COUNTING PRIMES

VIV

1. Day 1: Introduction, big-O and asymptotic notation

Our goal for this class is getting a sense of, roughly, how many primes there are. Today we're going to begin by making this goal precise, and understanding what we *should* be aiming for. Hopefully that'll make aiming for the goal easier!

Let's first define the *prime-counting function*, $\pi(x)$.

Definition 1.1. For $x \in \mathbb{N}$, $x \geq 0$, let $\pi(x)$ be the number of primes p with $p \leq x$.

So for example we'd have $\pi(2) = 1$ and $\pi(3) = 2$, and $\pi(1000) = 168$.

We'd like to get some sense of how big $\pi(x)$ is, as a function of x.

Remark 1.2. $\pi(x)$ is a step function; it only ever changes value at primes, and we can identify a prime number $n \in \mathbb{N}$ by saying that n is prime if and only if $\pi(n) = \pi(n-1) + 1$.

One immediate hurdle is that we won't be able to get a perfectly precise formula. If we instead were counting even numbers below x, we could just write that $\epsilon(x) = \lfloor \frac{x}{2} \rfloor$, and if we were counting squares below x, we could write that $\sigma(x) = \lfloor \sqrt{x} \rfloor$. But for primes, we shouldn't hope for anything of the sort.

Example 1.3. For any $a, b \in \mathbb{Q}$,

$$\pi(x) \neq \lfloor ax + b \rfloor.$$

Proof. Assume not, and we'll get a contradiction. Let $a = \frac{p}{q}$ and let $b = \frac{r}{s}$ for $p, q, r, s \in \mathbb{Z}$; assume without loss of generality that $q, s > 0$.

I claim that for any prime number n, $n + q$ is also prime. Why is this? We'll use the criterion above, that n is prime if and only if $\pi(n) = \pi(n-1) + 1$. Say n is prime. Then

$$\begin{aligned}
\pi(n+q) = \left\lfloor \frac{p}{q}(n+q) + \frac{r}{s} \right\rfloor &= \left\lfloor \frac{p}{q}n + p + \frac{r}{s} \right\rfloor \\
&= \left\lfloor \frac{p}{q}n + \frac{r}{s} \right\rfloor + p \\
&= \pi(n) + p \\
&= \pi(n-1) + 1 + p \\
&= \left\lfloor \frac{p}{q}(n-1) + \frac{r}{s} \right\rfloor + 1 + p \\
&= \left\lfloor \frac{p}{q}(n-1) + p + \frac{r}{s} \right\rfloor \\
&= \left\lfloor \frac{p}{q}(n-1+q) + \frac{r}{s} \right\rfloor + 1 \\
&= \pi(n+q-1) + 1.
\end{aligned}$$

Thus $n+q$ is also prime. But this is a problem! If q is even, then $2+q$ is even and bigger than 2, so $2+q$ can't be prime even though 2 is. But if q is odd, then $3+q$ is even and bigger than 2, so $3+q$ can't be prime even though 3 is. Here we've derived a contradiction. □

Our contradiction is getting at a more general philosophy.

Philosophy 1.4. It's hard for us to predict which numbers are prime.

You could argue that even if we can't get a perfect formula *of the form* $\lfloor ax+b \rfloor$, we could still get some different formula. Our proof above doesn't even work if a isn't rational, and I'm completely excluding lots of explicit expressions, like $\left\lfloor \frac{x}{\ln x} \right\rfloor$, or something with entirely different symbols. But the philosophy works against us no matter what: it's hard for us to predict which numbers are prime. If we had a perfectly exact formula, we'd be able to pick out primes very easily, and this shouldn't be possible.

So instead, we want to aim for a formula that isn't perfectly exact. So now our goal is to express $\pi(x)$

- in terms of functions that are easier for us to understand and compute,
- which, by our philosophy above, must have some wiggle room,
- but not so much wiggle room that it's meaningless.

To quantify this, we'll define *big-O notation* and *little-O notation*.

Definition 1.5. Let $f : \mathbb{N} \to \mathbb{R}_{\geq 0}$ and $g : \mathbb{N} \to \mathbb{R}_{\geq 0}$ be any two functions. We say that $f(n)$ *is* $O(g(n))$ and write $f(n) = O(g(n))$ if there exists a real constant $C > 0$, and some $N \in \mathbb{N}$ so that for all $n \geq N$, $|f(n)| \leq Cg(n)$.

We say that $f(n)$ *is* $o(g(n))$ if for every real $C > 0$, there exists an $N \in \mathbb{N}$ so that for all $n \geq N$, $|f(n)| \leq Cg(n)$.

Remark 1.6. We can make the same definition for functions on $\mathbb{R}$, i.e. for $f : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, if we like. We will make use of this sometimes!

The first definition is close to saying that $\lim_{n\to\infty} \frac{|f(n)|}{g(n)} < \infty$, but it's OK for us if the limit doesn't converge. The second definition is the same as saying that $\lim_{n\to\infty} \frac{|f(n)|}{g(n)} = 0$.

Proposition 1.7 (Big-O arithmetic)**.** *Let f_1, f_2, g_1, and g_2 be functions from $\mathbb{N} \to \mathbb{R}_{\geq 0}$.*

(a) Assume that $f_1(n) = O(g_1(n))$ and $f_2(n) = O(g_2(n))$. Show that $f_1 \cdot f_2(n) = O(g_1 \cdot g_2(n))$, where $f_1 \cdot f_2(n) = f_1(n) \cdot f_2(n)$. In particular, show that $f_1 \cdot f_2(n) = O(f_1 \cdot g_2(n))$.

(b) Show that $(f_1 + f_2)(n) = O(\max\{f_1, f_2\}(n))$ and that $\max\{f_1, f_2\}(n) = O((f_1 + f_2)(n))$.

(c) Assume that $f_1(n) = O(g_1(n))$ and $f_2(n) = O(g_2(n))$. Show that $(f_1 + f_2)(n) = O(\max\{g_1, g_2\}(n))$.

(d) Assume that $f_1(n) = O(g_1(n))$ and $C \in \mathbb{R}$ is any constant. Show that $C * f_1(n) = O(g_1(n))$.

(e) Assume that $\lim_{n\to\infty} f_1(n) = \infty$. Show that $f_1(n) + 1 = O(f_1(n))$.

(f) Assume that f, g are functions from $\mathbb{R}_{>0} \to \mathbb{R}_{\geq 0}$. Assume that $\lim_{x\to\infty} f(x) = \infty$, that $\lim_{x\to\infty} g(x) = \infty$, and that $f = O(g')$, where g' is the derivative of g. Show that

$$\int_1^n f(t)\mathrm{d}t = O(g(n) - g(1)) = O(g).$$

Proof. Exercise! □

Example 1.8.
- The number $\epsilon(n)$ of even numbers below n is $\lfloor \frac{n}{2} \rfloor$, which is $O(n)$ but not $o(n)$.
- The number $\sigma(n)$ of perfect squares below n is $\lfloor \sqrt{n} \rfloor$, which is $O(n)$ and $o(n)$ as well as $O(\sqrt{n})$, but it is not $o(\sqrt{n})$.
- If $f(n) = 2n^2 + 4n + 8$, then $f(n) = O(n^2)$, but not $O(n)$. In fact $f(n) = O(n^a)$ if and only if $a \geq 2$, and $f(n) = o(n^a)$ if and only if $n > 2$.
- If $f(n) = 1000000n^2 + (-1)^n$, then $f(n) = O(n^2)$.
- Returning to the number of even numbers below n; we have $\epsilon(n) = \lfloor \frac{n}{2} \rfloor$, which is great, but it'd be even nicer to have an expression without these pesky floors. Good news: $\lfloor \frac{n}{2} \rfloor$ is very close to $\frac{n}{2}$. In fact, it's either $= \frac{n}{2}$ or $= \frac{n}{2} - \frac{1}{2}$. So we get that $\epsilon(n) = \frac{n}{2} + O(1)$. This also implies the weaker statement, that $\epsilon(n) = \frac{n}{2}(1 + o(1))$.

Exercise 1.9 (Equivalence Relation # 1)**.** Throughout, let $f, g, h : \mathbb{N} \to \mathbb{R}_{\geq 0}$.

(a) Assume that $f = O(g)$ and $g = O(h)$. Show that $f = O(h)$.
(b) Show that $f = O(f)$.
(c) An *equivalence relation* $\sim$ on a set S is a subset of $S \times S$ (so for $a, b \in S$ we write $a \sim b$ if (a, b) is in the relation and $a \not\sim b$ if (a, b) is not in the relation) satisfying the following three properties.
   i. *Reflexivity.* For all $a \in S$, $a \sim a$.
   ii. *Symmetry.* For all $a, b \in S$, if $a \sim b$, then $b \sim a$.
   iii. *Transitivity.* For all $a, b, c \in S$, if $a \sim b$ and $b \sim c$, then $a \sim c$.

   Define the relation $\asymp$ on the set of functions from $\mathbb{N} \to \mathbb{R}_{\geq 0}$ by saying that $f \asymp g$ if $f = O(g)$ and $g = O(f)$. Show that $\asymp$ is an equivalence relation.

Exercise 1.10 (Equivalence Relation # 2)**.** Define a relation $\sim$ on the set of functions from $\mathbb{N} \to \mathbb{R}_{\geq 0}$ by $f \sim g$ if $f = g(1 + o(1)) = g + o(g)$. Show that $\sim$ is an equivalence relation.

Exercise 1.11. Show that if $f \sim g$, then $f \asymp g$. Give an example of f and g where $f \asymp g$, but $f \not\sim g$.

Exercise 1.12. Give examples of functions f, g from $\mathbb{N} \to \mathbb{R}_{\geq 0}$ such that $f \neq O(g)$ *and* $g \neq O(f)$.

Exercise 1.13. Order the following functions in by big-O sizing.

- $100000n$
- $e^n + 1000\sqrt{n}$
- $n^{1.01}$
- $\frac{n}{\ln n}$
- $4(\ln n)^2$
- $5(\ln \ln n)^{10000}$
- $\ln n$
- $n!$

This last example is key: we see that we can take a complicated function (in this case $\lfloor \frac{n}{2} \rfloor$), and approximate it by a simpler function, as long as we allow ourselves a big-O "fudge factor." Crucially our fudge factor is big-O smaller than our approximation, which is one way of saying that the approximation is good. In this case it's *much* smaller; not only $o(n)$ but in fact $O(1)$!

Theorem 1.14 (Prime Number Theorem)**.** *For $n \geq 2$,*

$$\pi(n) = \frac{n}{\ln n}(1 + o(1)).$$

This is the precise statement about the number of primes we were looking for. We won't prove it in this course, but we'll build up to a result that gets pretty close.

Remark 1.15. Above, we said that $\epsilon(n) = \frac{n}{2} + O(1)$. In that case we not only have a good approximation, but we know that the error is much smaller than it needs to be; it's only $O(1)$ instead of just barely being $o(n)$. In other words, $\epsilon(n) = \frac{n}{2}(1 + O(1/n))$, which is much better than multiplying by $(1 + o(1))$. Here in the Prime Number Theorem, we're not specifying how good the error is; the $(1 + o(1))$ could be basically the best possible error term. What's the truth? Well, as we'll see tomorrow, the $n/\ln n$ is itself an approximation of a more complicated function, called the *logarithmic integral*. Another way of stating the Prime Number Theorem (which we'll show tomorrow is equivalent) is

$$\pi(n) = \mathrm{li}(n)(1 + o(1)).$$

But here again, we have this $o(1)$, which we can prove, but we don't know if we can do better. In fact there's a difficult conjecture that one can do much better! It's conjectured that

$$\pi(n) = \mathrm{li}(n)(1 + O(\sqrt{n}\ln n)),$$

which is really significantly smaller. This conjecture is called the *Riemann Hypothesis*.

Let's expand on the remark above by defining $\mathrm{li}(n)$, and then actually showing that it's OK for us to go between $\mathrm{li}(n)$ and $\frac{n}{\ln n}$. This OK-ness is going to come in the form of a little-o error term.

Definition 1.16. The *logarithmic integral* is a function defined as

$$\mathrm{li}(x) = \int_2^x \frac{1}{\ln t}\mathrm{d}t.$$

Lemma 1.17.

$$\mathrm{li}(n) = \frac{n}{\ln n}(1 + o(1)).$$

Proof. Here we're going to use two main tactics: the first is integrating by parts, and the second is our newfound big-O powers. Let's first integrate by parts, where we'll say here that $u = \frac{1}{\ln t}$ and $\mathrm{d}v = \mathrm{d}t$. Then

$$\begin{aligned}\int_2^n \frac{1}{\ln t}\mathrm{d}t &= \frac{t}{\ln t}\bigg|_2^n - \int_2^n t \cdot \frac{(-1)}{(\ln t)^2} \cdot \frac{1}{t}\mathrm{d}t \\ &= \frac{n}{\ln n} - \frac{2}{\ln 2} + \int_2^n \frac{1}{(\ln t)^2}\mathrm{d}t \\ &= \frac{n}{\ln n} + \int_2^n \frac{1}{(\ln t)^2}\mathrm{d}t + o(1).\end{aligned}$$

At this point we just want to show that

$$\int_2^n \frac{1}{(\ln t)^2}\mathrm{d}t = o(\frac{n}{\ln n}).$$

This should work; we want this integral to be smaller than our original integral, and we've saved an extra power of ln in the integrand. But to be sure, let's do this out carefully.

One strategy here would be to repeat integration by parts, to get a full series expansion. We won't go that route; instead let's understand directly the error term at this step. If we were to integrate by parts, we would get a $\frac{t}{(\ln t)^2}$ term, so let's see how close that is to being an antiderivative of $\frac{1}{(\ln t)^2}$.

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{t}{(\ln t)^2}\right) &= \frac{1}{(\ln t)^2} + t\cdot\frac{(-2)}{(\ln t)^3}\cdot\frac{1}{t}\\ &= \frac{1}{(\ln t)^2} - \frac{2}{(\ln t)^3}.\end{aligned}$$

So what we get is that

$$\frac{1}{(\ln t)^2} = \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{t}{(\ln t)^2}\right) + \frac{2}{(\ln t)^3} = O\left(\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{t}{(\ln t)^2}\right)\right).$$

Thus we have

$$\begin{aligned}\int_2^n \frac{1}{(\ln t)^2}\mathrm{d}t &= \int_2^n O\left(\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{t}{(\ln t)^2}\right)\right)\mathrm{d}t\\ &= O\left(\frac{n}{(\ln n)^2}\right) = o\left(\frac{n}{\ln n}\right).\end{aligned}$$

This is exactly what we needed, so we are done! □

2. Summation by parts

We're now going to explore another key tool in our toolbox: namely, summation by parts. What we'll use specifically is sometimes called *Abel's summation formula*, but we'll call it *summation by parts*, which should be a suggestive term because it should sound to you like *integration by parts*. We just saw that $\mathrm{li}(n) = \frac{n}{\ln n}(1+o(1))$, and we just proved it using integration by parts. This tells us in particular that it's equivalent for us to prove that

$$\pi(x) = \sum_{n\leq x} \mathbb{1}_{\text{prime}}(n) = \mathrm{li}(x)(1+o(1))$$

as it is for us to prove that

$$\pi(x) = \sum_{n\leq x} \mathbb{1}_{\text{prime}}(n) = \frac{x}{\ln x}(1+o(1)),$$

so we can pick whichever is easiest for us.

Here $\mathbb{1}_{\text{prime}}(n)$ is the indicator function of the primes: it is 1 if n is prime, and 0 otherwise. This as it turns out is kind of a strange function! In particular, for reasons we'll see soon, we'd like instead to work with sums over the *von Mangoldt* function instead.

Definition 2.1. The *von Mangoldt* function is defined as

$$\Lambda(n) = \begin{cases}\ln p & \text{if } n = p^k, \text{ and } p \text{ is prime}\\ 0 & \text{otherwise.}\end{cases}$$

We write

$$\psi(x) = \sum_{n \leq x} \Lambda(n).$$

ψ is sometimes called the *Chebyshev function*, but here we will mostly just call it ψ.

So $\Lambda(2) = \ln 2, \Lambda(3) = \ln 3, \Lambda(4) = \ln 2, \Lambda(8) = \ln 2$, and $\Lambda(6) = 0$.

On the face of it, this seems like a much worse function to look at than the indicator function of the primes. We want to count primes, we don't want some weird weighted count that includes numbers like 4 and 9 that are not even prime! However, $\Lambda(n)$ has some nice properties that we'll see tomorrow, which make it in fact much more convenient than counting primes. Just as we showed before that integration by parts means that we don't care about the difference between $\mathrm{li}(n)$ and $\frac{n}{\ln n}$, we're now going to use summation by parts, to show that we won't care about the difference between $\Lambda(n)$ and $\mathbb{1}_{\text{prime}}(n)$, so we also won't care about the difference between $\sum_{n \leq x} \Lambda(n)$ and $\pi(x)$.

Let's start by clarifying what we mean by summation by parts, and proving that it works.

Theorem 2.2 (Summation by parts/Abel's summation formula)**.** *Let $a : \mathbb{N} \to \mathbb{R}$ be any function, and let*

$$A(x) = \sum_{n \leq x} a(n),$$

where $A(x) = 0$ for $x < 1$. Let $f : \mathbb{R}_{>0} \to \mathbb{R}$ be a function with continuous derivative on the interval $[x, y]$, for $0 < x < y$ integers. Then

$$\sum_{x < n \leq y} a(n) f(n) = A(y) f(y) - A(x) f(x) - \int_x^y A(t) f'(t) \mathrm{d}t.$$

Proof. The key observation here is that we've defined A and f over everything, not just integers. However, $A(x) = A(\lfloor x \rfloor)$, so we know the integers are relevant. Let $k = \lfloor x \rfloor$ and let $\ell = \lfloor y \rfloor$. Then

$$\begin{aligned}
\sum_{x < n \leq y} a(n) f(n) &= \sum_{n=k+1}^{\ell} a(n) f(n) \\
&= \sum_{n=k+1}^{\ell} (A(n) - A(n-1)) f(n) \\
&= \sum_{n=k+1}^{\ell} A(n) f(n) - \sum_{n=k}^{\ell-1} A(n) f(n+1) \\
&= \sum_{n=k+1}^{\ell-1} A(n)(f(n) - f(n-1)) + A(\ell) f(\ell) - A(k) f(k+1) \\
&= -\sum_{n=x+1}^{y-1} \int_n^{n+1} A(t) f'(t) \mathrm{d}t + A(\ell) f(\ell) - A(k) f(k+1) \\
&= -\int_{x+1}^{y} A(t) f'(t) \mathrm{d}t + A(\ell) f(\ell) - A(k) f(k+1)
\end{aligned}$$

As a last step we note that A is constant on $[x, k+1)$ and $[\ell, y]$, so that

$$\int_x^{k+1} A(t)f'(t)\mathrm{d}t = A(k)f(k+1) - A(x)f(x) \quad \text{and} \quad \int_\ell^y A(t)f'(t)\mathrm{d}t = A(y)f(y) - A(\ell)f(\ell),$$

which when we plug in gives that the sum we want is

$$\begin{aligned} &= -\int_{x+1}^y A(t)f'(t)\mathrm{d}t - \int_x^{k+1} A(t)f'(t)\mathrm{d}t - \int_\ell^y A(t)f'(t)\mathrm{d}t + A(y)f(y) - A(x)f(x) \\ &= A(y)f(y) - A(x)f(x) - \int_y^x A(t)f'(t)\mathrm{d}t, \end{aligned}$$

as desired. □

Tongue-in-cheek proof. Here's an alternate proof of summation by parts, which might seem like cheating. It's not, but we won't talk about why it's not in this course.

The function $A(x)$ is a step function that jumps $a(n)$ at each integer n. So if we want to think of the derivative of $A(x)$, it is zero almost everywhere, except at integers when it is an $a(n)$-flavored infinity. So in particular, we might feel inspired to write

$$\sum_{x<n\leq y} a(n)f(n) = \int_x^y f(t)\mathrm{d}A(t).$$

The "area under the curve" on the right is going to be exactly the sum on the left. Now let's just treat this as an integral, and we'll see how summation by parts ties into integration by parts. Integrating by parts gives us

$$\begin{aligned} \sum_{x<n\leq y} a(n)f(n) &= f(y)A(y) - f(x)A(x) - \int_x^y A(t)\mathrm{d}f(t) \\ &= f(y)A(y) - f(x)A(x) - \int_x^y A(t)f'(t)\mathrm{d}t. \end{aligned}$$

If you want to know more about why this isn't cheating, the integral we wrote above is a *Riemann-Stieltjes integral*, and it makes this logic precise. Ask me about it at TAU! □

Now let's get to our previous goal: we want to use summation by parts to show that it's OK for us to consider the sum $\psi(x) = \sum_{n\leq x} \Lambda(n)$ instead of the sum $\pi(x) = \sum_{n\leq x} \mathbb{1}_{\text{prime}}(n)$. To make this a little easier on ourselves, we're going to define an in-between function ϑ:

Definition 2.3. Let $\vartheta(x) = \sum_{p\leq x} \ln p$.

The point here is that the von Mangoldt function is strange for two reasons: first it weights everything by $\ln p$, and second it includes prime powers. This sum ϑ only does the first strange thing, so we'll understand how it relates to π first, and then get to ψ and Λ.

Lemma 2.4. *We have*

$$\vartheta(x) = \pi(x)\ln x - \int_2^x \frac{\pi(t)}{t}\mathrm{d}t$$

and

$$\pi(x) = \frac{\vartheta(x)}{\ln x} + \int_2^x \frac{\vartheta(t)}{t(\ln t)^2}\mathrm{d}t.$$

Proof. Each of these is a direct application of summation by parts. We'll prove the first one, leaving the second as an exercise. For the first one, let's take $a(n) = \mathbb{1}_{\text{prime}}(n)$ and $f(x) = \ln x$. Summation by parts tells us that

$$\vartheta(x) = \sum_{1<n\leq x} \mathbb{1}_{\text{prime}}(n) \ln n = \pi(x)\ln x - \pi(1)\ln 1 - \int_1^x \pi(t)\frac{1}{t}\mathrm{d}t,$$

which, noting that $\pi(t) = 0$ for $t < 2$ and $\ln 1 = 0$, gives us exactly what we want. □

Theorem 2.5. *The following are equivalent.*

(a) $\pi(x) = \frac{x}{\ln x}(1 + o(1))$
(b) $\vartheta(x) = x(1 + o(1))$
(c) $\psi(x) = x(1 + o(1))$

Proof. We'll do a most of these implications here, but not all of them; all that remains is that (a) implies (b), which is homework.

(b) ⇒ (a): Assume (b). By our lemma, we have

$$\begin{aligned}\pi(x) &= \frac{\vartheta(x)}{\ln x} + \int_2^x \frac{\vartheta(t)}{t(\ln t)^2}\mathrm{d}t \\ &= \frac{x}{\ln x}(1 + o(1)) + \int_2^x \frac{\vartheta(t)}{t(\ln t)^2}\mathrm{d}t.\end{aligned}$$

So let's consider this integral; note that we're assumint $\vartheta(t) = t(1 + o(1)) = O(t)$, so the integrand is $O\left(\frac{1}{(\ln t)^2}\right)$. We actually already saw in our proof of $\mathrm{li}(x) \sim \frac{x}{\ln x}$ that

$$\int_2^x O\left(\frac{1}{(\ln t)^2}\right)\mathrm{d}t = O\left(\frac{x}{(\ln x)^2}\right) = o\left(\frac{x}{\ln x}\right),$$

so $\pi(x) = \frac{x}{\ln x}(1 + o(1))$, as desired.

(c) ⟺ (b): The key idea here is that *there aren't very many prime powers that aren't primes, so they just don't matter*. Let's look carefully at $\psi(x)$. We want to split up our sum over which prime power we're looking at (primes, squares of primes, cubes of primes, etc). Doing so gives

$$\psi(x) = \sum_{p^k\leq x} \ln p = \sum_{p\leq x} \ln p + \sum_{p^2\leq x} \ln p + \sum_{p^3\leq x} \ln p + \cdots$$

But in fact, this is a bunch of slightly different copies of ϑ. The first sum over primes is just $\vartheta(x)$; the second one is $\vartheta(x^{1/2})$, since $p^2 \leq x \iff p \leq \sqrt{x}$. In all, we get

$$\psi(x) = \sum_{1\leq k\leq \log_2 x} \vartheta(x^{1/k}),$$

or

$$\psi(x) - \vartheta(x) = \sum_{2\leq k\leq \log_2 x} \vartheta(x^{1/k}).$$

Here we'll use the following bound on ϑ:

$$\vartheta(x) = \sum_{p\leq x} \ln p < \sum_{p\leq x} \ln x \leq x \ln x.$$

Note that we weren't trying very hard to get this bound! We were extremely careless, but it won't matter. Thus

$$\begin{aligned}\psi(x) - \vartheta(x) &\leq \sum_{2 \leq k \leq \log_2 x} x^{1/k} \ln x^{1/k} \\ &\leq (\log_2 x)\sqrt{x}\ln(\sqrt{x}) \\ &= \frac{\sqrt{x}(\ln x)^2}{2\ln 2} = O(\sqrt{x}(\ln x)^2).\end{aligned}$$

Thus $\psi(x) - \vartheta(x) = o(x)$; in fact it's much much smaller than x! So if $\psi(x) = x + o(x)$, then so is $\vartheta(x)$, and vice versa. □

Exercise 2.6. (a) Show that $\sum_{n=1}^{\infty} \frac{1}{n}$ diverges.

(b) For $k \geq 2$, show that $\sum_{n=2}^{\infty} \frac{1}{n^k}$ converges, and in fact that $\sum_{n=2}^{\infty} \frac{1}{n^k} = O\left(\frac{1}{2^{k-1}}\right)$.

(c) Show that $\sum_{k=2}^{\infty} \sum_{n=2}^{\infty} \frac{1}{n^k}$ converges.

Remark 2.7. This argument is tailor-made for the precision of the prime number theorem. Since we'll be proving something weaker, it's worth noting that it would apply as well for the equivalence relation $\asymp$, not just the equivalence relation $1 + o(1)$. In other words, if $\psi(x) = O(x)$ and $x = O(\psi(x))$, then $\pi(x) = O(x/\ln x)$ and $x/\ln x = O(\pi(x))$.

3. Möbius inversion

Now let's talk about why we care about $\Lambda(n)$. To do so, we're going to introduce our third very powerful tool: Möbius inversion. For anyone who feels like we've been doing nothing but calculus for the past few days, be comforted (or beware!): here lies arithmetic, and algebra.

Before we get to Möbius inversion, let's define the real star of the show: the *Möbius function* μ.

Definition 3.1. A number n is *squarefree* if it is not divisible by the square of any prime number. In other words, if $n = p_1^{e_1} \cdots p_k^{e_k}$ is the prime factorization of n, then n is squarefree if and only if $e_i = 1$ for all i.

Definition 3.2. The *Möbius function* μ is a function from $\mathbb{N} \to \{-1, 0, 1\}$ given as follows. For each integer n, we write the prime factorization of n as $n = p_1^{e_1} \cdots p_k^{e_k}$. Then

$$\mu(n) = \begin{cases} (-1)^k & \text{if } n \text{ is squarefree} \\ 0 & \text{otherwise.} \end{cases}$$

So, $\mu(n)$ is zero if n is not squarefree, and otherwise is -1 if n has an odd number of distinct prime factors, and 1 if n has an even number of distinct prime factors. For exmaple, $\mu(2) = -1$, $\mu(3) = -1$, $\mu(4) = 0$, $\mu(6) = 1$, $\mu(8) = 0$, $\mu(12) = 0$, and $\mu(30) = -1$. For $n = 1$, we will say that $\mu(1) = 1$; in this case we're really saying that 1 has 0 distinct prime factors, and $(-1)^0 = 1$.

The Möbius function is a truly magical function, with a lot of special properties. Here's one of them:

Lemma 3.3. *Let $n \in \mathbb{N}_{\geq 1}$. Then*

$$\sum_{d|n} \mu(d) = \begin{cases} 1 & \text{if } n = 1 \\ 0 & \text{if } n > 1. \end{cases}$$

Proof. Assume $n \neq 1$; then $n = p_1^{e_1} \cdots p_k^{e_k}$, with $k \geq 1$. Let m be the "squarefree portion of n," so $m = p_1 \cdots p_k$. Note that if $d|n$ but $d|m$, then for some i, $p_i^2|d$. Thus d is not squarefree, so $\mu(d) = 0$. This gives

$$\sum_{d|n} \mu(d) = \sum_{d|m} \mu(d).$$

Note $n = 1 \iff m = 1$, so at this point it suffices for us to show that if m is squarefree, the lemma holds.

Now the divisors of m are each given by a subset of $\{p_1, \ldots, p_k\}$; there are $\binom{k}{j}$ divisors d with j prime factors, and for each of those $\mu(d) = (-1)^j$. Thus

$$\sum_{d|m} \mu(d) = \sum_{j=0}^{k} (-1)^j \binom{k}{j} = (1-1)^j = 0,$$

by the binomial theorem. Another way to see this is that

$$\begin{aligned} \sum_{d|m} \mu(d) &= \sum_{\substack{d|m \\ p_1|d}} \mu(d) + \sum_{\substack{d|m \\ p_1 \nmid d}} \mu(d) \\ &= \sum_{d|(m/p_1)} \mu(p_1 \cdot d) + \mu(d) = 0. \end{aligned}$$

On the other hand, if $n = 1$, then $\sum_{d|n} \mu(d) = \mu(1) = 1$. □

So this is very handy; the Möbius function gives us a nice way to understand the indicator function of 1. This leads to Möbius inversion, which is even handier. In this case we want to consider two functions f, g from $\mathbb{N} \to \mathbb{R}$ (or even $\mathbb{C}$), with the property that for all n,

$$f(n) = \sum_{d|n} g(d).$$

The question is then, how can we express $g(n)$ in terms of values of f? Let's first look at a few examples.

Example 3.4. Say p is a prime. Then $f(p) = \sum_{d|p} g(d) = g(1) + g(d)$. How do we solve for $g(p)$? Well, $g(p) = f(p) - g(1)$, and $f(1) = g(1)$, so $g(d) = f(p) - f(1)$.

Example 3.5. Now say p and q are two primes. How do we write $g(pq)$ in terms of f? We have $f(pq) = \sum_{d|pq} g(d) = g(pq) + g(p) + g(q) + g(1)$, but also from above that $g(p) = f(p) - f(1)$, and the same for q, as well as $g(1) = f(1)$. Then

$$\begin{aligned} g(pq) &= f(pq) - g(p) - g(q) - g(1) \\ &= f(pq) - f(p) + f(1) - f(q) + f(1) - f(1) \\ &= f(pq) - f(p) - f(q) + f(1). \end{aligned}$$

What do these examples have in common? Well, it's an alternating sum–and the alternation is based on the number of prime factors! I smell a Möbius function.

Proposition 3.6 (Möbius inversion)**.** *Let f, g be two functions from $\mathbb{N} \to \mathbb{R}$, and suppose that for all $n \in \mathbb{N}$,*

$$f(n) = \sum_{d|n} g(d).$$

Then for all $n \in \mathbb{N}$,

$$g(n) = \sum_{d|n} f(d)\mu(n/d).$$

Proof. Let's start with the right-hand side of the formula, and see if we can simplify to $g(n)$. We start by plugging in our formula for $f(n)$:

$$\begin{aligned}\sum_{d|n} f(d)\mu(n/d) &= \sum_{d|n}\sum_{k|d} g(k)\mu(n/d) \\ &= \sum_{k|n} g(k) \sum_{k|d|n} \mu(n/d) \\ &= \sum_{k|n} g(k) \sum_{\ell|n/k} \mu(n/(k\ell)) \\ &= \sum_{k|n} g(k) \sum_{\ell|n/k} \mu(\ell).\end{aligned}$$

Here we've noted that we had a sum over all divisors of n/k of $\mu((n/k)/\ell)$; since we're summing over everything, we can swap ℓ with $(n/k)/\ell$ in the sum to get the last step. Now we apply our Lemma, to get that the only nonzero term is when $n/k = 1$, or when $k = n$. But now we're left with nothing but $g(n)$. □

Remark 3.7. Note that by swapping the roles of d and n/d, we can also write

$$g(n) = \sum_{d|n} \mu(d) f(n/d).$$

Remark 3.8. In fact this also goes the other way: if we know that $g(n) = \sum_{d|n} f(d)\mu(n/d)$, then we also know that $f(n) = \sum_{d|n} g(d)$. The proof is an exercise!

Example 3.9. Here's a silly, but important, example of Möbius inversion. Let $f(n)$ be 1 if $n = 1$ and 0 otherwise. By our lemma, we have

$$f(n) = \sum_{d|n} \mu(d),$$

so let's apply Möbius inversion with $g = \mu$. This gives that

$$\mu(n) = \sum_{d|n} f(d)\mu(n/d) = f(1)\mu(n/1) = \mu(n),$$

which sure is good!

Another way to phrase this is that in proving the lemma, we were really using reverse Möbius inversion, for this silly example.

Let's see a medium application of Möbius inversion, before the big one.

Definition 3.10. The *totient function* $\phi(n)$ is the number of $k \leq n$ with $\gcd(k, n) = 1$.

For example, $\phi(1) = 1$, $\phi(2) = 1$, $\phi(3) = 2$, $\phi(5) = 4$, $\phi(6) = 2$, $\phi(7) = 6$, $\phi(9) = 6$.

We'll start with the following fact about ϕ:

Fact 3.11. If $n \geq 1$,

$$\sum_{d|n} \phi(d) = n.$$

Proof. Let's consider the set $S = \{1, 2, \ldots, n\}$. We're going to split up S as a bunch of disjoint sets as follows. For $d|n$, let

$$A(d) = \{k : \gcd(k, n) = d, 1 \leq k \leq n\}.$$

Then $S = \bigsqcup_{d|n} A(d)$, so

$$n = \sum_{d|n} |A(d)|.$$

What's the size of $A(d)$? Note that $\gcd(k, n) = d \iff \gcd(k/d, n/d) = 1$, and $1 \leq k \leq n \iff 1 \leq k/d \leq n/d$. So $A(d)$ is in bijection with the set of integers q such that $1 \leq q \leq n/d$, $(q, n/d) = 1$. There are precisely $\phi(n/d)$ such q, so

$$n = \sum_{d|n} \phi(n/d) = \sum_{d|n} \phi(d).$$

□

Möbius inversion then transforms this fact into the fact that

Fact 3.12. If $n \geq 1$,

$$\phi(n) = \sum_{d|n} \mu(d) \frac{n}{d}.$$

Now we'll prove the following.

Theorem 3.13. *For $x > 1$, we have*

$$\sum_{n \leq x} \phi(n) = \frac{3}{\pi^2} x^2 + O(x \ln x).$$

If the π^2 came out of nowhere, that's because we're relying on the following black box, which is a deep fact about the Riemann zeta, and which we won't prove this week.

Black Box 3.14.

$$\sum_{n=1}^{\infty} \frac{\mu(n)}{n^2} = \frac{6}{\pi^2}.$$

Using this black box, let's prove the theorem!

Proof. We're going to make use of our Möbius inversion formula. Plugging it in yields

$$\begin{aligned}\sum_{n\leq x}\phi(n) &= \sum_{n\leq x}\sum_{d|n}\mu(d)\frac{n}{d}\\ &= \sum_{d\leq x}\sum_{q\leq x/d}\mu(d)q, \text{ where } q = n/d\\ &= \sum_{d\leq x}\mu(d)\left(\frac{1}{2}\left(\frac{x}{d}\right)^2 + O\left(\frac{x}{d}\right)\right)\\ &= \frac{1}{2}x^2\sum_{d\leq x}\frac{\mu(d)}{d^2} + O\left(x\sum_{d\leq x}\frac{1}{d}\right)\end{aligned}$$

Note that, for example by the integral test, $\sum_{d=1}^{\infty}\frac{\mu(d)}{d^2}$ converges absolutely. In particular, (also by the integral test), we can bound $\sum_{d\geq x}\frac{1}{d^2} = O\left(\frac{1}{d}\right)$. Thus

$$\sum_{d\leq x}\frac{\mu(d)}{d^2} = \sum_{d=1}^{\infty}\frac{\mu(d)}{d^2} + O\left(\frac{1}{d}\right) = \frac{6}{\pi^2} + O\left(\frac{1}{d}\right),$$

using our black box. Meanwhile, integrating can also give us the bound that

$$\sum_{d\leq x}\frac{1}{d} = O(\ln x).$$

Putting everything together yields

$$\sum_{n\leq x}\phi(n) = \frac{1}{2}x^2\left(\frac{6}{\pi^2} + O\left(\frac{1}{x}\right)\right) + O(x\ln x) = \frac{3}{\pi^2} + O(x\ln x).$$

□

Let's come back to the von Mangoldt function, and see what Möbius inversion gives us.

Fact 3.15. For $n \geq 1$,

$$\ln n = \sum_{d|n}\Lambda(d).$$

Proof. First note that for $n = 1$, we have $\ln 1 = 0$, and the same is true for $\Lambda(1)$, as desired. Now let $n \geq 2$. Let $n = p_1^{e_1}\cdots p_k^{e_k}$ be the prime factorization of n. The divisors of n with $\Lambda(d) \neq 0$ are precisely the ones that are a prime power, so $p_1, p_1^2, \ldots, p_1^{e_1}, p_2, \ldots$. Thus

$$\begin{aligned}\sum_{d|n}\Lambda(d) &= \sum_{i=1}^{k}\left(\Lambda(p_i) + \Lambda(p_i^2) + \cdots + \Lambda(p_i^k)\right)\\ &= \sum_{i=1}^{k} e_i\ln p_i = \sum_{i=1}^{k}\ln p_i^{e_i}\\ &= \ln\prod_{i=1}^{k}p_i^{e_i} = \ln n.\end{aligned}$$

□

This is a great fact, because it's perfect for Möbius inversion. Möbius inversion gives us:

Important Formula 3.16. For $n \geq 1$,

$$\Lambda(n) = \sum_{d|n} \mu(d) \ln \frac{n}{d}.$$

4. Grand Finale: Chebyshev's theorem

Theorem 4.1 (Chebyshev)**.** *$\psi(x) \asymp x$. That is to say, $\psi(x) = O(x)$, and $x = O(\psi(x))$.*

Note that this is weaker than the PNT! The PNT says that $\psi(x) \sim x$, but we're showing something weaker here: just that $\psi(x) \asymp x$. (See Exercise 1.11)

Let's prove this. The proof will be the whole section, so it's not isolated in a proof box. We'll start with our important formula. Adding it up over $n \leq x$ gives that

$$\psi(x) = \sum_{n \leq x} \sum_{d|n} \mu(d) \ln \frac{n}{d} = \sum_{d \leq x} \mu(d) \sum_{m \leq x/d} \ln m,$$

where here we've written $n = md$ while swapping summations. Let $T(x) = \sum_{n \leq x} \ln n$, so that $\psi(x) = \sum_{d \leq x} \mu(d) T(x/d)$. The benefit of T is that it's a nice summatory function that we can understand via the integral test:

$$\int_1^N \ln u \mathrm{d}u \leq T(N) \leq \int_1^{N+!} \ln u \mathrm{d}u$$

for any $N \in \mathbb{N}$. This is an integral we can really understand:

$$\int \ln u \mathrm{d}u = u \ln u - u.$$

Thus

$$N \ln N - N + 1 \leq T(N) \leq (N+1) \ln(N+1) - N - 1 + 1,$$

so in particular on expanding both sides we get

$$T(x) = x \ln x - x + O(\ln 2x).$$

Here we're playing fast and loose with the difference between N and x, which is allowed because $\ln u$ is monotonically increasing. This is a great estimate, but what happens when we plug it in? We get

$$\begin{aligned}
\psi(x) &= \sum_{d \leq x} \mu(d) \left(\frac{x}{d} \ln \frac{x}{d} - xd + O\left(\ln 2\frac{x}{d} \right) \right) \\
&= (x \ln x - x) \sum_{d \leq x} \frac{\mu(d)}{d} - x \sum_{d \leq x} \frac{\mu(d) \ln d}{d} + O\left(\sum_{d \leq x} \mu(d) \ln 2\frac{x}{d} \right).
\end{aligned}$$

Unfortunately our amazing integral approximation hasn't gotten us all the way there, since we still need to get a handle on these sums $\sum_{d \leq x} \frac{\mu(d)}{d}$ and $\sum_{d \leq x} \frac{\mu(d) \ln d}{d}$.

To do this, we're going to need one last handy yet fundamental idea. We can't understand these sums with $\mu(d)$, so we're going to replace $\mu(d)$ by a different function $a(d)$, so that

- a is close enough to μ that we don't ruin our approximation, but
- a is easier to sum than μ.

In order to make a close to μ, let's make a something like a truncation of μ; like μ, but eventually a is just 0 instead. Now, the problem is that at the moment, we have too many options for a; we don't know how to pick a good one. So let's think carefully about what we want a to do and what that means, and gather up some constraints for ourselves, until we aren't left with so many options.

Let $\mathcal{D}$ be a finite set of numbers, and let's assume that $a(d) = 0$ for all $d \notin \mathcal{D}$. Then

$$\sum_{d\in\mathcal{D}} a(d)T(x/d) = (x\ln x - x)\sum_{d\le x}\frac{a(d)}{d} - x\sum_{d\le x}\frac{a(d)\ln d}{d} + O\left(\sum_{d\le x} a(d)\ln 2\frac{x}{d}\right). \tag{4.1}$$

Now looking at this gives us the first constraint that we'll really want to assume: let's assume that

$$\sum_{d\in\mathcal{D}}\frac{a(d)}{d} = 0.$$

Note that we can guess that this is reasonably close to true for the Möbius function itself: if not, we'd have an $x\ln x$ term, when we want the biggest term to be about x. We then will hope that

$$-\sum_{d\in\mathcal{D}}\frac{a(d)\ln d}{d} \approx 1.$$

Meanwhile we also want $\sum_{d\in\mathcal{D}} a(d)T(x/d)$ to be reasonably close to $\psi(x)$. By how we defined $T(x)$, we have

$$\begin{aligned}\sum_{d\in\mathcal{D}} a(d)T(x/d) &= \sum_{dn\le x} a(d)\ln n\\ &= \sum_{dn\le x} a(d)\sum_{k|n}\Lambda(k)\\ &= \sum_{dkm\le x} a(d)\Lambda(k)\\ &= \sum_{k\le x}\Lambda(k)\left(\sum_{dm\le x/k} a(d)\right).\end{aligned}$$

Again we'll define the thing in parentheses as its own function: namely we'll define

$$E(y) = \sum_{dm\le y} a(d) = \sum_{d\le y} a(d)\left\lfloor\frac{y}{d}\right\rfloor,$$

so that above we have $E(x/k)$. Our whole expression above is going to be close to $\psi(x)$ if $E(y)$ is close to 1. Comparing to $\mu(d)$, note that this plays out quite nicely: if $y \ge 1$, then

$$\sum_{d\le y}\mu(d)\lfloor y/d\rfloor = \sum_{dk\le y}\mu(d) = \sum_{n\le y}\sum_{d|n}\mu(d) = 1.$$

Returning to E itself, note that if $a(d) = 0$ whenever $d \notin \mathcal{D}$, then

$$\begin{aligned} E(y) &= -\sum_{d\in\mathcal{D}} a(d)\left\lfloor \frac{y}{d}\right\rfloor \\ &= -y\sum_{d\in\mathcal{D}}\frac{a(d)}{d} - \sum_{d\in\mathcal{D}} a(d)\left(\frac{y}{d} - \left\lfloor\frac{y}{d}\right\rfloor\right) \\ &= -\sum_{d\in\mathcal{D}} a(d)\left(\frac{y}{d} - \left\lfloor\frac{y}{d}\right\rfloor\right), \end{aligned}$$

since we had the assumption that $\sum_{d\in\mathcal{D}} a(d)/d = 0$. But note that $\left(\frac{y}{d} - \left\lfloor\frac{y}{d}\right\rfloor\right)$ repeats as y varies, and in particular it repeats every $\operatorname{lcm}_{d\in\mathcal{D}} d$. So we can understand $E(y)$ by doing a finite calculation from a given $a(d)$.

Now finally let's take a specific choice of $a(d)$, and see how "reasonably close to 1" we get for $E(y)$. As a simple choice, let's let

$$a(d) = \begin{cases} 1 & \text{if } d = 1 \\ -2 & \text{if } d = 2 \\ 0 & \text{if } d > 2 \end{cases}.$$

Then

$$\sum_{d\in\mathcal{D}}\frac{a(d)}{d} = \frac{1}{1} + \frac{(-2)}{2} = 1 - 1 = 0,$$

so we've satisfied that constraint. Now let's look at what happens with $E(y)$, and in turn what happens with ψ. For $0 \le y < 1$,

$$E(y) = -\left(1\cdot y + (-2)\cdot\frac{y}{2}\right) = 0,$$

and for $1 \le y < 2$,

$$E(y) = -\left(1\cdot(y-1) + (-2)\cdot\frac{y}{2}\right) = 1.$$

So we get that

$$E(y) = \begin{cases} 0 & \text{if } 0 \le y < 1 \\ 1 & \text{if } 1 \le y < 2 \end{cases},$$

and then $E(y)$ repeats at $y = 2$. Thus for our choice of $E(y)$, we have

$$\psi(x) - \psi(x/2) = \sum_{x/2<k\le x}\Lambda(k) \le \sum_{k\le x}\Lambda(k)E(x/k) \le \sum_{k\le x}\Lambda(k) = \psi(x),$$

so returning to our formula 4.1, we get

$$\begin{aligned}\sum_{k\le x}\Lambda(k)E(x/k) &= (x\ln x - x)\sum_{d\le x}\frac{a(d)}{d} - x\sum_{d\le x}\frac{a(d)\ln d}{d} + O\left(\sum_{d\le x}a(d)\ln 2\frac{x}{d}\right)\\ \Rightarrow \psi(x)-\psi(x/2) &= 0 - x\sum_{d\le x}\frac{a(d)\ln d}{d} + O\left(\sum_{d\le x}a(d)\ln\left(2\frac{x}{d}\right)\right)\\ &= -x\frac{\ln 1}{1} - x\sum_{d\le x}\frac{(-2)\ln 2}{2} + O\left(\ln 2x - 2\ln x\right)\\ &= x\ln 2 + O(\ln x).\end{aligned}$$

Thus

$$\psi(x)-\psi(x/2)\le(\ln 2)x + O(\ln x)\le\psi(x).$$

The second inequality here shows that $x = O(\psi(x)$, since $x \le \frac{1}{\ln 2}\psi(x) + O(\ln x)$, so we're halfway there! Now note that we can plug in whatever we want for x! Let's plug in $x/2^r$. The first inequality gives

$$\psi(x/2^r)-\psi(x/2^{r+1})\le(\ln 2)\frac{x}{2^r} + O(\ln(x/2^r)).$$

Summing over r gives

$$\begin{aligned}\sum_{r=1}^{\log_2 x}\psi(x/2^r)-\psi(x/2^{r+1}) &\le \sum_{r=1}^{\log_2 x}(\ln 2)\frac{x}{2^r} + O(\ln(x/2^r))\\ \Rightarrow\psi(x) &\le (\ln 2)x\sum_{r=1}^{\log_2 x}\frac{1}{2^r} + \sum_{r=1}^{\log_2 x}O(\ln(x/2^r))\\ &\le 2(\ln 2)x + O((\ln x)^2).\end{aligned}$$

Thus $\psi(x) = O(x)$, as desired.

In numbers, we've shown that eventually,

$$0.693x\le\psi(x)\le 1.3863x.$$

Exercise 4.2. By using

$$a(d) = \begin{cases}1 & \text{if } d = 1,30\\ -1 & \text{if } d = 2,3,5\\ 0 & \text{otherwise,}\end{cases}$$

show that

$$(0.9212)x + O(\ln x)\le\psi(x)\le(1.1056)x + O((\ln x)^2).$$