum is $\geq \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{4}{8} = \frac{1}{2}$. And so on. From this we can see that the sum of this series gets arbitrarily large as we take more and more terms. Of course it gets larger and larger "slower and slower", but the partial sums still eventually diverge to ∞: they get larger than any specified number if we are willing to wait long enough.

So this series diverges, but slowly. We can get an idea of how slowly it diverges by looking at the function $f(x) = \frac{1}{x}$. We know that the integral of this function from 1 to x is just $\log x$, the natural logarithm of x (see Figure 1).

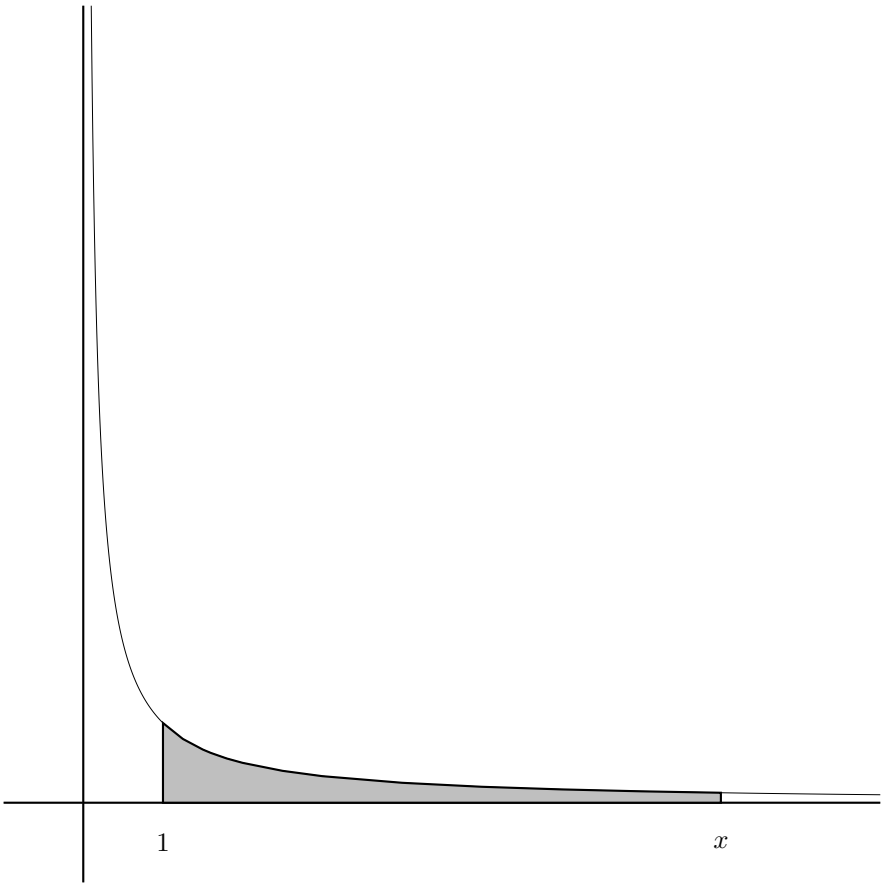


Figure 1: The shaded area is $\log x$.

In Figure 2, we show how to build rectangles underneath this curve whose sizes are just the terms of the harmonic series.

From this figure, we can see that $\log x$ grows at least as fast as the harmonic series, and so it tends to ∞ as $x \to \infty$. It may be hard to believe that the shaded area under the curve in Figure 1 gets very large as $x \to \infty$, but that is what we have just proved.

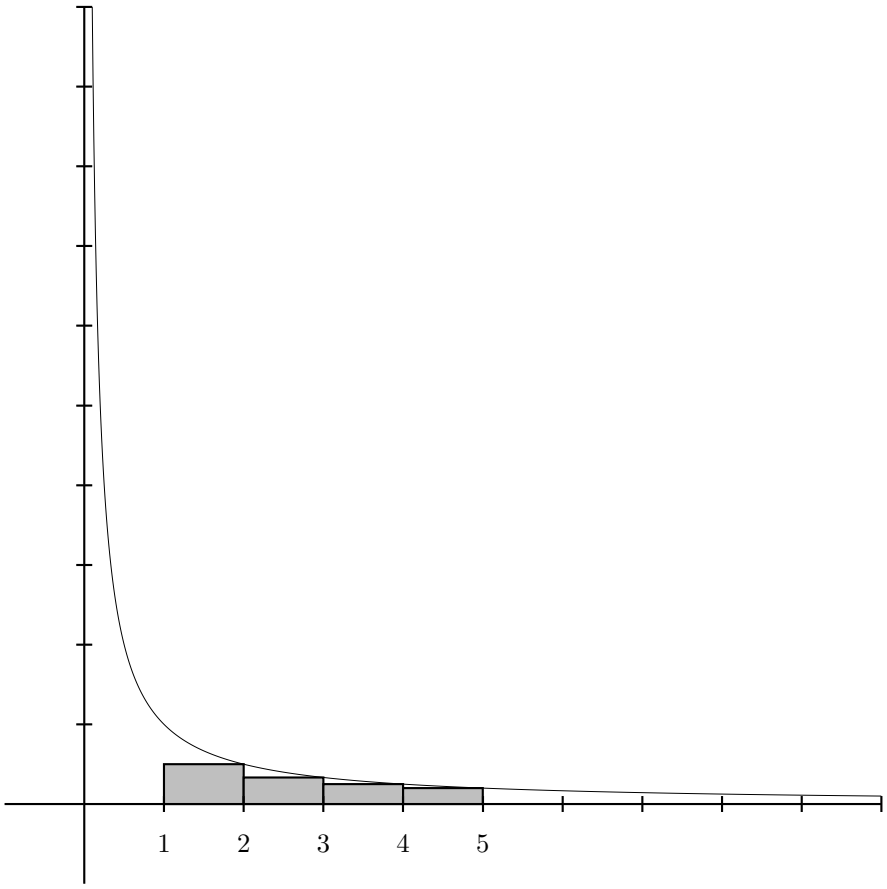


Figure 2: The rectangles have area $\frac{1}{2}$, $\frac{1}{3}$, $\frac{1}{4}$, ...

2.2 Euler's proof that there is an infinite number of prime numbers

Euler used the fact that the harmonic series diverges as a basis of a proof that there are an infinite number of prime numbers. Here is how he did it:

Remember that every positive integer is the product of prime numbers, and this product is unique. That is, there is only one such product, up to the order of the factors. For instance,

$$2352 = 2^4 \cdot 3^1 \cdot 7^2$$

and the exponents (4, 1, 2) are uniquely determined by the number 2352; no other exponents would work. We can use exponents of 0 to fill in the "missing primes", like this:

$$2352 = 2^4 \cdot 3^1 \cdot 5^0 \cdot 7^2$$

In fact, we could include all the primes this way:

$$2352 = 2^4 \cdot 3^1 \cdot 5^0 \cdot 7^2 \cdot 11^0 \cdot 13^0 \cdot 17^0 \cdots$$

This looks like an infinite product, but in fact, all but a finite number of its factors are 1, so it is really just an ordinary finite product.

We can express all this a little more generally. We number the primes in order, like this

n	p_n
1	2
2	3
3	5
4	7
5	11
6	13
7	17
⋮	⋮

Then using this numbering, any positive integer a has a unique representation

$$a = p_1^{\alpha_1} p_2^{\alpha_2} p_3^{\alpha_3} \cdots = \prod_{n=1}^{\infty} p_n^{\alpha_n}$$

where all but a finite number of the exponents α_n are 0 (so the corresponding factors $p_n^{\alpha_n}$ are 1).

Now here is what Euler did: Suppose that there are only a finite number of primes. Say there are N of them. As above, we enumerate them as $\{p_1, p_2, \ldots, p_N\}$. Now we know that for any n we can write

$$\frac{1}{1 - \frac{1}{p_n}} = 1 + \frac{1}{p_n} + \frac{1}{p_n^2} + \frac{1}{p_n^3} + \cdots = \sum_{i=0}^{\infty} \frac{1}{p_n^i}$$

(We know this because this is just the sum of the geometric series $1 + a + a^2 + a^3 + \ldots$, with $a = \frac{1}{p_n}$.)

So now let us take the product of these expressions over all N primes:

$$\prod_{n=1}^{N} \frac{1}{1 - \frac{1}{p_n}} = \prod_{n=1}^{N} \sum_{i=0}^{\infty} \frac{1}{p_n^i}$$

The left hand side is a finite product of finite numbers. We don't really care what its value is. All that matters is that it is obviously a finite number.

Now let us look at the right-hand side. It is a product of sums. It looks like this:

$$(1) \quad \left(1 + \frac{1}{p_1} + \frac{1}{p_1^2} + \cdots\right)\left(1 + \frac{1}{p_2} + \frac{1}{p_2^2} + \cdots\right)\left(1 + \frac{1}{p_3} + \frac{1}{p_3^2} + \cdots\right) \cdots \left(1 + \frac{1}{p_N} + \frac{1}{p_N^2} + \cdots\right)$$

If we imagine multiplying this product out completely, we would take one term from each of the N sums and multiply them together to get a product. We would form a product from each possible

combination in this way, and add those products up. That would give us the value of the expression on the right hand side.

Now each of these products is of the form

$$\frac{1}{p_1^{\alpha_1}} \frac{1}{p_2^{\alpha_2}} \cdots \frac{1}{p_N^{\alpha_N}} = \prod_{n=1}^{N} \frac{1}{p_n^{\alpha_n}}$$

Every power of every prime occurs exactly once in one of the sums in the expression (1). Therefore, there will be exactly one term like this for each different collection of exponents $\{\alpha_1, \alpha_2, \ldots, \alpha_N\}$.

We know that each positive integer a has a representation

$$a = \prod_{n=1}^{N} p_n^{\alpha_n}$$

and so

$$\frac{1}{a} = \prod_{n=1}^{N} \frac{1}{p_n^{\alpha_n}}$$

That is, for each positive integer a, there is precisely one of the multiplied-out terms on the right-hand side that equals $\frac{1}{a}$, and every multiplied-out term on the right-hand side is $\frac{1}{a}$ for some such a. Therefore, the right-hand side must equal

$$1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \cdots$$

That is, the right-hand side must equal the sum of the harmonic series. But we know this series diverges—it does not have a finite sum. Therefore our assumption that there is only a finite number N of primes must be incorrect, and the proof is complete.

Now this proof of Euler's was by no means the first proof that the number of primes is infinite. Euclid had already written down a simple and very elegant proof 2000 years previously. Neverthless, Euler's proof is quite clever, and the ideas in the proof also lead to some other remarkable results, as we shall see below.

2.3 The series $\sum_{n=1}^{\infty} \frac{1}{n^2}$

We saw above that the harmonic series

$$\sum_{n=1}^{\infty} \frac{1}{n} = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \cdots$$

diverges, even though the individual terms tend to 0. Now if a is any number between 0 and 1, then $0 < a^2 < a$. That is, squaring a small number makes it smaller. So it might occur to us to see

if forming a new series by squaring each term of the harmonic series yields a series that converges. That is, we consider the series

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = 1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \frac{1}{25} + \cdots$$

It turns out that this series actually does converge. We can see this by playing the same game with graphs that we did with the harmonic series. If we build rectangles whose areas are the terms in this series, then these rectangles lie beneath the graph of the function $f(x) = \frac{1}{x^2}$. Figure 3 shows what this looks like.

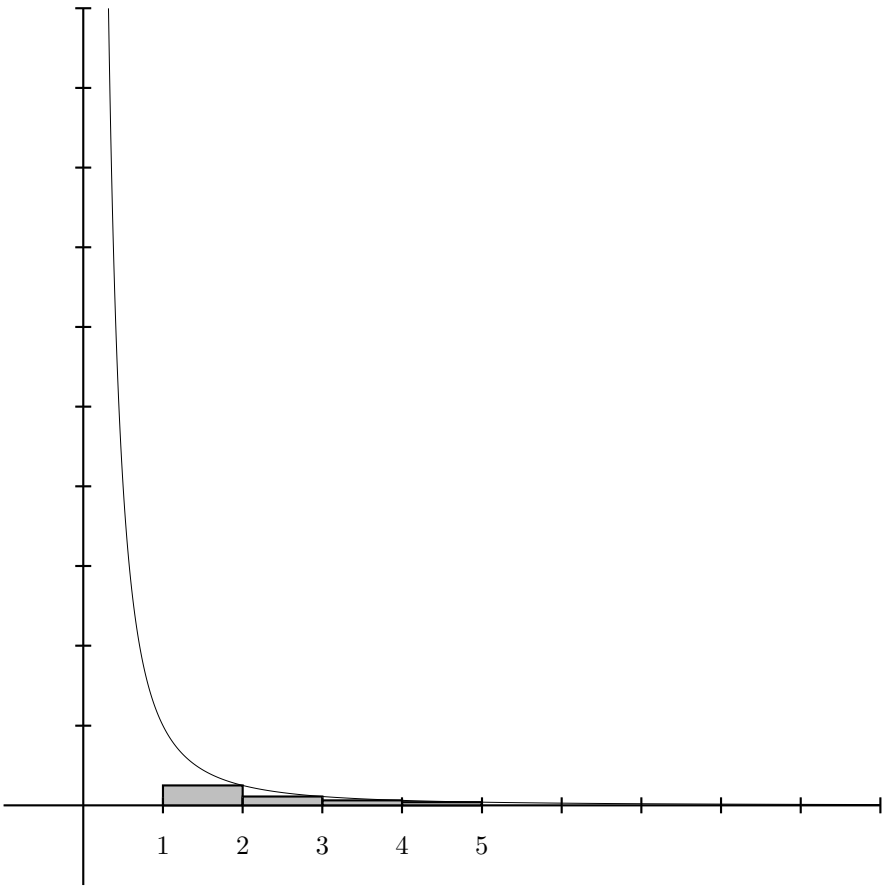


Figure 3: The graph of the function $f(x) = \frac{1}{x^2}$. The rectangles have area $\frac{1}{2^2}$, $\frac{1}{3^2}$, $\frac{1}{4^2}$, ...

It is apparent from the graph that this function tends to 0 as $x \to \infty$ much more quickly than the function $\frac{1}{x}$. And we also see that the terms of this series also get smaller much quicker than the terms of the harmonic series. (And by the way, the terms of the geometric series get smaller much quicker than the terms of either of these series.)

Now in fact, we can see that the total area under the graph of the function $f(x) = \frac{1}{x^2}$ from 1 to ∞ is finite. For it is just

$$\int_1^\infty \frac{1}{x^2}\,dx = -\frac{1}{x}\Big|_1^\infty = 1$$

Therefore, the sum of all the terms of this series except the first, which is represented by the sum of the areas of all the rectangles, is less than this, and so must be finite—less than 1, in fact. This shows that the series converges, and we see that actually, the sum of the whole series is less than 1 + the sum of the rest of the terms, and so is less than 2. And it's also greater than 1, since the first term is 1. So the sum is a number between 1 and 2.

Well, that's not bad. It shows how calculus can be applied to give us a handle on the convergence of a series. Of course, the question then arises: can we find a simple expression for the *exact* sum of the series?

Many people wondered about this question. It was finally answered by Euler, using another ingenious argument, which we will now present.

2.4 Linear factors of polynomials

First, we have to say a few things about polynomials.

Polynomials have the following property: If $p(x)$ is a polynomial, and if a is any number, then $x - a$ is a factor of $p(x)$ if and only if $p(a) = 0$. This is just because, no matter what the number a is, we can divide $p(x)$ by $x - a$ to get a quotient polynomial $q(x)$ and a remainder r. (The remainder r is always a number, not a polynomial.)

For instance, if $p(x)$ is the polynomial $x^5 - 5x^4 - 11x^3 + 57x^2 - 7x - 12$ and a is 5, we can divide $p(x)$ by $x - 5$, just as we learned in high school (see Figure 4). In that figure, the remainder r is 3. We can also write the result of the division in that figure as follows:

$$x^5 - 5x^4 - 11x^3 + 57x^2 - 7x - 12 = (x^4 - 11x^2 + 2x + 3)(x - 5) + 3$$

This kind of division can be done in general: we always can divide $p(x)$ by $x - a$ to get a quotient $q(x)$ and a remainder r, and we can write this as

$$p(x) = q(x)(x - a) + r$$

From this equation, we see that $x - a$ is a factor of the polynomial $p(x)$ if and only if $r = 0$.

Also, if we substitute a for x in this equation, we get $p(a) = r$. This shows that

$$\begin{aligned} x - a \text{ is a factor of } p(x) &\iff \text{the remainder } r \text{ is } 0 \\ &\iff p(a) = 0 \end{aligned}$$

A number a such that $p(a) = 0$ is called a *zero* of the polynomial $p(x)$.

Thus, if we know the zeros of a polynomial (i.e., if we know where a polynomial takes the value 0), we also know its linear factors. For instance, if we have been given the polynomial

$$p(x) = x^4 - 2x^3 - 41x^2 + 42x + 360$$

$$
\begin{array}{rr|l}
 & x^4 - 11x^2 + 2x + 3 & \\
\hline
x - 5 & x^5 - 5x^4 - 11x^3 + 57x^2 - 7x - 12 & \\
 & \underline{x^5 - 5x^4} & \\
 & -11x^3 & \\
 & \underline{-11x^3 + 55x^2} & \\
 & 2x^2 - 7x & \\
 & \underline{2x^2 - 10x} & \\
 & 3x - 12 & \\
 & \underline{3x - 15} & \\
 & 3 &
\end{array}
$$

Figure 4: Dividing polynomials, as we learned to do it in high school. The remainder is 3.

and we know it vanishes when x is -3, 4, -5, and 6, then we know at once that p factors as

$$p(x) = A(x+3)(x-4)(x+5)(x-6)$$

for some value of A.

This is in fact the complete factorization of the polynomial $p(x)$ because both sides of this equation have degree 4, and any other factor would increase the degree of the right-hand side. Furthermore, A must be 1, because the highest order term in $p(x)$ is x^4, and on the other hand, the highest order term on the right-hand side, when we multiply it out, is Ax^4.

This is very useful as a way of finding the factors of a polynomial, because in many cases it is not hard to find the zeros of the polynomial.

There is another way this can be written: in the example we just used, we could also write

$$\begin{aligned} x+3 \quad &\text{as} \quad 3\left(1+\frac{x}{3}\right) \\ x-4 \quad &\text{as} \quad -4\left(1-\frac{x}{4}\right) \end{aligned}$$

and so on. This way, we can write

$$p(x) = 360\left(1+\frac{x}{3}\right)\left(1-\frac{x}{4}\right)\left(1+\frac{x}{5}\right)\left(1-\frac{x}{6}\right)$$

Further, the constant 360 is just $p(0)$, as we can see by substituting in 0 for x, and in general, we see that if a polynomial $p(x)$ of degree n has zeros at $\{a_1, a_2, \ldots, a_n\}$, and if none of the numbers

a_i is 0, then

$$p(x) = A \prod_{i=1}^{n} \left(1 - \frac{x}{a_i}\right) \tag{2}$$

where in this case $A = p(0)$.

If $p(x)$ has a factor x—that is, if $p(0) = 0$—then we can divide $p(x)$ by x to get a polynomial $q(x)$, and we could continue dividing until we get a polynomial that does not have a factor of x. The product representation (2) then applies to that polynomial.

In the case $p(x)$ has exactly one factor of x—that is, in the case that x is a factor of $p(x)$ but x^2 is not—we see that we get by this method

$$\frac{p(x)}{x} = A \prod_{i=1}^{n} \left(1 - \frac{x}{a_i}\right)$$

or equivalently,

$$p(x) = Ax \prod_{i=1}^{n} \left(1 - \frac{x}{a_i}\right)$$

In this case, the degree of the polynomial is $n + 1$.

2.5 How Euler evaluated $\sum_{n=1}^{\infty} \frac{1}{n^2}$

Euler, as well as all mathematicians in his generation, knew that many functions have series expansions

$$f(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots$$

Series like this are called *power series*. Euler and his contemporaries also knew that there is a nice formula for the coefficients of such a series. In fact, you may already know, as they did, the power series for $\sin x$:

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots$$

But even if you don't, this series is easy to derive, just by finding successive derivatives and evaluating at 0:

First we find the derivatives. We know how to take the derivative of a power series (i.e., term-by-term), and we know how to take the derivative of a trigonometric function, so we do this both

ways:

$$\begin{aligned}
\sin x &= a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots \\
\frac{d}{dx}\sin x &= a_1 + 2a_2 x + 3a_3 x^2 + \cdots && \frac{d}{dx}\sin x = \cos x \\
\frac{d^2}{dx^2}\sin x &= 2a_2 + 6a_3 x + \cdots && \frac{d^2}{dx^2}\sin x = -\sin x \\
\frac{d^3}{dx^3}\sin x &= 6a_3 + \cdots && \frac{d^3}{dx^3}\sin x = -\cos x
\end{aligned}$$

Thus, substituting 0 for x, we have

$$\begin{aligned}
a_0 &= \sin 0 = 0 \\
a_1 &= \cos 0 = 1 \\
2a_2 &= -\sin 0 = 0 \\
6a_3 &= -\cos 0 = -1
\end{aligned}$$

Putting this all together, we have

$$\sin x = x - \frac{1}{6}x^3 + \text{terms of degree 4 or higher}$$

or, dividing by x,

$$\frac{\sin x}{x} = 1 - \frac{1}{6}x^2 + \text{terms of degree 3 or higher} \tag{3}$$

Now a power series can be thought of as a polynomial of infinite degree. When you look at the graph of the function $f(x) = \sin x$ (see Figure 5), you can see that it has zeros at the points $\{0, \pm\pi, \pm 2\pi, \pm 3\pi, \dots\}$. Euler had the remarkable idea that these zeros of $\sin x$ should correspond to linear factors of the power series for $\sin x$, just as the zeros of ordinary polynomials correspond to linear factors of those polynomials. Thus, if we take account of the factor x (which corresponds,

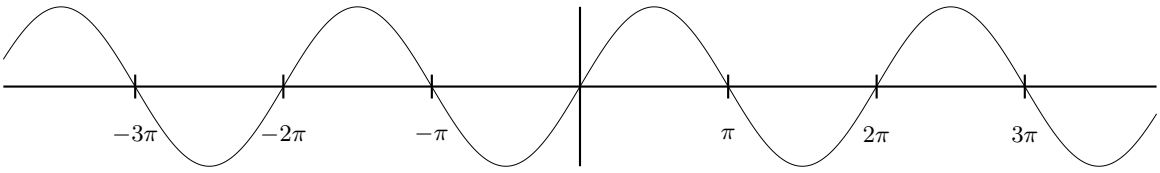


Figure 5: The graph of the function $f(x) = \sin x$.

as in the case of polynomials, to the zero at 0), we can write—formally, at least—

$$\sin x = Ax \prod_{n=1}^{\infty} \left(1 + \frac{x}{n\pi}\right)\left(1 - \frac{x}{n\pi}\right) \tag{4}$$

This was really a bold thing to do. It immediately raises all sorts of questions, such as

- Does the infinite product actually converge?
- Does the order in which we have written the factors matter?
- Even if it does converge, does it really converge to $\sin x$?

Let us ignore these questions for the moment and proceed formally. The first thing to do is to find the value of the constant multiplier A. Now if the equation (4) is true, then dividing by x, we have

$$\frac{\sin x}{x} = A \prod_{n=1}^{\infty} \left(1 + \frac{x}{n\pi}\right)\left(1 - \frac{x}{n\pi}\right)$$

We know that the limit of the left-hand side as $x \to 0$ is 1. And if we just substitute $x = 0$ in the right-hand side, we get A (since each of the factors in parentheses reduces to 1). So we must have $A = 1$, and we have

$$\sin x = x \prod_{n=1}^{\infty} \left(1 + \frac{x}{n\pi}\right)\left(1 - \frac{x}{n\pi}\right)$$

Using the fact that $(1 + t)(1 - t) = 1 - t^2$, we can rewrite this as

$$\sin x = x \prod_{n=1}^{\infty} \left(1 - \frac{x^2}{n^2\pi^2}\right) \tag{5}$$

This is a wonderful identity. It turns out that all this can be rigorously justified by means of techniques that were developed in the 19^{th} century, well after Euler's death. However, Euler did examine this identity critically, including testing it numerically, and convinced himself and others that it was in fact true.

Now let us compare this to what we get from the identity (3). First, let us rewrite (5) as a product for $\frac{\sin x}{x}$, to make it consistent with the power series (3):

$$\frac{\sin x}{x} = \prod_{n=1}^{\infty} \left(1 - \frac{x^2}{n^2\pi^2}\right) \tag{6}$$

We can imagine multiplying out this infinite product of binomials and adding up all the terms to get a power series, which must of course be the power series (3).

To multiply out the infinite product, we take one term from each factor and multiply them together. The only way to get a constant term is to multiply all the constant terms (i.e., the first term in each factor). These terms are all 1, so their product is 1, and that is indeed the constant term in (3).

To get the next term in (3) (i.e, the term $-\frac{1}{6}x^2$), we take the constant terms from *all but one* factor in (6), and the other term from that factor. All the constant terms are 1, and if the factor is the n^{th} factor, we get $-\frac{x^2}{n^2\pi^2}$. Adding all these products together and setting the result equal to the second term in the power series (3), we get

$$-\frac{1}{6}x^2 = \sum_{n=1}^{\infty} -\frac{x^2}{n^2\pi^2}$$

We can get rid of the minus sign on both sides. Then we can divide out by x^2 and multiply through by π^2. We come up with

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$$

That is the ingenious way in which Euler found the sum of this series. Note, by the way, that $\pi^2/6 = 1.644934\ldots$, which is indeed between 1 and 2.

3 Dirichlet

3.1 Some simple ideas of probability

Suppose we have 35 marbles. Some are translucent and the rest are opaque. Some of each kind are red and the rest are blue. Say they are divided up as in Figure 6.

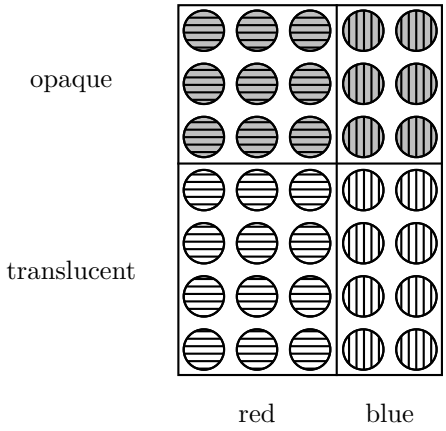


Figure 6: 35 marbles

We can talk about probabilities with respect to this set. For instance, we can say that the probability that a marble chosen at random from this set is red is 3/5. Similarly, the probability that a blue marble is opaque is 3/7.

Mathematical probability consists of the following generalization of this kind of model: We start with

- A set X. Let us make things simple and assume in this section that X is a finite set.

  An example is the set X consisting of the set of 35 marbles.

  There are of course many other finite sets. For instance, there is a set Y consisting of the ordered pairs $\langle n, m\rangle$, where n and m are both integers between 1 and 30. This set has $30^2 = 900$ elements.

- A function P (for "probability") that maps subsets of X into real numbers between 0 and 1. That is, it assigns to each subset A of X a number $P(A)$, which we call the probability of the set A. The intuitive idea is that "the probability that an element of X should belong to the subset A of X" is given by the value $P(A)$.

  For example, in the set X of 35 marbles defined above, the subset R of red marbles contains $3/5$ of the elements of X, and so it is reasonable to say that $P(R) = 3/5$. That is, the probability that a marble chosen at random from X is red is $3/5$.

  Of course, this depends on how the marble was chosen. We are assuming that all choices of marbles are equally likely. Another way of saying this is to say that P assigns to each 1-element subset of X the value $1/35$.

  To be a little more precise, let us use the notation $|A|$ to denote the number of elements in the set A. So for instance $|R| = 21$. Then our probability function P in this marble example is defined by

  $$P(A) = \frac{|A|}{|X|}$$

  for any set A.

- There is a certain use of language that is typical of this field: a subset of X corresponds to a property. For instance, the subset R corresponds to the property "being red". In fact, we have four obvious subsets:

R	being red
B	being blue
T	being translucent
O	being opaque

  The probability that a marble is red is the probability that it is in the set R, which is $P(R)$.

  Combinations of these sets correspond to combinations of their properties. For instance, a marble is red *and* translucent if and only if it is an element of the set $R \cap T$. That is, the "and" of two properties corresponds to the intersection of their two subsets.

  Similarly, the "or" of two properties corresponds to the union of their two subsets. For instance, a marble is red *or* translucent (or both—by "or" we mean the "inclusive or") if and only if it belongs to the subset $R \cup T$.

- The function P must have the following properties:

  1. $P(\varnothing) = 0$.
  2. $P(X) = 1$.
  3. If A and B are *disjoint* subsets of X (i.e., they have no elements in common—equivalently, $A \cap B = \varnothing$), then

     $$P(A \cup B) = P(A) + P(B)$$

  The first two properties are obvious.

The third property is also easy to see: If for instance

$$A = R \cap T = \text{set of red translucent marbles}$$
$$B = B \cap O = \text{set of blue opaque marbles}$$

then

$$P(A) = 12/35$$
$$P(B) = 6/35$$
$$P(A \cup B) = 18/35$$

and clearly $P(A \cup B) = P(A) + P(B)$.

So these three properties are exactly what we would expect of a function P in order for it to give us a notion of probability.

Note that if for any subset A we denote its complement $X - A$ by $\overline{A}$, then A and $\overline{A}$ are disjoint and so, using items 2 and 3, we have

$$1 = P(X) = P(A \cup \overline{A}) = P(A) + P(\overline{A})$$

so

$$P(\overline{A}) = 1 - P(A)$$

This also makes intuitive sense. For instance, the probability that a marble in X is red is $3/5$, and the probability that it is *not* red is $1 - 3/5 = 2/5$.

Similarly, the same reasoning shows that if $B \subset A$, then $P(A) = P(B) + P(A - B)$.

Now this kind of example is so simple that it seems trivial. The mathematical theory of probability is developed much farther, along the following lines:

- It is not necessary that all points of X (i.e., all single-element subsets of X) have the same probability. There are many natural cases in which they do not. (It is always true, however, that—if X is a finite set—the probabilities of all the single-element subsets of X must add up to 1.)
- The theory can be extended—although the mathematics becomes much more sophisticated—to sets X that are infinite. Many subtle questions then arise. For instance, what is the probability that a real number between 0 and 1 is rational? What is the probability that a continuous function is differentiable? Questions like this are deep. We are going to ignore all of them.

The reason we are not going to go very deeply into the theory of probability is that we need it mainly to motivate what we are going to do in the rest of this paper. We are not actually going to prove anything rigorously with it.

We need one important idea, however. We need to know what it means for two subsets to be *independent*, or equivalently, for two properties to be independent.

The idea is that two properties are independent if knowing one does not give you any information about the other. For instance, consider the two properties "red" and "translucent" in our marble example.

The probability that a marble is red is $P(R) = 3/5$. Now just considering the marbles that are translucent (this amounts to momentarily letting X be the set T), the probability that such a translucent marble is red is also $3/5$. That is to say, knowing that a marble is translucent does not make it more or less probable that the marble is red. For this reason, we say that the property of being translucent is *independent* of the property of being red, or equivalently, the sets T and R are independent.

On the other hand, if we consider the properties "red" and "blue", we see that they cannot be independent. The probability that a marble (chosen at random from the set X) is blue is $2/5$. But if we know that the marble is red to begin with (so the set X is in effect the set R), then the probability that it is also blue is 0. So knowing whether or not a marble is red gives us *a lot* of information about whether it is blue.

Now let us return to considering the independence of the properties "red" and "translucent". We can make our analysis more precise in the following way: We have defined our probability function P in the marbles example by

$$P(A) = \frac{|A|}{|X|}$$

What we have just seen is that the probability of R is

$$P(R) = \frac{|R|}{|X|} = 21/35 = 3/5$$

and the probability that a translucent marble is red is

$$\frac{|R \cap T|}{|T|} = 12/20 = 3/5$$

(Note that in this last equation, T takes the place of X, since we are considering only the set of translucent marbles.)

That is, we have

$$\frac{|R|}{|X|} = \frac{|R \cap T|}{|T|} = \frac{|R \cap T|}{|X|} \Big/ \frac{|T|}{|X|}$$

Equivalently,

$$P(R) = \frac{P(R \cap T)}{P(T)}$$

or simply

$$P(R \cap T) = P(R)P(T)$$

We turn this into a general definition: two sets A and B are *independent* if and only if

$$P(A \cap B) = P(A)P(B)$$

Note that independence is a very special property that a pair of sets may have. It corresponds to the sets being "orthogonal" in some representation such as in Figure 6. For example, in Figure 7, the sets R and T are *not* independent—one can see that knowing that a marble is translucent makes it less likely to be red, and in fact

$$12/35 = P(R \cap T) < P(R)P(T) = 24/35 \cdot 20/35$$

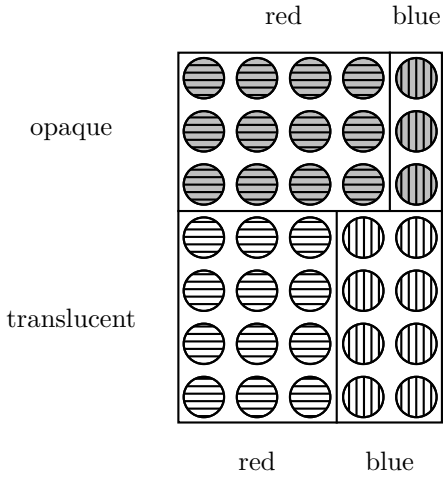


Figure 7: 35 marbles. In this example, "red" and "translucent" are not independent properties.

We can generalize this definition of independence to handle more than two sets. We say that a family of sets $\{A_i : i = 1, 2, \cdots\}$ is independent if for every subfamily $\{A_{i_j} : j = 1, 2, \cdots\}$ we have

$$P\left(\bigcap_j A_{i_j}\right) = \prod_j P(A_{i_j})$$

That is, for each subfamily of $\{A_i\}$, the probability of the intersection of the sets in that subfamily must equal the product of the probabilities of the sets.

Finally, we need the following result: if A and B are independent, then also $\overline{A}$ and B are independent. For,

$$\begin{aligned}
P(\overline{A} \cap B) &= P\big((X - A) \cap B\big) \\
&= P(B - A \cap B) \qquad \text{(just draw a Venn diagram)} \\
&= P(B) - P(A \cap B) \qquad \text{(because } A \cap B \subset B\text{)} \\
&= P(B) - P(A)P(B) \qquad \text{(because } A \text{ and } B \text{ are independent)} \\
&= \big(1 - P(A)\big)P(B) \\
&= P(\overline{A})P(B)
\end{aligned}$$

Of course, we can then apply this result to B to show that also $\overline{A}$ and $\overline{B}$ are independent.

The result can be extended in a straightforward manner to the case of more than two sets: for instance, if the sets $\{A_i\}$ are independent, then also the sets $\left\{\overline{A_i}\right\}$ are independent. We will use this fact in the next section.

3.2 The probability that two numbers are relatively prime

Given two positive integers picked "at random", what is the probability that they are relatively prime? This question is not really well-defined, because there is no unique way to specify what it means to pick a positive integer at random. However, we can give the question a conventional meaning as follows:

Given a number N, let $r(N)$ denote the probability that two integers chosen at random in the range 1 to N are relatively prime. More specifically, we assume that each integer has equal probability of being $1, 2, \ldots, N$—that is to say, the probability of each value between 1 and N is $1/N$. In addition, we assume that the two integers are chosen independently.

Of course $r(N)$ can be computed—there are only a finite number of different possibilities to consider for any given N. Then we can look to see if $r(N)$ has a limit as $N \to \infty$. If it does, we can say that this limit is the probability that two integers picked at random are relatively prime.

Dirichlet showed how to do this in 1849, and we will also perform this computation, although not precisely the way Dirichlet did it. (Actually, Dirichlet did not evaluate $r(N)$, but he evaluated a related quantity. The basic idea is his.)

Before doing this, however, we will give a plausibility argument for the result. This plausibility argument will use the fundamental ideas of elementary probability theory that we developed in the last section. In particular, we will see that the probability of being divisible by a number a is independent of the probability of being divisible by a number b provided a and b are relatively prime. We will use this to give a non-rigorous but plausible derivation of Dirichlet's result.

Then in the remainder of these notes we will show how all this can be made rigorous.

For convenience, when we say "number" in the following, we mean "positive integer". Let us use the notations

$$P(a \mid n) = \text{ the probability that } a \text{ divides } n$$

$$P\big((a \mid n) \text{ and } (a \mid m)\big) = \text{ the probability that } a \text{ divides } n \textit{ and } a \text{ divides } m$$

1. If a is any fixed number, the probability that a number n (chosen at random) is divisible by a is $1/a$. That is, $P(a \mid n) = 1/a$.

   This is just because in the first N numbers, $\lfloor N/a \rfloor$ are divisible by a. Therefore, the probability that one of the first N numbers is divisible by a is

   $$\frac{1}{N}\left\lfloor \frac{N}{a} \right\rfloor \tag{7}$$

   We know by definition that

   $$\frac{N}{a} - 1 < \left\lfloor \frac{N}{a} \right\rfloor \leq \frac{N}{a}$$

and so as $N \to \infty$, the expression in (7) tends to $1/a$.

2. If a is any fixed number, and if n and m are two numbers chosen at random, the probability that n and m are both divisible by a is $1/a^2$.

   This is because if we restrict n and m to be between 1 and N, there are N^2 possible pairs of numbers $\langle n, m\rangle$. Of these pairs, the number having both n and m divisible by a is $\lfloor N/a \rfloor^2$. Therefore, the probability that a pair $\langle n, m\rangle$, both elements of which are between 1 and N has both elements divisible by a is

   $$\frac{1}{N^2}\left\lfloor \frac{N}{a} \right\rfloor^2$$

   and as $N \to \infty$, this tends to $1/a^2$.

3. if a and b are two fixed numbers that are relatively prime and n is a number chosen at random, the properties

   - n is divisible by a
   - n is divisible by b

   are independent.

   This is true by the following reasoning: We know first that

   - The probability that n is divisible by a is $1/a$.
   - The probability that n is divisible by b is $1/b$.

   To say that these properties are independent is to say that the probability that n is divisible by *both* a and b is the product of these probabilities, namely, $1/(ab)$, which as we have seen above is just the probability that n is divisible by ab.

   That is, we have to show that the following two probabilities are the same:

   - The probability that n is divisible by both a and b.
   - The probability that n is divisible by ab.

   But these probabilities *are* the same, because since a and b are relatively prime, n is divisible by both a and b if and only if n is divisible by ab. This proves the result. The reasoning can be recast as follows:

   $$\begin{aligned} P\big((a \mid n) \text{ and } (b \mid n)\big) &= P(ab \mid n) \qquad \text{(since } a \text{ and } b \text{ are relatively prime)} \\ &= \frac{1}{ab} \\ &= \frac{1}{a}\frac{1}{b} \\ &= P(a \mid n)P(b \mid n) \end{aligned}$$

4. If a and b are two fixed numbers that are relatively prime and n and m are two numbers chosen at random, the probabilities

   - n and m are both divisible by a

- n and m are both divisible by b

are independent.

The reasoning is the same:

$$\begin{aligned}
&P\big(\{(a \mid n) \text{ and } (a \mid m)\} \text{ and } \{(b \mid n) \text{ and } (b \mid m)\}\big) \\
&\qquad = P\big(\{(a \mid n) \text{ and } (b \mid n)\} \text{ and } \{(a \mid m) \text{ and } (b \mid m)\}\big) \\
&\qquad = P\big((ab \mid n) \text{ and } (ab \mid m)\big) \qquad \text{(since } a \text{ and } b \text{ are relatively prime)} \\
&\qquad = \frac{1}{(ab)^2} \qquad \text{(by item 2)} \\
&\qquad = \frac{1}{a^2}\frac{1}{b^2} \\
&\qquad = P\big((a \mid n) \text{ and } (a \mid m)\big)P\big((b \mid n) \text{ and } (b \mid m)\big)
\end{aligned}$$

This result can be extended: the same reasoning shows that if $\{p_1, p_2, \ldots, p_n\}$ are distinct prime numbers, then the probabilities that any p_i divides both n and m are all independent.

5. Passing to complements in item 4, if p and q are two primes, and n and m are two numbers, then the probability that p does *not* divide both n and m (although it may divide one or the other) is independent of the probability that q does *not* divide both n and m.

   And similarly, we can prove that if $\{p_1, p_2, \ldots, p_n\}$ are distinct prime numbers, the probabilities that p_i does not divide both n and m are all independent.

Now to say that two numbers n and m are relatively prime is to say that there is no prime number that divides them both. We have seen that the probability that p divides both n and m is $1/p^2$. Therefore, the probability that p does *not* divide both n and m (i.e., it may divide one or the other, or neither, but not both) is

$$1 - \frac{1}{p^2}$$

A similar probability holds for each prime p, and these probabilities are all independent. Therefore, the probability that n and m are relatively prime is

$$\prod_{k=1}^{\infty}\left(1 - \frac{1}{p_k^2}\right) \tag{8}$$

This product should start to look familiar. In fact, if we call it P, then

$$\frac{1}{P} = \prod_{k=1}^{\infty}\frac{1}{1 - \frac{1}{p_k^2}}$$

Now remember we showed that

$$\prod_{k=1}^{\infty}\frac{1}{1 - \frac{1}{p_k}} = \sum_{n=1}^{\infty}\frac{1}{n} = \infty$$

We can use the same reasoning here. We have

$$\prod_{k=1}^{\infty} \frac{1}{1-\frac{1}{p_k^2}} = \left(1+\frac{1}{p_1^2}+\frac{1}{p_1^4}+\cdots\right)\left(1+\frac{1}{p_2^2}+\frac{1}{p_2^4}+\cdots\right)\cdots\left(1+\frac{1}{p_k^2}+\frac{1}{p_k^4}+\cdots\right)\cdots$$

The right-hand side is evaluated by taking one term from each of the expressions in parentheses and multiplying those terms together. Then the resulting products are added up. There is one such product for every *squared* number, since only even powers of primes are represented on the right-hand side of this identity.

Therefore, we must have

$$\frac{1}{P} = \prod_{k=1}^{\infty} \frac{1}{1-\frac{1}{p_k^2}} = \sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$$

Therefore, $P = 6/\pi^2$. That is, the probability that two numbers chosen at random are relatively prime is

$$\frac{6}{\pi^2}$$

This number evaluates to $0.6079271\ldots$. Thus, the probability that two numbers picked at random are relatively prime is about 3/5.

We mentioned that this was not a real proof but only a plausibility argument. Where does it break down? It turns out that everything here is quite rigorous with one exception: the formation of the expression (8). The problem is that we are multiplying infinitely many probabilities here, one for each prime number. Each of these probabilities is a limit probability (as $N \to \infty$). But what we really should have done is compute the probability $r(N)$ for each finite N (each such $r(N)$ will involve only a finite number of primes) and *then* let $N \to \infty$. It turns out that this is not particularly difficult to do. We need two techniques, however, and the next two sections are devoted to them. Then in the last section we use them to present a short but rigorous proof of Dirichlet's result.

As we will see, the entire rigorous proof, even with all the preliminary material, takes up somewhat less space than the plausibility argument based on probability that was just presented. Nevertheless, the probabilistic argument is important, because it provides us with a powerful intuitive sense of why the result is true.

3.3 The inclusion-exclusion principle

As before, if S is any finite set, we denote the number of elements of S by $|S|$.

Now suppose we have n subsets of X; call them A_i $(1 \le i \le n)$. We want to derive an expression for the number of elements of X that are not in *any* of the sets A_i. That is, we want to find an expression for

$$\left| X - \bigcup_{i=1}^{n} A_i \right|$$

Here is the idea: To make things simple, let us just start with two sets, and let us call them A and B. If A and B are disjoint, then it is clear that

$$|X - (A \cup B)| = |X| - |A| - |B|$$

But if A and B have any elements in common, the right-hand side of the equation above would subtract those common elements twice. So to fix that, we have to add them back in, and we get

$$|X - (A \cup B)| = |X| - |A| - |B| + |A \cap B|$$

or equivalently,

$$|X - (A_1 \cup A_2)| = |X| - |A_1| - |A_2| + |A_1 \cap A_2|$$

This is our result when the number n of sets A_i is 2.

When $n = 3$, we get, similarly, (see Figure 8)

$$|X - (A_1 \cup A_2 \cup A_3)| = |X| - |A_1| - |A_2| - |A_3| + |A_1 \cap A_2| + |A_1 \cap A_3| + |A_2 \cap A_3| - |A_1 \cap A_2 \cap A_3|$$

We could also write this as follows:

$$\left| X - \bigcup_{i=1}^{3} A_i \right| = |X| - \sum_{1 \le i_1 \le 3} |A_{i_1}| + \sum_{1 \le i_1 < i_2 \le 3} |A_{i_1} \cap A_{i_2}| - \sum_{1 \le i_1 < i_2 < i_3 \le 3} |A_{i_1} \cap A_{i_2} \cap A_{i_3}|$$

(Note that the last sum has only 1 term: $|A_1 \cap A_2 \cap A_3|$.)

It's pretty clear that this formula generalizes to any n. To derive it in general, we need a little more notation:

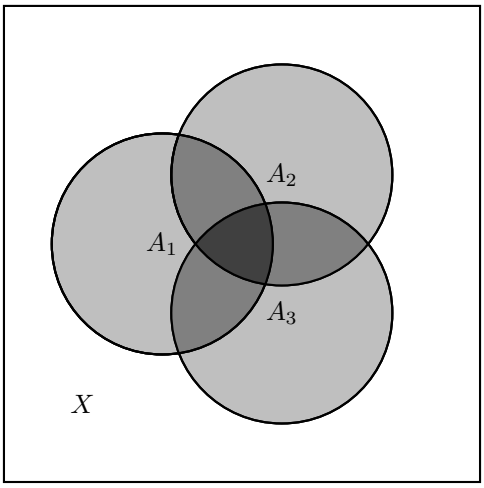
Suppose X is a finite set and $f : X \to \mathbf{R}$ is any function. That is, f is just an assignment of a real number $f(x)$ to each point x of the finite set X. Suppose we add all these numbers up to get a sum. We can think of this sum as being the "integral" of f over the set X. We write this using the integral sign, like this:

$$\int_X f = \sum_{x \in X} f(x)$$

(This use of the integration notation is entirely correct: what we are actually doing is taking the integral of f over X with respect to the "counting measure" on X. However, that terminology is overkill for what we are doing here.)

The usual rules of integration apply. For instance, if f and g are two such functions, and if a is any real number, we have

$$\int_X (f + g) = \int_X f + \int_X g$$
$$\int_X (af) = a \int_X f$$



$$|X - (A_1 \cup A_2 \cup A_3)| = |X| - |A_1| - |A_2| - |A_3| + |A_1 \cap A_2| + |A_1 \cap A_3| + |A_2 \cap A_3| - |A_1 \cap A_2 \cap A_3|$$

$$\left| X - \bigcup_{i=1}^{3} A_i \right| = |X| - \sum_{1 \le i_1 \le 3} |A_{i_1}| + \sum_{1 \le i_1 < i_2 \le 3} |A_{i_1} \cap A_{i_2}| - \sum_{1 \le i_2 < i_2 < i_3 \le 3} |A_{i_1} \cap A_{i_2} \cap A_{i_3}|$$

Figure 8: How to compute the size of the complement of a union of three sets.

For a particular example, let us define the *characteristic function* χ_A of a set A to be the function that is 1 on A and zero elsewhere:

$$\chi_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases}$$

Then it is clear that the integral of χ_A is just the number of elements of A:

$$\int_X \chi_A = |A|$$

Notice that if A and B are two subsets of X, then

$$\chi_{A \cap B}(x) = \chi_A(x) \chi_B(x)$$

and in fact this holds for any number of sets:

$$\chi_{\cap A_i}(x) = \prod_i \chi_{A_i}(x)$$

It is also easy to see that

$$\chi_{X-A}(x) = 1 - \chi_A(x)$$

Now we can use this notation to derive the formula for $|X - \cup A_i|$ that we started out considering above. First, we know that

$$X - \bigcup_{i=1}^{n} A_i = \bigcap_{i=1}^{n} (X - A_i)$$

Therefore we have

$$\begin{aligned}
&1 - \chi_{\cup_{i=1}^n A_i}(x) \\
&\quad = \prod_{i=1}^{n} (1 - \chi_{A_i}(x)) \\
&\quad = 1 - \sum_{1 \le i_1 \le n} \chi_{A_{i_1}}(x) + \sum_{1 \le i_1 < i_2 \le n} \chi_{A_{i_1}}(x)\chi_{A_{i_2}}(x) - \sum_{1 \le i_1 < i_2 < i_3 \le n} \chi_{A_{i_1}}(x)\chi_{A_{i_2}}(x)\chi_{A_{i_3}}(x) + \cdots \\
&\quad = 1 - \sum_{1 \le i_1 \le n} \chi_{A_{i_1}}(x) + \sum_{1 \le i_1 < i_2 \le n} \chi_{A_{i_1} \cap A_{i_2}}(x) - \sum_{1 \le i_1 < i_2 < i_3 \le n} \chi_{A_{i_1} \cap A_{i_2} \cap A_{i_3}}(x) + \cdots
\end{aligned}$$

and therefore, by integrating, we get

$$\left| X - \bigcup_{i=1}^{n} A_i \right| = |X| - \sum_{1 \le i_1 \le n} |A_{i_1}| + \sum_{1 \le i_1 < i_2 \le n} |A_{i_1} \cap A_{i_2}| - \sum_{1 \le i_1 < i_2 < i_3 \le n} |A_{i_1} \cap A_{i_2} \cap A_{i_3}| + \cdots$$

This identity is known as the *inclusion-exclusion principle*.

Note that since X is finite, there are really only a finite number of terms in this sum, as we already saw above when n was 2 or 3.

3.4 The bounded convergence theorem for series

There is an immensely useful theorem due to Lebesgue, known as the bounded convergence theorem. Here we are going to state and prove a simple version of this theorem.

Suppose that we have a sequence of pairs of numbers

$$a_1^{(N)}, a_2^{(N)} \qquad (1 \le N < \infty)$$

and suppose that as $N \to \infty$, both the first and second elements of each pair tend to a limit:

$$a_1^{(N)} \to a_1$$
$$a_2^{(N)} \to a_2$$

Then it is certainly true that

$$a_1^{(N)} + a_2^{(N)} \to a_1 + a_2$$

The same is true if, instead of a sequence of pairs of numbers, we have a sequence of K-tuples, where K is any fixed finite number: if as $N \to \infty$

$$a_n^{(N)} \to a_n \qquad \text{(for all } 1 \le n \le K\text{)}$$

then

$$\sum_{n=1}^{K} a_n^{(N)} \to \sum_{n=1}^{K} a_n \qquad \text{as } N \to \infty$$

Now suppose we let K "become infinite". That is, suppose instead of K-tuples of numbers, we have a family of series, indexed by N, so the n^{th} term of the N^{th} series is $a_n^{(N)}$. The first series looks like this:

$$a_1^{(1)} + a_2^{(1)} + a_3^{(1)} + \dots$$

The second series looks like this:

$$a_1^{(2)} + a_2^{(2)} + a_3^{(2)} + \dots$$

and so on.

Suppose that each of these series converges; say

$$\sum_{n=1}^{\infty} a_n^{(N)} = s_N$$

and furthermore, suppose that for each fixed n, $a_n^{(N)}$ converges to a limit a_n as $N \to \infty$. (That is, the first terms of the series converge to a_1, the second terms converge to a_2, and so on. We express this by saying that the series converge *termwise* to the series $\sum_{n=1}^{\infty} a_n$. We ask (as above in the case of series with a finite number of terms) if it is necessarily true that the sums s_N also converge to a limit s and if

$$\sum_{n=1}^{\infty} a_n = s$$

It is easy to see that in this generality, the answer is no. For consider the family of series defined as follows:

$$a_n^{(N)} = \begin{cases} 1 & \text{if } n = N \\ 0 & \text{otherwise} \end{cases}$$

Then the N^{th} series is identically 0 except for the N^{th} term, which is 1. The first series is

$$1, 0, 0, 0, \dots$$

The second series is

$$0, 1, 0, 0, \dots$$

and so on.

Certainly each sum s_N is 1, and the limit of these sums is clearly $s = 1$. On the other hand, the limit (as $N \to \infty$) of the n^{th} terms of these series is 0 for each n. (For as soon as $n > N$, $a_n^{(N)} = 0$.) Therefore, $a_n = 0$ for all n, and so the sum $\sum a_n = 0 \neq 1 = s$.

There is, however, a simple additional condition that, when it is satisfied, does away with this problem and allows us to conclude that s exists and $\sum a_n = s$.

The way it works is this: the only reason that our counterexample failed to work "correctly" was that the series "escaped to infinity". That is, each successive series $a^{(N)}$ was concentrated farther and farther out. And that is really the only thing that can go wrong.

To prevent the series from escaping to infinity, we introduce the following constraint:

Suppose that there is a convergent series $\sum_{n=1}^{\infty} b_n$ where all the b_n are ≥ 0, and suppose that for *every* N,

$$\left|a_n^{(N)}\right| \leq b_n \qquad \text{for all } n$$

In this case, we say that each series $\sum_{n=1}^{\infty} a_n^{(N)}$ is *majorized* (or *bounded*) by the series $\sum_{n=1}^{\infty} b_n$. Note that in such a case each series $\sum_{n=1}^{\infty} a_n^{(N)}$ must be absolutely convergent (since it is bounded termwise by the nonnegative convergent series $\sum_{n=1}^{\infty} b_n$), and so it is automatically convergent.

The bounding series $\sum_{n=1}^{\infty} b_n$ is used like this: since this series converges, we know that for each $\epsilon > 0$ ("no matter how small"), there is some number K (depending on ϵ, of course) such that

$$\sum_{n=K}^{\infty} b_n < \epsilon$$

Now we can imagine splitting each of the series $a^{(N)}$ into two parts. The first part consists of the first K terms, and the second part (which we can call the "tail") consists of the rest of the terms. Just as we reasoned above, the first K terms of each series converge pointwise, and so the first part of each series (i.e., the K^{th} partial sum) also converges. But the rest of each series is bounded by ϵ, since it is majorized by the "tail" $\sum_{n=K+1}^{\infty} b_n$ of the majorizing series.

So here is the complete proof:

3.1 Theorem *If*

- $\sum_{n=1}^{\infty} b_n$ *is a convergent series of non-negative terms, and*
- $\sum_{n=1}^{\infty} a_n^{(N)}$ *is a family of series that are majorized by the series* $\sum_{n=1}^{\infty} b_n$ *(and therefore have sums which we denote by* s_N*), and*
- *For each* n, $\lim_{N\to\infty} a_n^{(N)} = a_n$ *exists,*

then $lim_{N\to\infty} s_N = s$ *exists, and*

$$\sum_{n=1}^{\infty} a_n = s$$

PROOF. As a preliminary remark, note that since for each n all $\left|a_n^{(N)}\right| \leq b_n$, the limit of the $a_n^{(N)}$ must also be bounded in absolute value by b_n; that is, $|a_n| \leq b_n$. Therefore, $\sum a_n$ converges absolutely, and hence converges. (That is, it therefore also converges "non-absolutely".) Let us denote its sum by s:

$$\sum_{n=1}^{\infty} a_n = s$$

We will be done once we have shown that $s_N \to s$. Here's how we do this:

Given $\epsilon > 0$,

- Let K be so large that $\sum_{n=K}^{\infty} b_n < \epsilon$.
- Then, having fixed K, let N_0 be so large that for $N \geq N_0$ and $1 \leq n \leq K$, $\left|a_n^{(N)} - a_n\right| < \epsilon/K$. (We can do this because K is finite. Of course, N_0 depends on ϵ.)

Now for $N \geq N_0$, we have

$$\begin{aligned}
|s_N - s| &= \left|\sum_{n=1}^{\infty}(a_n^{(N)} - a_n)\right| \\
&\leq \sum_{n=1}^{K}\left|a_n^{(N)} - a_n\right| + \sum_{n=K+1}^{\infty}\left|a_n^{(N)} - a_n\right| \\
&\leq \sum_{n=1}^{K}\frac{\epsilon}{K} + 2\sum_{n=K+1}^{\infty} b_n \\
&\leq K\frac{\epsilon}{K} + 2\epsilon \\
&= 3\epsilon
\end{aligned}$$

That is, for large enough N, $|s_N - s| < 3\epsilon$. This completes the proof. □

As an aside, the two-step technique we used to pick K and N_0 so as to make $|s_N - s|$ small occurs often in analysis. It is often expressed simply by writing, "First make K large, then make N_0 large."

3.5 Making the proof rigorous

Remember that we have defined $r(N)$ to be the probability that two numbers picked independently with uniform probability between 1 and N are relatively prime. We will now compute $r(N)$.

The number of such pairs of numbers is N^2, and the number of such pairs that are both multiples of a prime p is $\left\lfloor \frac{N}{p} \right\rfloor^2$. Similarly, the number of such pairs that are both multiples of two primes p and q is $\left\lfloor \frac{N}{pq} \right\rfloor^2$, and so on.

Let $\{p_n : 1 \le i \le m\}$ denote the set of primes that are $\le N$. By the inclusion-exclusion principle, the number of pairs of numbers that are not both divisible by any of these primes is

$$N^2 - \sum_{1\le n_1\le m} \left\lfloor \frac{N}{p_{n_1}} \right\rfloor^2 + \sum_{1\le n_1<n_2\le m} \left\lfloor \frac{N}{p_{n_1}p_{n_2}} \right\rfloor^2 - \sum_{1\le n_1<n_2<n_3\le m} \left\lfloor \frac{N}{p_{n_1}p_{n_2}p_{n_3}} \right\rfloor^2 + \cdots$$

Note that although the sum is formally infinite, in actuality it is a finite sum for any specific value of N. In fact, we could substitute ∞ for m in this equation without changing anything, because by assumption, if $k > m$, $p_k > N$, so $\lfloor N/p_k \rfloor = 0$. So that is how we will write it, using ∞ instead of m.

The probability $r(N)$ then is this expression divided by N^2. That is,

$$r(N) =$$
$$1 - \sum_{1\le n_1<\infty} \frac{1}{N^2}\left\lfloor \frac{N}{p_{n_1}} \right\rfloor^2 + \sum_{1\le n_1<n_2<\infty} \frac{1}{N^2}\left\lfloor \frac{N}{p_{n_1}p_{n_2}} \right\rfloor^2 - \sum_{1\le n_1<n_2<n_3<\infty} \frac{1}{N^2}\left\lfloor \frac{N}{p_{n_1}p_{n_2}p_{n_3}} \right\rfloor^2 + \cdots$$

Now we know that, whatever the value of a,

$$\frac{1}{N^2}\left\lfloor \frac{N}{a} \right\rfloor^2 \to \frac{1}{a^2} \qquad \text{as } N \to \infty$$

If we could be justified in substituting these limiting values in the above sum, we would have

$$\begin{aligned}\lim_{N\to\infty} r(N) &= 1 - \sum_{1\le n_1<\infty} \frac{1}{p_{n_1}^2} + \sum_{1\le n_1<n_2<\infty} \frac{1}{(p_{n_1}p_{n_2})^2} - \sum_{1\le n_1<n_2<n_3<\infty} \frac{1}{(p_{n_1}p_{n_2}p_{n_3})^2} + \cdots \\ &= \prod_{k=1}^{\infty}\left(1 - \frac{1}{p_k^2}\right) \\ &= \frac{6}{\pi^2}\end{aligned}$$

So everything hinges on showing that we are allowed to pass to the limit "under the summation sign". (This is essentially what we did in a previous section in computing the elementary probabilities as limits first, and then combining them. That's why the reasoning in that section was not a rigorous proof but a plausibility argument.)

We show this by using the bounded convergence theorem of the last section.

First, we take the absolute value of every term in the series for $r(N)$. In this way we find that the series for $r(N)$ is majorized by the series

$$1 + \sum_{1\le n<\infty} \frac{1}{N^2}\left\lfloor \frac{N}{p_{n_1}} \right\rfloor^2 + \sum_{1\le n_1<n_2<\infty} \frac{1}{N^2}\left\lfloor \frac{N}{p_{n_1}p_{n_2}} \right\rfloor^2 + \sum_{1\le n_1<n_2<n_3<\infty} \frac{1}{N^2}\left\lfloor \frac{N}{p_{n_1}p_{n_2}p_{n_3}} \right\rfloor^2 + \cdots \tag{9}$$

Next, we use the fact that $\lfloor x \rfloor \le x$ for all x. This shows that the series (9) (and hence also the series for $r(N)$) is majorized by the series

$$1 + \sum_{1\le n_1<\infty} \frac{1}{p_{n_1}^2} + \sum_{1\le n_1<n_2<\infty} \frac{1}{(p_{n_1}p_{n_2})^2} + \sum_{1\le n_1<n_2<n_2<\infty} \frac{1}{(p_{n_1}p_{n_2}p_{n_3})^2} + \cdots \tag{10}$$

Now this series does not depend on N. In addition, it converges; in fact, it equals

$$\begin{aligned}\prod_{k=1}^{\infty}\left(1+\frac{1}{p_k^2}\right) &\leq \prod_{k=1}^{\infty}\left(1+\frac{1}{p_k^2}+\frac{1}{p_k^4}+\cdots\right)\\ &= \prod_{k=1}^{\infty}\frac{1}{1-\frac{1}{p_k^2}}\\ &= \frac{\pi^2}{6}\end{aligned}$$

The actual value doesn't matter; the point is that this series converges.

So $r(N)$ is given by a series which is majorized by the convergent series (10). Using this fact, the bounded convergence theorem assures us that we are justified in taking the limit of the series for $r(N)$ term-by-term before summing it, and that completes the proof.