5
The distribution of primes

This chapter concerns itself with the question: how many primes are there? In Chapter 1, we proved that there are infinitely many primes; however, we are interested in a more quantitative answer to this question; that is, we want to know how "dense" the prime numbers are.

This chapter has a bit more of an "analytical" flavor than other chapters in this text. However, we shall not make use of any mathematics beyond that of elementary calculus.

5.1 Chebyshev's theorem on the density of primes

The natural way of measuring the density of primes is to count the number of primes up to a bound x, where x is a real number. For a real number $x \geq 0$, the function $\pi(x)$ is defined to be the number of primes up to x. Thus, $\pi(1) = 0$, $\pi(2) = 1$, $\pi(7.5) = 4$, and so on. The function π is an example of a "step function," that is, a function that changes values only at a discrete set of points. It might seem more natural to define π only on the integers, but it is the tradition to define it over the real numbers (and there are some technical benefits in doing so).

Let us first take a look at some values of $\pi(x)$. Table 5.1 shows values of $\pi(x)$ for $x = 10^{3i}$ and $i = 1, \ldots, 6$. The third column of this table shows the value of $x/\pi(x)$ (to five decimal places). One can see that the differences between successive rows of this third column are roughly the same—about 6.9—which suggests that the function $x/\pi(x)$ grows logarithmically in x. Indeed, as $\log(10^3) \approx 6.9$, it would not be unreasonable to guess that $x/\pi(x) \approx \log x$, or equivalently, $\pi(x) \approx x/\log x$.

The following theorem is a first—and important—step towards making the above guesswork more rigorous:

Table 5.1. *Some values of* $\pi(x)$

x	$\pi(x)$	$x/\pi(x)$
10^3	168	5.95238
10^6	78498	12.73918
10^9	50847534	19.66664
10^{12}	37607912018	26.59015
10^{15}	29844570422669	33.50693
10^{18}	24739954287740860	40.42045

Theorem 5.1 (Chebyshev's theorem). *We have*

$$\pi(x) = \Theta(x/\log x).$$

It is not too difficult to prove this theorem, which we now proceed to do in several steps. Recalling that $\nu_p(n)$ denotes the power to which a prime p divides an integer n, we begin with the following observation:

Theorem 5.2. *Let n be a positive integer. For any prime p, we have*

$$\nu_p(n!) = \sum_{k\geq 1} \lfloor n/p^k \rfloor.$$

Proof. This follows immediately from the observation that the numbers $1, 2, \ldots, n$ include exactly $\lfloor n/p \rfloor$ multiplies of p, $\lfloor n/p^2 \rfloor$ multiplies of p^2, and so on (see Exercise 1.5). □

The following theorem gives a lower bound on $\pi(x)$.

Theorem 5.3. $\pi(n) \geq \frac{1}{2}(\log 2)n/\log n$ *for all integers* $n \geq 2$.

Proof. For positive integer m, consider the binomial coefficient

$$N := \binom{2m}{m} = \frac{(2m)!}{(m!)^2}.$$

Note that

$$N = \left(\frac{m+1}{1}\right)\left(\frac{m+2}{2}\right)\cdots\left(\frac{m+m}{m}\right),$$

from which it is clear that $N \geq 2^m$ and that N is divisible only by primes p not exceeding $2m$. Applying Theorem 5.2 to the identity $N = (2m)!/(m!)^2$, we have

$$\nu_p(N) = \sum_{k\geq 1} (\lfloor 2m/p^k \rfloor - 2\lfloor m/p^k \rfloor).$$

Each term in this sum is either 0 or 1 (see Exercise 1.4), and for $k > \log(2m)/\log p$, each term is zero. Thus, $\nu_p(N) \leq \log(2m)/\log p$.

So we have

$$\begin{aligned}\pi(2m)\log(2m) &= \sum_{p\leq 2m} \frac{\log(2m)}{\log p}\log p\\ &\geq \sum_{p\leq 2m} \nu_p(N)\log p = \log N \geq m\log 2,\end{aligned}$$

where the summations are over the primes p up to $2m$. Therefore,

$$\pi(2m) \geq \tfrac{1}{2}(\log 2)(2m)/\log(2m).$$

That proves the theorem for even n. Now consider odd $n \geq 3$, so $n = 2m-1$ for $m \geq 2$. Since the function $x/\log x$ is increasing for $x \geq 3$ (verify), and since $\pi(2m-1) = \pi(2m)$ for $m \geq 2$, we have

$$\begin{aligned}\pi(2m-1) &= \pi(2m)\\ &\geq \tfrac{1}{2}(\log 2)(2m)/\log(2m)\\ &\geq \tfrac{1}{2}(\log 2)(2m-1)/\log(2m-1).\end{aligned}$$

That proves the theorem for odd n. □

As a consequence of the above theorem, we have $\pi(x) = \Omega(x/\log x)$ for real $x \to \infty$. Indeed, for real $x \geq 2$, setting $c := \frac{1}{2}(\log 2)$, we have

$$\pi(x) = \pi(\lfloor x\rfloor) \geq c\lfloor x\rfloor/\log\lfloor x\rfloor \geq c(x-1)/\log x = \Omega(x/\log x).$$

To obtain a corresponding upper bound for $\pi(x)$, we introduce an auxiliary function, called Chebyshev's theta function:

$$\vartheta(x) := \sum_{p\leq x}\log p,$$

where the sum is over all primes p up to x.

Chebyshev's theta function is an example of a summation over primes, and in this chapter, we will be considering a number of functions that are defined in terms of sums or products over primes. To avoid excessive tedium, we adopt the usual convention used by number theorists: if not explicitly stated, summations and products over the variable p are always understood to be over primes. For example, we may write $\pi(x) = \sum_{p\leq x} 1$.

The next theorem relates $\pi(x)$ and $\vartheta(x)$. Recall the "$\sim$" notation from §3.1: for two functions f and g such that $f(x)$ and $g(x)$ are positive for all sufficiently large x, we write $f \sim g$ to mean that $\lim_{x\to\infty} f(x)/g(x) = 1$, or equivalently, for all $\epsilon > 0$ there exists x_0 such that $(1-\epsilon)g(x) < f(x) < (1+\epsilon)g(x)$ for all $x > x_0$.

Theorem 5.4. *We have*

$$\pi(x) \sim \frac{\vartheta(x)}{\log x}.$$

Proof. On the one hand, we have

$$\vartheta(x) = \sum_{p\le x} \log p \le \log x \sum_{p \le x} 1 = \pi(x) \log x.$$

So we have

$$\pi(x) \ge \frac{\vartheta(x)}{\log x}.$$

On the other hand, for every $x > 1$ and δ with $0 < \delta < 1$, we have

$$\begin{aligned} \vartheta(x) &\ge \sum_{x^\delta < p \le x} \log p \\ &\ge \delta \log x \sum_{x^\delta < p \le x} 1 \\ &= \delta \log x \, (\pi(x) - \pi(x^\delta)) \\ &\ge \delta \log x \, (\pi(x) - x^\delta). \end{aligned}$$

Hence,

$$\pi(x) \le x^\delta + \frac{\vartheta(x)}{\delta \log x}.$$

Since by the previous theorem, the term x^δ is $o(\pi(x))$, we have for all sufficiently large x (depending on δ), $x^\delta \le (1-\delta)\pi(x)$, and so

$$\pi(x) \le \frac{\vartheta(x)}{\delta^2 \log x}.$$

Now, for any $\epsilon > 0$, we can choose δ sufficiently close to 1 so that $1/\delta^2 < 1 + \epsilon$, and for this δ, and for all sufficiently large x, we have $\pi(x) < (1+\epsilon)\vartheta(x)/\log x$, and the theorem follows. □

Theorem 5.5. $\vartheta(x) < 2x \log 2$ *for all real numbers* $x \ge 1$.

Proof. It suffices to prove that $\vartheta(n) < 2n \log 2$ for integers $n \ge 1$, since then $\vartheta(x) = \vartheta(\lfloor x \rfloor) < 2 \lfloor x \rfloor \log 2 \le 2x \log 2$.

For positive integer m, consider the binomial coefficient

$$M := \binom{2m+1}{m} = \frac{(2m+1)!}{m!(m+1)!}.$$

One sees that M is divisible by all primes p with $m+1 < p \leq 2m+1$. As M occurs twice in the binomial expansion of $(1+1)^{2m+1}$, one sees that $M < 2^{2m+1}/2 = 2^{2m}$. It follows that

$$\vartheta(2m+1) - \vartheta(m+1) = \sum_{m+1<p\leq 2m+1} \log p \leq \log M < 2m \log 2.$$

We now prove the theorem by induction. For $n = 1$ and $n = 2$, the theorem is trivial. Now let $n > 2$. If n is even, then we have

$$\vartheta(n) = \vartheta(n-1) < 2(n-1)\log 2 < 2n \log 2.$$

If $n = 2m+1$ is odd, then we have

$$\begin{aligned}\vartheta(n) &= \vartheta(2m+1) - \vartheta(m+1) + \vartheta(m+1)\\ &< 2m\log 2 + 2(m+1)\log 2 = 2n\log 2. \quad \square\end{aligned}$$

Another way of stating the above theorem is:

$$\prod_{p\leq x} p < 4^x.$$

Theorem 5.1 follows immediately from Theorems 5.3, 5.4 and 5.5. Note that we have also proved:

Theorem 5.6. *We have*

$$\vartheta(x) = \Theta(x).$$

EXERCISE 5.1. If p_n denotes the nth prime, show that $p_n = \Theta(n \log n)$.

EXERCISE 5.2. For integer $n > 1$, let $\omega(n)$ denote the number of distinct primes dividing n. Show that $\omega(n) = O(\log n / \log\log n)$.

EXERCISE 5.3. Show that for positive integers a and b,

$$\binom{a+b}{b} \geq 2^{\min(a,b)}.$$

5.2 Bertrand's postulate

Suppose we want to know how many primes there are of a given bit length, or more generally, how many primes there are between m and $2m$ for a given integer m. Neither the statement, nor the proof, of Chebyshev's theorem imply that there are *any* primes between m and $2m$, let alone a useful density estimate of such primes.

Bertrand's postulate is the assertion that for all positive integers m, there exists a prime between m and $2m$. We shall in fact prove a stronger result, namely, that not only is there one prime, but the number of primes between m and $2m$ is $\Omega(m/\log m)$.

Theorem 5.7 (Bertrand's postulate). *For any positive integer m, we have*

$$\pi(2m) - \pi(m) > \frac{m}{3\log(2m)}.$$

The proof uses Theorem 5.5, along with a more careful re-working of the proof of Theorem 5.3. The theorem is clearly true for $m \leq 2$, so we may assume that $m \geq 3$. As in the proof of the Theorem 5.3, define $N := \binom{2m}{m}$, and recall that N is divisible only by primes strictly less than $2m$, and that we have the identity

$$\nu_p(N) = \sum_{k\geq 1} (\lfloor 2m/p^k \rfloor - 2\lfloor m/p^k \rfloor), \tag{5.1}$$

where each term in the sum is either 0 or 1. We can characterize the values $\nu_p(N)$ a bit more precisely, as follows:

Lemma 5.8. *Let $m \geq 3$ and $N = \binom{2m}{m}$ as above. For all primes p, we have*

$$p^{\nu_p(N)} \leq 2m; \tag{5.2}$$

$$\text{if } p > \sqrt{2m}, \text{ then } \nu_p(N) \leq 1; \tag{5.3}$$

$$\text{if } 2m/3 < p \leq m, \text{ then } \nu_p(N) = 0; \tag{5.4}$$

$$\text{if } m < p < 2m, \text{ then } \nu_p(N) = 1. \tag{5.5}$$

Proof. For (5.2), all terms with $k > \log(2m)/\log p$ in (5.1) vanish, and hence $\nu_p(N) \leq \log(2m)/\log p$, from which it follows that $p^{\nu_p(N)} \leq 2m$.

(5.3) follows immediately from (5.2).

For (5.4), if $2m/3 < p \leq m$, then $2m/p < 3$, and we must also have $p \geq 3$, since $p = 2$ implies $m < 3$. We have $p^2 > p(2m/3) = 2m(p/3) \geq 2m$, and hence all terms with $k > 1$ in (5.1) vanish. The term with $k = 1$ also vanishes, since $1 \leq m/p < 3/2$, from which it follows that $2 \leq 2m/p < 3$, and hence $\lfloor m/p \rfloor = 1$ and $\lfloor 2m/p \rfloor = 2$.

For (5.5), if $m < p < 2m$, it follows that $1 < 2m/p < 2$, so $\lfloor 2m/p \rfloor = 1$. Also, $m/p < 1$, so $\lfloor m/p \rfloor = 0$. It follows that the term with $k = 1$ in (5.1) is 1, and it is clear that $2m/p^k < 1$ for all $k > 1$, and so all the other terms vanish. □

We need one more technical fact, namely, a somewhat better lower bound on N than that used in the proof of Theorem 5.3:

Lemma 5.9. *Let $m \geq 3$ and $N = \binom{2m}{m}$ as above. We have*

$$N > 4^m/(2m). \tag{5.6}$$

Proof. We prove this for all $m \geq 3$ by induction on m. One checks by direct calculation that it holds for $m = 3$. For $m > 3$, by induction we have

$$\begin{aligned}\binom{2m}{m} &= 2\frac{2m-1}{m}\binom{2(m-1)}{m-1} > \frac{(2m-1)4^{m-1}}{m(m-1)} \\ &= \frac{2m-1}{2(m-1)}\frac{4^m}{2m} > \frac{4^m}{2m}. \quad \square\end{aligned}$$

We now have the necessary technical ingredients to prove Theorem 5.7. Define

$$P_m := \prod_{m<p<2m} p,$$

and define Q_m so that

$$N = Q_m P_m.$$

By (5.4) and (5.5), we see that

$$Q_m = \prod_{p \leq 2m/3} p^{\nu_p(N)}.$$

Moreover, by (5.3), $\nu_p(N) > 1$ for at most those $p \leq \sqrt{2m}$, so there are at most $\sqrt{2m}$ such primes, and by (5.2), the contribution of each such prime to the above product is at most $2m$. Combining this with Theorem 5.5, we obtain

$$Q_m < (2m)^{\sqrt{2m}} \cdot 4^{2m/3}.$$

We now apply (5.6), obtaining

$$P_m = NQ_m^{-1} > 4^m(2m)^{-1}Q_m^{-1} > 4^{m/3}(2m)^{-(1+\sqrt{2m})}.$$

It follows that

$$\begin{aligned}\pi(2m) - \pi(m) \geq \log P_m/\log(2m) &> \frac{m\log 4}{3\log(2m)} - (1+\sqrt{2m}) \\ &= \frac{m}{3\log(2m)} + \frac{m(\log 4 - 1)}{3\log(2m)} - (1+\sqrt{2m}). \end{aligned} \tag{5.7}$$

Clearly, the term $(m(\log 4 - 1))/(3\log(2m))$ in (5.7) dominates the term $1 + \sqrt{2m}$, and so Theorem 5.7 holds for all sufficiently large m. Indeed, a simple calculation shows that (5.7) implies the theorem for $m \geq 13{,}000$, and one can verify by brute force (with the aid of a computer) that the theorem holds for $m < 13{,}000$.

5.3 Mertens' theorem

Our next goal is to prove the following theorem, which turns out to have a number of applications.

Theorem 5.10. *We have*

$$\sum_{p \leq x} \frac{1}{p} = \log\log x + O(1).$$

The proof of this theorem, while not difficult, is a bit technical, and we proceed in several steps.

Theorem 5.11. *We have*

$$\sum_{p \leq x} \frac{\log p}{p} = \log x + O(1).$$

Proof. Let $n := \lfloor x \rfloor$. By Theorem 5.2, we have

$$\log(n!) = \sum_{p \leq n} \sum_{k \geq 1} \lfloor n/p^k \rfloor \log p = \sum_{p \leq n} \lfloor n/p \rfloor \log p + \sum_{k \geq 2} \sum_{p \leq n} \lfloor n/p^k \rfloor \log p.$$

We next show that the last sum is $O(n)$. We have

$$\begin{aligned} \sum_{p \leq n} \log p \sum_{k \geq 2} \lfloor n/p^k \rfloor &\leq n \sum_{p \leq n} \log p \sum_{k \geq 2} p^{-k} \\ &= n \sum_{p \leq n} \frac{\log p}{p^2} \cdot \frac{1}{1 - 1/p} = n \sum_{p \leq n} \frac{\log p}{p(p-1)} \\ &\leq n \sum_{k \geq 2} \frac{\log k}{k(k-1)} = O(n). \end{aligned}$$

Thus, we have shown that

$$\log(n!) = \sum_{p \leq n} \lfloor n/p \rfloor \log p + O(n).$$

Further, since $\lfloor n/p \rfloor = n/p + O(1)$, applying Theorem 5.5, we have

$$\log(n!) = \sum_{p \leq n} (n/p) \log p + O(\sum_{p \leq n} \log p) + O(n) = n \sum_{p \leq n} \frac{\log p}{p} + O(n). \quad (5.8)$$

We can also estimate $\log(n!)$ using a little calculus (see §A2). We have

$$\log(n!) = \sum_{k=1}^{n} \log k = \int_1^n \log t \, dt + O(\log n) = n \log n - n + O(\log n). \quad (5.9)$$

Combining (5.8) and (5.9), and noting that $\log x - \log n = o(1)$, we obtain

$$\sum_{p \le x} \frac{\log p}{p} = \log n + O(1) = \log x + O(1),$$

which proves the theorem. □

We shall also need the following theorem, which is a very useful tool in its own right:

Theorem 5.12 (Abel's identity). *Suppose that $c_k, c_{k+1}, \ldots$ is a sequence of numbers, that*

$$C(t) := \sum_{k \le i \le t} c_i,$$

and that $f(t)$ has a continuous derivative $f'(t)$ on the interval $[k, x]$. Then

$$\sum_{k \le i \le x} c_i f(i) = C(x)f(x) - \int_k^x C(t)f'(t)\,dt.$$

Note that since $C(t)$ is a step function, the integrand $C(t)f'(t)$ is piecewise continuous on $[k, x]$, and hence the integral is well defined (see §A3).

Proof. Let $n := \lfloor x \rfloor$. We have

$$\begin{aligned}\sum_{i=k}^{n} c_i f(i) &= C(k)f(k) + [C(k+1) - C(k)]f(k+1) + \cdots \\ &\qquad + [C(n) - C(n-1)]f(n) \\ &= C(k)[f(k) - f(k+1)] + \cdots + C(n-1)[f(n-1) - f(n)] \\ &\qquad + C(n)f(n) \\ &= C(k)[f(k) - f(k+1)] + \cdots + C(n-1)[f(n-1) - f(n)] \\ &\qquad + C(n)[f(n) - f(x)] + C(x)f(x).\end{aligned}$$

Observe that for $i = k, \ldots, n-1$, we have $C(t) = C(i)$ for $t \in [i, i+1)$, and so

$$C(i)[f(i) - f(i+1)] = -\int_i^{i+1} C(t)f'(t)\,dt;$$

likewise,

$$C(n)[f(n) - f(x)] = -\int_n^x C(t)f'(t)\,dt,$$

from which the theorem directly follows. □

Proof of Theorem 5.10. For $i \geq 2$, set

$$c_i := \begin{cases} (\log i)/i & \text{if } i \text{ is prime,} \\ 0 & \text{otherwise.} \end{cases}$$

By Theorem 5.11, we have

$$C(t) := \sum_{2 \leq i \leq t} c_i = \sum_{p \leq t} \frac{\log p}{p} = \log t + O(1).$$

Applying Theorem 5.12 with $f(t) = 1/\log t$, we obtain

$$\begin{aligned} \sum_{p \leq x} \frac{1}{p} &= \frac{C(x)}{\log x} + \int_2^x \frac{C(t)}{t(\log t)^2} dt \\ &= \Big(1 + O(1/\log x)\Big) + \Big(\int_2^x \frac{dt}{t \log t} + O(\int_2^x \frac{dt}{t(\log t)^2})\Big) \\ &= 1 + O(1/\log x) + (\log\log x - \log\log 2) + O(1/\log 2 - 1/\log x) \\ &= \log\log x + O(1). \quad \square \end{aligned}$$

Using Theorem 5.10, we can easily show the following:

Theorem 5.13 (Mertens' theorem). *We have*

$$\prod_{p \leq x} (1 - 1/p) = \Theta(1/\log x).$$

Proof. Using parts (i) and (iii) of §A1, for any fixed prime p, we have

$$-\frac{1}{p^2} \leq \frac{1}{p} + \log(1 - 1/p) \leq 0. \tag{5.10}$$

Moreover, since

$$\sum_{p \leq x} \frac{1}{p^2} \leq \sum_{i \geq 2} \frac{1}{i^2} < \infty,$$

summing the inequality (5.10) over all primes $p \leq x$ yields

$$-C \leq \sum_{p \leq x} \frac{1}{p} + \log U(x) \leq 0,$$

where C is a positive constant, and $U(x) := \prod_{p \leq x}(1 - 1/p)$. From this, and from Theorem 5.10, we obtain

$$\log\log x + \log U(x) = O(1).$$

This means that

$$-D \leq \log\log x + \log U(x) \leq D$$

for some positive constant D and all sufficiently large x, and exponentiating this yields

$$e^{-D} \leq (\log x)U(x) \leq e^{D},$$

and hence, $U(x) = \Theta(1/\log x)$, and the theorem follows. □

EXERCISE 5.4. Let $\omega(n)$ be the number of distinct prime factors of n, and define $\overline{\omega}(x) = \sum_{n \leq x} \omega(n)$, so that $\overline{\omega}(x)/x$ represents the "average" value of ω. First, show that $\overline{\omega}(x) = \sum_{p \leq x} \lfloor x/p \rfloor$. From this, show that $\overline{\omega}(x) \sim x \log\log x$.

EXERCISE 5.5. Analogously to the previous exercise, show that $\sum_{n \leq x} \tau(n) \sim x \log x$, where $\tau(n)$ is the number of positive divisors of n.

EXERCISE 5.6. Define the sequence of numbers $n_1, n_2, \ldots$, where n_k is the product of all the primes up to k. Show that as $k \to \infty$, $\phi(n_k) = \Theta(n_k/\log\log n_k)$. Hint: you will want to use Mertens' theorem, and also Theorem 5.6.

EXERCISE 5.7. The previous exercise showed that $\phi(n)$ could be as small as (about) $n/\log\log n$ for infinitely many n. Show that this is the "worst case," in the sense that $\phi(n) = \Omega(n/\log\log n)$ as $n \to \infty$.

EXERCISE 5.8. Show that for any positive integer constant k,

$$\int_2^x \frac{dt}{(\log t)^k} = \frac{x}{(\log x)^k} + O\left(\frac{x}{(\log x)^{k+1}}\right).$$

EXERCISE 5.9. Use Chebyshev's theorem and Abel's identity to show that

$$\sum_{p \leq x} \frac{1}{\log p} = \frac{\pi(x)}{\log x} + O(x/(\log x)^3).$$

EXERCISE 5.10. Use Chebyshev's theorem and Abel's identity to prove a stronger version of Theorem 5.4:

$$\vartheta(x) = \pi(x)\log x + O(x/\log x).$$

EXERCISE 5.11. Show that

$$\prod_{2 < p \leq x} (1 - 2/p) = \Theta(1/(\log x)^2).$$

EXERCISE 5.12. Show that if $\pi(x) \sim cx/\log x$ for some constant c, then we must have $c = 1$. Hint: use either Theorem 5.10 or 5.11.

Exercise 5.13. Strengthen Theorem 5.10, showing that $\sum_{p\leq x} 1/p \sim \log\log x + A$ for some constant A. (Note: $A \approx 0.261497212847643$.)

Exercise 5.14. Strengthen Mertens' theorem, showing that $\prod_{p\leq x}(1 - 1/p) \sim B_1/(\log x)$ for some constant B_1. Hint: use the result from the previous exercise. (Note: $B_1 \approx 0.561459483566885$.)

Exercise 5.15. Strengthen the result of Exercise 5.11, showing that

$$\prod_{2<p\leq x} (1-2/p) \sim B_2/(\log x)^2$$

for some constant B_2. (Note: $B_2 \approx 0.832429065662$.)

5.4 The sieve of Eratosthenes

As an application of Theorem 5.10, consider the **sieve of Eratosthenes**. This is an algorithm for generating all the primes up to a given bound k. It uses an array $A[2\ldots k]$, and runs as follows.

```
for $n \leftarrow 2$ to $k$ do $A[n] \leftarrow 1$
for $n \leftarrow 2$ to $\lfloor\sqrt{k}\rfloor$ do
    if $A[n] = 1$ then
        $i \leftarrow 2n$; while $i \leq k$ do { $A[i] \leftarrow 0$; $i \leftarrow i + n$ }
```

When the algorithm finishes, we have $A[n] = 1$ if and only if n is prime, for $n = 2, \ldots, k$. This can easily be proven using the fact (see Exercise 1.1) that a composite number n between 2 and k must be divisible by a prime that is at most $\sqrt{k}$, and by proving by induction on n that at the beginning of the nth iteration of the main loop, $A[i] = 0$ iff i is divisible by a prime less than n, for $i = n, \ldots, k$. We leave the details of this to the reader.

We are more interested in the running time of the algorithm. To analyze the running time, we assume that all arithmetic operations take constant time; this is reasonable, since all the quantities computed in the algorithm are bounded by k, and we need to at least be able to index all entries of the array A, which has size k.

Every time we execute the inner loop of the algorithm, we perform $O(k/n)$ steps to clear the entries of A indexed by multiples of n. Naively, we could bound the running time by a constant times

$$\sum_{n\leq\sqrt{k}} k/n,$$

which is $O(k \operatorname{len}(k))$, where we have used a little calculus (see §A2) to derive that

$$\sum_{n=1}^{\ell} 1/n = \int_1^{\ell} \frac{dy}{y} + O(1) \sim \log \ell.$$

However, the inner loop is executed only for prime values of n; thus, the running time is proportional to

$$\sum_{p \le \sqrt{k}} k/p,$$

and so by Theorem 5.10 is $\Theta(k \operatorname{len}(\operatorname{len}(k)))$.

Exercise 5.16. Give a detailed proof of the correctness of the above algorithm.

Exercise 5.17. One drawback of the above algorithm is its use of space: it requires an array of size k. Show how to modify the algorithm, without substantially increasing its running time, so that one can enumerate all the primes up to k, using an auxiliary array of size just $O(\sqrt{k})$.

Exercise 5.18. Design and analyze an algorithm that on input k outputs the table of values $\tau(n)$ for $n = 1, \ldots, k$, where $\tau(n)$ is the number of positive divisors of n. Your algorithm should run in time $O(k \operatorname{len}(k))$.

5.5 The prime number theorem . . . and beyond

In this section, we survey a number of theorems and conjectures related to the distribution of primes. This is a vast area of mathematical research, with a number of very deep results. We shall be stating a number of theorems from the literature in this section without proof; while our intent is to keep the text as self contained as possible, and to avoid degenerating into "mathematical tourism," it nevertheless is a good idea to occasionally have a somewhat broader perspective. In the following chapters, we shall not make any critical use of the theorems in this section.

5.5.1 The prime number theorem

The main theorem in the theory of the density of primes is the following.

Theorem 5.14 (Prime number theorem). *We have*

$$\pi(x) \sim x/\log x.$$

Proof. Literature—see §5.6. □

As we saw in Exercise 5.12, if $\pi(x)/(x/\log x)$ tends to a limit as $x \to \infty$, then the limit must be 1, so in fact the hard part of proving the prime number theorem is to show that $\pi(x)/(x/\log x)$ does indeed tend to some limit.

One simple consequence of the prime number theorem, together with Theorem 5.4, is the following:

Theorem 5.15. *We have*

$$\vartheta(x) \sim x.$$

Exercise 5.19. Using the prime number theorem, show that $p_n \sim n \log n$, where p_n denotes the nth prime.

Exercise 5.20. Using the prime number theorem, show that Bertrand's postulate can be strengthened (asymptotically) as follows: for all $\epsilon > 0$, there exist positive constants c and x_0, such that for all $x \geq x_0$, we have

$$\pi((1+\epsilon)x) - \pi(x) \geq c\frac{x}{\log x}.$$

5.5.2 The error term in the prime number theorem

The prime number theorem says that

$$|\pi(x) - x/\log x| \leq \delta(x),$$

where $\delta(x) = o(x/\log x)$. A natural question is: how small is the "error term" $\delta(x)$? It turns out that:

Theorem 5.16. *We have*

$$\pi(x) = x/\log x + O(x/(\log x)^2).$$

This bound on the error term is not very impressive. The reason is that $x/\log x$ is not really the best "simple" function that approximates $\pi(x)$. It turns out that a better approximation to $\pi(x)$ is the **logarithmic integral**, defined for real $x \geq 2$ by

$$\mathrm{li}(x) := \int_2^x \frac{dt}{\log t}.$$

It is not hard to show (see Exercise 5.8) that

$$\mathrm{li}(x) = x/\log x + O(x/(\log x)^2).$$

Table 5.2. *Values of* $\pi(x)$, $\mathrm{li}(x)$, *and* $x/\log x$

x	$\pi(x)$	$\mathrm{li}(x)$	$x/\log x$
10^3	168	176.6	144.8
10^6	78498	78626.5	72382.4
10^9	50847534	50849233.9	48254942.4
10^{12}	37607912018	37607950279.8	36191206825.3
10^{15}	29844570422669	29844571475286.5	28952965460216.8
10^{18}	24739954287740860	24739954309690414.0	24127471216847323.8

Thus, $\mathrm{li}(x) \sim x/\log x \sim \pi(x)$. However, the error term in the approximation of $\pi(x)$ by $\mathrm{li}(x)$ is much better. This is illustrated numerically in Table 5.2; for example, at $x = 10^{18}$, $\mathrm{li}(x)$ approximates $\pi(x)$ with a relative error just under 10^{-9}, while $x/\log x$ approximates $\pi(x)$ with a relative error of about 0.025.

The sharpest proven result is the following:

Theorem 5.17. *Let* $\kappa(x) := (\log x)^{3/5}(\log\log x)^{-1/5}$*. Then for some* $c > 0$*, we have*

$$\pi(x) = \mathrm{li}(x) + O(xe^{-c\kappa(x)}).$$

Proof. Literature—see §5.6. □

Note that the error term $xe^{-c\kappa(x)}$ is $o(x/(\log x)^k)$ for every fixed $k \geq 0$. Also note that Theorem 5.16 follows directly from the above theorem and Exercise 5.8.

Although the above estimate on the error term in the approximation of $\pi(x)$ by $\mathrm{li}(x)$ is pretty good, it is conjectured that the actual error term is much smaller:

Conjecture 5.18. *For all* $x \geq 2.01$*, we have*

$$|\pi(x) - \mathrm{li}(x)| < x^{1/2}\log x.$$

Conjecture 5.18 is equivalent to a famous conjecture called the **Riemann hypothesis**, which is an assumption about the location of the zeros of a certain function, called **Riemann's zeta function**. We give a *very* brief, high-level account of this conjecture, and its connection to the theory of the distribution of primes.

For real $s > 1$, the zeta function is defined as

$$\zeta(s) := \sum_{n=1}^{\infty} \frac{1}{n^s}. \tag{5.11}$$

Note that because $s > 1$, the infinite series defining $\zeta(s)$ converges. A simple, but important, connection between the zeta function and the theory of prime numbers is the following:

Theorem 5.19 (Euler's identity). *For real $s > 1$, we have*

$$\zeta(s) = \prod_p (1 - p^{-s})^{-1}, \tag{5.12}$$

where the product is over all primes p.

Proof. The rigorous interpretation of the infinite product on the right-hand side of (5.12) is as a limit of finite products. Thus, if $p_1, p_2, \ldots$ is the list of primes, we are really proving that

$$\zeta(s) = \lim_{r\to\infty} \prod_{i=1}^{r} (1 - p_i^{-s})^{-1}.$$

Now, from the identity

$$(1 - p_i^{-s})^{-1} = \sum_{e=0}^{\infty} p_i^{-es},$$

we have

$$\begin{aligned}\prod_{i=1}^{r}(1 - p_i^{-s})^{-1} &= \Big(1 + p_1^{-s} + p_1^{-2s} + \cdots\Big)\cdots\Big(1 + p_r^{-s} + p_r^{-2s} + \cdots\Big)\\ &= \sum_{e_1=0}^{\infty}\cdots\sum_{e_r=0}^{\infty}(p_1^{e_1}\cdots p_r^{e_r})^s\\ &= \sum_{n=1}^{\infty}\frac{g_r(n)}{n^s},\end{aligned}$$

where

$$g_r(n) := \begin{cases} 1 & \text{if } n \text{ is divisible only by the primes } p_1, \ldots, p_r;\\ 0 & \text{otherwise.}\end{cases}$$

Here, we have made use of the fact (see §A5) that we can multiply term-wise infinite series with non-negative terms.

Now, for any $\epsilon > 0$, there exists n_0 such that $\sum_{n=n_0}^{\infty} n^{-s} < \epsilon$ (because the series defining $\zeta(s)$ converges). Moreover, there exists an r_0 such that $g_r(n) = 1$ for all $n < n_0$ and $r \geq r_0$. Therefore, for $r \geq r_0$, we have

$$\left|\sum_{n=1}^{\infty}\frac{g_r(n)}{n^s} - \zeta(s)\right| \leq \sum_{n=n_0}^{\infty} n^{-s} < \epsilon.$$

It follows that

$$\lim_{r\to\infty}\sum_{n=1}^{\infty}\frac{g_r(n)}{n^s}=\zeta(s),$$

which proves the theorem. □

While Theorem 5.19 is nice, things become much more interesting if one extends the domain of definition of the zeta function to the complex plane. For the reader who is familiar with just a little complex analysis, it is easy to see that the infinite series defining the zeta function in (5.11) converges absolutely for complex numbers s whose real part is greater than 1, and that (5.12) holds as well for such s. However, it is possible to extend the domain of definition of ζ even further—in fact, one can extend the definition of ζ in a "nice way " (in the language of complex analysis, *analytically continue*) to the entire complex plane (except the point $s = 1$, where there is a simple pole). Exactly how this is done is beyond the scope of this text, but assuming this extended definition of ζ, we can now state the Riemann hypothesis:

Conjecture 5.20 (Riemann hypothesis). *For any complex number $s = x + yi$, where x and y are real numbers with $0 < x < 1$ and $x \neq 1/2$, we have $\zeta(s) \neq 0$.*

A lot is known about the zeros of the zeta function in the "critical strip," consisting of those points s whose real part is greater than 0 and less than 1: it is known that there are infinitely many of them, and there are even good estimates about their density. It turns out that one can apply standard tools in complex analysis, like contour integration, to the zeta function (and functions derived from it) to answer various questions about the distribution of primes. Indeed, such techniques may be used to prove the prime number theorem. However, if one assumes the Riemann hypothesis, then these techniques yield much sharper results, such as the bound in Conjecture 5.18.

Exercise 5.21. For any arithmetic function a, we can form the **Dirichlet series**

$$F_a(s) := \sum_{n=1}^{\infty}\frac{a(n)}{n^s}.$$

For simplicity we assume that s takes only real values, even though such series are usually studied for complex values of s.

(a) Show that if the Dirichlet series $F_a(s)$ converges absolutely for some real s, then it converges absolutely for all real $s' \geq s$.

(b) From part (a), conclude that for any given arithmetic function a, there is an **interval of absolute convergence** of the form (s_0, ∞), where we allow $s_0 = -\infty$ and $s_0 = \infty$, such that $F_a(s)$ converges absolutely for $s > s_0$, and does not converge absolutely for $s < s_0$.

(c) Let a and b be arithmetic functions such that $F_a(s)$ has an interval of absolute convergence (s_0, ∞) and $F_b(s)$ has an interval of absolute convergence (s_0', ∞), and assume that $s_0 < \infty$ and $s_0' < \infty$. Let $c := a \star b$ be the Dirichlet product of a and b, as defined in §2.6. Show that for all $s \in (\max(s_0, s_0'), \infty)$, the series $F_c(s)$ converges absolutely and, moreover, that $F_a(s)F_b(s) = F_c(s)$.

5.5.3 Explicit estimates

Sometimes, it is useful to have explicit estimates for $\pi(x)$, as well as related functions, like $\vartheta(x)$ and the nth prime function p_n. The following theorem presents a number of bounds that have been proved without relying on any unproved conjectures.

Theorem 5.21. *We have:*

(i) $\dfrac{x}{\log x}\left(1 + \dfrac{1}{2\log x}\right) < \pi(x) < \dfrac{x}{\log x}\left(1 + \dfrac{3}{2\log x}\right)$, *for* $x \geq 59$*;*

(ii) $n(\log n + \log\log n - 3/2) < p_n < n(\log n + \log\log n - 1/2)$, *for* $n \geq 20$*;*

(iii) $x(1 - 1/(2\log x)) < \vartheta(x) < x(1 + 1/(2\log x))$, *for* $x \geq 563$*;*

(iv) $\log\log x + A - \dfrac{1}{2(\log x)^2} < \sum_{p \leq x} 1/p < \log\log x + A + \dfrac{1}{2(\log x)^2}$, *for* $x \geq 286$*, where* $A \approx 0.261497212847643$*;*

(v) $\dfrac{B_1}{\log x}\left(1 - \dfrac{1}{2(\log x)^2}\right) < \prod_{p \leq x}\left(1 - \dfrac{1}{p}\right) < \dfrac{B_1}{\log x}\left(1 + \dfrac{1}{2(\log x)^2}\right)$, *for* $x \geq 285$*, where* $B_1 \approx 0.561459483566885$*.*

Proof. Literature—see §5.6. □

5.5.4 Primes in arithmetic progressions

The arithmetic progression of odd numbers $1, 3, 5, \ldots$ contains infinitely many primes, and it is natural to ask if other arithmetic progressions do as well. An arithmetic progression with first term a and common difference d consists of all integers of the form

$$md + a, \quad m = 0, 1, 2, \ldots.$$

If d and a have a common factor $c > 1$, then every term in the progression is divisible by c, and so there can be no more than one prime in the progression. So a necessary condition for the existence of infinitely many primes p with $p \equiv a \pmod{d}$ is that $\gcd(d, a) = 1$. A famous theorem due to Dirichlet states that this is a sufficient condition as well.

Theorem 5.22 (Dirichlet's theorem). *For any positive integer d and any integer a relatively prime to d, there are infinitely many primes p with $p \equiv a \pmod{d}$.*

Proof. Literature—see §5.6. □

We can also ask about the density of primes in arithmetic progressions. One might expect that for a fixed value of d, the primes are distributed in roughly equal measure among the $\phi(d)$ different residue classes $[a]_d$ with $\gcd(a, d) = 1$. This is in fact the case. To formulate such assertions, we define $\pi(x; d, a)$ to be the number of primes p up to x with $p \equiv a \pmod{d}$.

Theorem 5.23. *Let $d > 0$ be a fixed integer, and let $a \in \mathbb{Z}$ be relatively prime to d. Then*

$$\pi(x; d, a) \sim \frac{x}{\phi(d) \log x}.$$

Proof. Literature—see §5.6. □

The above theorem is only applicable in the case where d is fixed and $x \to \infty$. But what if we want an estimate on the number of primes p up to x with $p \equiv a \pmod{d}$, where x is, say, a fixed power of d? Theorem 5.23 does not help us here. The following conjecture does, however:

Conjecture 5.24. *For any real $x \geq 2$, integer $d \geq 2$, and $a \in \mathbb{Z}$ relatively prime to d, we have*

$$\left| \pi(x; d, a) - \frac{\mathrm{li}(x)}{\phi(d)} \right| \leq x^{1/2}(\log x + 2 \log d).$$

The above conjecture is in fact a consequence of a generalization of the Riemann hypothesis—see §5.6.

Exercise 5.22. Assuming Conjecture 5.24, show that for all α, ϵ, with $0 < \alpha < 1/2$ and $0 < \epsilon < 1$, there exists an x_0, such that for all $x > x_0$, for all $d \in \mathbb{Z}$ with $2 \leq d \leq x^{\alpha}$, and for all $a \in \mathbb{Z}$ relatively prime to d, the number of primes $p \leq x$ such that $p \equiv a \pmod{d}$ is at least $(1 - \epsilon)\,\mathrm{li}(x)/\phi(d)$ and at most $(1 + \epsilon)\,\mathrm{li}(x)/\phi(d)$.

It is an open problem to prove an unconditional density result analogous to Exercise 5.22 for any positive exponent α. The following, however, is known:

Theorem 5.25. *There exists a constant c such that for all integer $d \geq 2$ and $a \in \mathbb{Z}$ relatively prime to d, the least prime p with $p \equiv a \pmod{d}$ is at most $cd^{11/2}$.*

Proof. Literature—see §5.6. □

5.5.5 Sophie Germain primes

A **Sophie Germain prime** is a prime p such that $2p + 1$ is also prime. Such primes are actually useful in a number of practical applications, and so we discuss them briefly here.

It is an open problem to prove (or disprove) that there are infinitely many Sophie Germain primes. However, numerical evidence, and heuristic arguments, strongly suggest not only that there are infinitely many such primes, but also a fairly precise estimate on the density of such primes.

Let $\pi^*(x)$ denote the number of Sophie Germain primes up to x.

Conjecture 5.26. *We have*

$$\pi^*(x) \sim C \frac{x}{(\log x)^2},$$

where C is the constant

$$C := 2 \prod_{q>2} \frac{q(q-2)}{(q-1)^2} \approx 1.32032,$$

and the product is over all primes $q > 2$.

The above conjecture is a special case of a more general conjecture, known as **Hypothesis H**. We can formulate a special case of Hypothesis H (which includes Conjecture 5.26), as follows:

Conjecture 5.27. *Let $(a_1, b_1), \ldots, (a_k, b_k)$ be distinct pairs of integers such that $a_i > 0$, and for all primes p, there exists an integer m such that*

$$\prod_{i=1}^{k} (ma_i + b_i) \not\equiv 0 \pmod{p}.$$

Let $P(x)$ be the number of integers m up to x such that $ma_i + b_i$ are simultaneously prime for $i = 1, \ldots, k$. Then

$$P(x) \sim D \frac{x}{(\log x)^k},$$

where

$$D := \prod_p \left\{ \left(1 - \frac{1}{p}\right)^{-k} \left(1 - \frac{\omega(p)}{p}\right) \right\},$$

the product being over all primes p, and $\omega(p)$ being the number of distinct solutions m modulo p to the congruence

$$\prod_{i=1}^{k} (ma_i + b_i) \equiv 0 \pmod{p}.$$

The above conjecture also includes (a strong version of) the famous **twin primes conjecture** as a special case: the number of primes p up to x such that $p+2$ is also prime is $\sim Cx/(\log x)^2$, where C is the same constant as in Conjecture 5.26.

Exercise 5.23. Show that the constant C appearing in Conjecture 5.26 satisfies

$$2C = B_2/B_1^2,$$

where B_1 and B_2 are the constants from Exercises 5.14 and 5.15.

Exercise 5.24. Show that the quantity D appearing in Conjecture 5.27 is well defined, and satisfies $0 < D < \infty$.

5.6 Notes

The prime number theorem was conjectured by Gauss in 1791. It was proven independently in 1896 by Hadamard and de la Vallée Poussin. A proof of the prime number theorem may be found, for example, in the book by Hardy and Wright [44].

Theorem 5.21, as well as the estimates for the constants A, B_1, and B_2 mentioned in that theorem and Exercises 5.13, 5.14, and 5.15, are from Rosser and Schoenfeld [79].

Theorem 5.17 is from Walfisz [96].

Theorem 5.19, which made the first connection between the theory of prime numbers and the zeta function, was discovered in the 18th century by Euler. The Riemann hypothesis was made by Riemann in 1859, and to this day, remains one of the most vexing conjectures in mathematics. Riemann in fact showed that his conjecture about the zeros of the zeta function is equivalent to the conjecture that for each fixed $\epsilon > 0$, $\pi(x) = \mathrm{li}(x) + O(x^{1/2+\epsilon})$. This was strengthened by von Koch in 1901, who showed that the Riemann hypothesis is true if and only if $\pi(x) = \text{li}(x) + O(x^{1/2} \log x)$. See Chapter 1 of the book by Crandall and Pomerance [30] for more on the connection between the Riemann hypothesis and the theory of prime numbers; in particular, see Exercise 1.36 in that book for an outline of a proof that Conjecture 5.18 follows from the Riemann hypothesis.

A warning: some authors (and software packages) define the logarithmic integral using the interval of integration $(0, x)$, rather than $(2, x)$, which increases its value by a constant $c \approx 1.0452$.

Theorem 5.22 was proved by Dirichlet in 1837, while Theorem 5.23 was proved by de la Vallée Poussin in 1896. A result of Oesterlé [69] implies that Conjecture 5.24 for $d \geq 3$ is a consequence of an assumption about the location of the zeros of certain generalizations of Riemann's zeta function; the case $d = 2$ follows from the bound in Conjecture 5.18 under the ordinary Riemann hypothesis. Theorem 5.25 is from Heath-Brown [45].

Hypothesis H is from Hardy and Littlewood [43].

For the reader who is interested in learning more on the topics discussed in this chapter, we recommend the books by Apostol [8] and Hardy and Wright [44]; indeed, many of the proofs presented in this chapter are minor variations on proofs from these two books. Our proof of Bertrand's postulate is based on the presentation in Section 9.2 of Redmond [76]. See also Bach and Shallit [12] (especially Chapter 8), Crandall and Pomerance [30] (especially Chapter 1) for a more detailed overview of these topics.

The data in Tables 5.1 and 5.2 was obtained using the computer program *Maple.*