Chapter Two

Notation and preliminaries

The main purpose of this chapter is to set some notation and review some standard material quickly, as one inevitably does when setting notation. All or nearly all we shall see is standard. The particular way in which alternative Diophantine approximations a/q, a'/q' are set out in Lemma 2.2 may not be as common as it should be.

2.1 GENERAL NOTATION

2.1.1 Basic functions on $\mathbb{Z}$, $\mathbb{R}/\mathbb{Z}$ and $\mathbb{R}$

Given positive integers m, n, we write $m|n$ ("m divides n") to mean that n/m is an integer. We write $m|n^\infty$ to mean that every prime dividing m also divides n. We say a positive integer n is *square-free* if $p^2 \nmid n$ for every prime p. For p prime, n a non-zero integer, we define $v_p(n)$ to be the largest non-negative integer α such that $p^\alpha|n$.

When we write $\sum_{n\leq x}$, we mean $\sum_{1\leq n\leq x}$. Whether $\sum_n$ means $\sum_{n=-\infty}^{\infty}$ (that is, $\sum_{n\in\mathbb{Z}}$) or $\sum_{n=1}^{\infty}$ (that is, $\sum_{n\in\mathbb{Z}^+}$) will be clear from context; we will avoid the use of $\sum_n$ when there is any risk of confusion.

As always, $\Lambda(n)$ denotes the *von Mangoldt function*:

$$\Lambda(n) = \begin{cases} \log p & \text{if } n = p^\alpha \text{ for some prime } p \text{ and some integer } \alpha \geq 1,\\ 0 & \text{otherwise,}\end{cases}$$

and μ denotes the *Möbius function*

$$\mu(n) = \begin{cases} (-1)^k & \text{if } n = p_1 p_2 \cdots p_k, \text{ all } p_i \text{ distinct}\\ 0 & \text{if } p^2|n \text{ for some prime } p.\end{cases}$$

It is easy to show that

$$\sum_{d|n} \mu(d) = \begin{cases} 0 & \text{if } n > 1\\ 1 & \text{if } n = 1,\end{cases} \tag{2.1}$$

which implies (and is a special case of) the *Möbius inversion formula*. The Möbius inversion formula states that, for any $f : \mathbb{Z}^+ \to \mathbb{C}$ and $g(n) = \sum_{d|n} f(d)$,

$$f(n) = \sum_{d|n} \mu(d) g\left(\frac{n}{d}\right). \tag{2.2}$$

Euler's function ϕ is defined by

$$\phi(n) = n \cdot \prod_{p|n} \left(1 - \frac{1}{p}\right).$$

We let $\tau(n)$ be the number of divisors of an integer n, $\omega(n)$ the number of prime divisors of n, and $\sigma(n)$ the sum of the divisors of n. For n square-free,

$$\sigma(n) = n \cdot \prod_{p|n} \left(1 + \frac{1}{p}\right).$$

We write (a, b) for the greatest common divisor of a and b. If there is any risk of confusion with the pair (a, b) or the interval (a, b), we write $\gcd(a, b)$. Denote by (a, b^∞) the divisor $\prod_{p|b} p^{v_p(a)}$ of a. (Thus, $a/(a, b^\infty)$ is coprime to b, and is in fact the maximal divisor of a with this property.) We write $[a, b]$ for the least common multiple of a and b; again, if there were any risk of confusion with an interval, we would write $\mathrm{lcm}(a, b)$ instead.

As is customary, we write $e(x)$ for $e^{2\pi i x}$ and $\log^+ x$ for $\max(\log x, 0)$. For x real, we let $\lfloor x \rfloor$ ("integer part" or "floor") denote the largest integer $n \leq x$, and let $\lceil x \rceil$ ("ceiling") denote the smallest integer $n \geq x$.

For $a, b \in \mathbb{Z}$ and $m \in \mathbb{Z}^+$, we write $a \equiv b \bmod m$ to mean, of course, that m divides $b - a$. Given x real, we write $x \bmod 1$ to mean $x \bmod \mathbb{Z}$. By $x \equiv y \bmod 1$, or $x = y \bmod 1$, we mean that $x + \mathbb{Z} = y + \mathbb{Z}$, i.e., $x - y \in \mathbb{Z}$.

Let $x \in \mathbb{R}$. There are two common conventions on $\{x\}$ ("fractional part"): for some writers, $\{x\}$ denotes $x - \lfloor x \rfloor$, whereas, for others, $\{x\}$ denotes the element y of $(-1/2, 1/2]$ such that $x \equiv y \bmod 1$. We follow the first convention, as we more or less have to if we use Euler-Maclaurin and Bernoulli polynomials in their standard form (3.1.2). We will write $d(x, \mathbb{Z})$ for the distance of x to the closest integer; evidently, $d(x, \mathbb{Z})$ equals $|y|$, where $y \in (-1/2, 1/2]$ is as above. Since $d(x, \mathbb{Z})$ depends only on $x \bmod \mathbb{Z}$, we can also define $d(x, \mathbb{Z})$ for $x \in \mathbb{R}/\mathbb{Z}$.

Given two elements $\alpha_1, \alpha_2 \in \mathbb{R}/\mathbb{Z}$, we will write $d(\alpha_1, \alpha_2)$ for the distance between them, meaning the unique element $\delta \in [0, 1/2]$ such that at least one of the two statements $\alpha_1 = \alpha_2 + \delta \bmod 1$, $\alpha_1 = \alpha_2 - \delta \bmod 1$ holds. It is clear that $d(\alpha_1, \alpha_2) = d(\alpha_1 - \alpha_2, \mathbb{Z})$.

Given a set S, we write 1_S for its *characteristic function*[1], also called *indicator function*:

$$1_S(x) = \begin{cases} 1 & \text{if } x \in S, \\ 0 & \text{otherwise.} \end{cases}$$

The symbol $\delta_{x,y}$ (the *Kronecker delta*) is defined to equal 1 when $x = y$ and 0 when $x \neq y$. Usually, x and y here are discrete variables – most often integers. We

[1] Probabilists who use "characteristic function" to mean "Fourier transform" (of a Lebesgue measure) owe other people an explanation.

will soon discuss (Dirac) delta functions in $\mathbb{R}$; they are related to but not the same as the Kronecker delta.

There is a much more general bit of notation, namely, the *Iverson bracket* $[P]$, defined to be 1 when P is true and 0 when P is false. Thus, for instance, $\delta_{x,y} = [x = y]$. Usage of the Iverson bracket is unfortunately not universal, as it arguably should be. We will use it when it would be awkward to do without it.

2.1.2 Asymptotics

As is usual, for $f : \mathbb{R}^+ \to \mathbb{C}$, $g : \mathbb{R}^+ \to \mathbb{R}$, the statements $f(t) \ll g(t)$ means that there is an unspecified constant $C > 0$ such that $|f(t)| \leq Cg(t)$ for all large enough t. The constant C is called the *implied constant*. We define $f(t) = O(g(t))$ and $g(t) \gg f(t)$ to mean the same as $f(t) \ll g(t)$, though we would generally write $g(t) \gg f(t)$ only if $f(t)$ and $g(t)$ are both real and non-negative for t large enough. The asymptotic notation $f \sim g$ means that $\lim_{t\to\infty} f(t)/g(t) = 1$, whereas $f \asymp g$ means that $1 \ll f(t)/g(t) \ll 1$.

We can also use $\ll$, $\gg$, $O(\cdot)$ and $\sim$ when $t \to t_0$, $t \to t_0^+$ or $t \to t_0^-$ rather than $t \to \infty$, provided that we say so explicitly. The same notation applies when f, g are functions of a complex variable $s \to s_0$, or of an integer variable $n \to \infty$. When we write O_ϵ, $\ll_\epsilon$ or $\gg_\epsilon$ (say), we mean that the implied constant depends on ϵ.

We can write simply $O(g(t))$, rather than $f(t) = O(g(t))$, to mean a quantity – such as an error term – bounded by $Cg(t)$. It is also acceptable usage to write, say, "$f(x,y) \ll g(x,y)$ provided that $x \ll y$", with the condition $x \ll y$ meaning that $x \leq Cy$ for some $C > 0$, where x, y are positive variables. At any rate, since, aiming at explicit results, we cannot generally work with statements with unspecified constants, we will use $\ll$, $\gg$ and $O(\cdot)$ mainly conversationally and fairly rarely. It will be more common for us to write $O^*(R)$, which means a quantity s such that $|s| \leq R$.

We will write $f(x^+)$ to mean $\lim_{y\to x^+} f(y)$ and $f(x^-)$ to mean $\lim_{y\to x^-} f(y)$.

2.1.3 Support and convexity

For a function $f : X \to \mathbb{C}$ on a set X, the *support* $\operatorname{supp}(f)$ of f is the set $\{x \in X : f(x) \neq 0\}$. If X is endowed with a topology, then its support $\operatorname{supp}(f)$ is instead defined to be the closure of $\{x \in X : f(x) \neq 0\}$. Of course the two definitions coincide if X is given the discrete topology (as is usually the case for X finite, or for $X = \mathbb{Z}$). The support of a sequence $\{a_n\}_{n=1}^\infty$ is the set $\{n \in \mathbb{Z}^+ : a_n \neq 0\}$. We say that that a function f is *supported* on a set S if its support is contained in S. Similarly, a sequence $\{a_n\}_{n=1}^\infty$ is *supported* on $S \subset \mathbb{Z}^+$ if its support is contained in S.

A function is of *compact support*, or *compactly supported*, if its support is compact. A function $f : \mathbb{Z}^+ \to \mathbb{C}$ has *prime support* if it is supported on the set of primes.

We say a property holds almost everywhere (a.e., for short) if it holds outside a set of measure zero.

We follow convention by calling a function $f : \mathbb{R} \to \mathbb{R}$ *convex* if $f(tx+(1-t)y) \leq tf(x) + (1-t)f(y)$ for all $x, y \in \mathbb{R}$ and all $0 \leq t \leq 1$, and *concave* if the inequality is reversed: $f(tx+(1-t)y) \geq tf(x)+(1-t)f(y)$ for all $0 \leq t \leq 1$. Non-mathematicians follow the opposite convention; they do not call a hill *concave*, except precisely when one of its sides is convex in our sense. It may be that we look at functions from on high.

2.1.4 Pedestrian matters

1. Following the habits of all carbon-based life forms, when we write $x/2q$, say, we mean $x/(2q)$, and not $(x/2) \cdot q$.
2. We will use references to displayed quantities to mean the said quantities. Thus, if we have written
$$x^2 + 3C + 0.5, \tag{2.3}$$
and we say "(2.3) is less than 1", we mean that $x^2 + 3C + 0.5 < 1$, and not that the real number 2.3 is less than 1.
3. There is nothing wrong with using footnotes here and there.
4. This subsection could have been called "prosaic matters", but the rest of the book is not in verse, either literally or metaphorically.

With that said, let us hop onto our quadrigas, or rather τέϑριπποι.

2.2 DIOPHANTINE APPROXIMATION

The following result is often stated with the condition that Q be an integer. Let us prove it for all real $Q \geq 1$. (This variant is of course well-known; see, e.g., [Vau97, Lemma 2.1].)

Lemma 2.1 (Dirichlet's approximation theorem)**.** *Let $\alpha \in \mathbb{R}$, $Q \geq 1$. Then there are integers a, $1 \leq q \leq Q$ with $(a, q) = 1$ such that*
$$\left|\alpha - \frac{a}{q}\right| < \frac{1}{qQ}. \tag{2.4}$$

If only a non-strict inequality $\leq$ is desired in (2.4), the proof below can be modified so as to give a strict inequality for q, that is, $q < Q$.

Proof. Let $m = \lfloor Q \rfloor$. Consider the following $m + 1$ points in the circle $\mathbb{R}/\mathbb{Z}$:
$$0 \cdot \alpha, 1 \cdot \alpha, \dots, m \cdot \alpha \mod \mathbb{Z}.$$
At least two of them must be at distance no more than $1/(m+1)$ from each other. Call them $b_1\alpha \bmod \mathbb{Z}$, $b_2\alpha \bmod \mathbb{Z}$, $b_1 < b_2$. Then
$$d((b_2 - b_1)\alpha, \mathbb{Z}) = d(b_2\alpha - b_1\alpha, \mathbb{Z}) \leq \frac{1}{m+1} < \frac{1}{Q},$$
i.e., there is an integer a such that $|(b_2 - b_1)\alpha - a| < 1/Q$. Define $q = b_2 - b_1$. Then $|\alpha - a/q| < 1/qQ$. If a, q are not coprime, divide both of them by (a, q). □

It is also possible to prove the same lemma using the basic theory of continued fractions; see [HW79, Thm. 164] or [Khi64, Thm. 9].

Lemma 2.1 tells us that, for any $\alpha \in \mathbb{R}$, there are good approximations a/q to α with q not too large. However, sometimes we want a q that is neither too large nor too small. The following lemma shows that we can fix that, provided that the approximation a/q is not too good (nor too bad).

Lemma 2.2. *Let $\alpha \in \mathbb{R}$. Let a and $q \geq 1$ be coprime integers such that $\epsilon = |\alpha - a/q|$ satisfies $0 < \epsilon \leq 1/q^2$. Then, for $Q = 1/\epsilon q$, there exist coprime integers a', q' such that*

$$\left|\alpha - \frac{a'}{q'}\right| \leq \frac{1}{q'Q} \qquad \textit{and} \qquad \frac{Q}{2} < q' \leq Q.$$

Proof. By the definition of ϵ and Q, we have $|\alpha - a/q| = \epsilon = 1/qQ$. Since $\epsilon \leq 1/q^2$, we see that $q \leq Q$. Dirichlet's approximation theorem (Lem. 2.1) assures us that there exist coprime integers a_1, $1 \leq q_1 \leq Q$ such that $|\alpha - a_1/q_1| < 1/q_1 Q$. Then $a/q \neq a_1/q_1$, and so

$$\left|\frac{a}{q} - \frac{a_1}{q_1}\right| \geq \frac{1}{qq_1}.$$

At the same time,

$$\left|\frac{a}{q} - \frac{a_1}{q_1}\right| \leq \left|\alpha - \frac{a}{q}\right| + \left|\alpha - \frac{a_1}{q_1}\right| < \frac{1}{qQ} + \frac{1}{q_1 Q}.$$

Hence $1/qQ + 1/q_1 Q > 1/qq_1$, and so $q + q_1 > Q$. If $q_1 > Q/2$, let $a'/q' = a_1/q_1$. If $q_1 \leq Q/2$, then $q > Q/2$, and we may let $a'/q' = a/q$. □

The idea of using different approximations to the same α is, of course, not new, even within the context of the circle method: see, e.g., [Vau97, §2.8, Ex. 2]. Lemma 2.2 can be seen as a convenient formulation of an existing general idea.

On a different but related note: we will not use continued fractions, and thus need not review the basic nomenclature of the area; however, the following remark may be useful for some readers. The approximations a/q, a'/q' are not necessarily "successive approximants"; we do not want to assume that the approximation a/q we are given arises from a continued fraction. At the same time, it follows from [Khi64, Thm. 13]. that, if a/q is in fact a continued-fraction approximant, then either the next approximant is a valid value for a'/q' or a/q itself is a valid value for a'/q'.

2.3 BASICS ON NORMS IN ANALYSIS

2.3.1 Norms and the inner product

We write $|f|_r$ for the L^r norm of a function $f : X \to \mathbb{C}$:

$$|f|_r = \left(\int_X |f(x)|^r d\mu\right)^{1/r}.$$

The measure $d\mu$ and the space X will generally be clear from context; we are almost always speaking of functions f over $\mathbb{R}$ (with the usual measure $d\mu = dx$) or over $\mathbb{Z}$ or $\mathbb{Z}^+$ (again with the usual measure, and so $|f|_r = (\sum_n |f(n)|^r)^{1/r}$). It is understood that by "measure" here we mean "positive measure", that is, $d\mu$ takes only non-negative values. As is always the case, we say that f is in L^r (or: $f \in L^r$) if $|f|_r$ is well-defined and finite.

Strictly speaking, $|\cdot|_r$ is only a seminorm on functions f, in that $|f|_r = 0$ for functions f that are zero almost everywhere. Thus, the right way to define the space $L^r = L^r(X)$ is as the quotient of the vector space of functions $f : X \to \mathbb{C}$ with finite L^r norm by the vector space of functions that are zero almost everywhere. Then $|\cdot|_r$ is really a norm on $L^r(X)$.

We will treat "L^r" and "ℓ^r" as essentially synonymous; however, as is common, we will prefer to use ℓ^r when the space in question is discrete, and L^r when it is continuous. Thus, for instance, we use L^r when working over $\mathbb{R}$ or $\mathbb{R}/\mathbb{Z}$, and ℓ^r when working over $\mathbb{Z}$.

A function $f : \mathbb{R} \to \mathbb{C}$ is *integrable* if $|f|_1$ is well-defined and finite.

We define the inner product $\langle f, g\rangle$ of two complex-valued functions f, g in the usual way:

$$\langle f, g\rangle = \int \overline{f(x)} g(x) d\mu.$$

Obviously, $\langle f, f\rangle = |f|_2^2$. It is an easy exercise to show that one can express an inner product as a linear combination of squares of ℓ^2-norms:

$$\langle f, g\rangle = \frac{1}{2}(|f+g|_2^2 - i|f+ig|_2^2) - \frac{1-i}{2}(|f|_2^2 + |g|_2^2).$$

Of course, we can also define an inner product $\langle\cdot,\cdot\rangle$ on an abstract vector space V over $\mathbb{C}$ as a bilinear map $V \times V \to V$ that is positive definite ($\langle v, v\rangle \geq 0$, with equality if and only if $v = 0$) and satisfies $\langle v, w\rangle = \overline{\langle w, v\rangle}$. The inner product defines a norm $|v|_2 = \sqrt{\langle v, v\rangle}$ and the norm defines a metric $|v - w|_2$. If V is complete with respect to this metric, we say it is a *Hilbert space*. As we were seeing, a space $L^2(X)$ defined with respect to any positive measure μ has an inner product; in fact, it is a Hilbert space.

The Cauchy-Schwarz[2] inequality $\langle v, w\rangle \leq |v|_2|w|_2$ is true for any inner-product space. The usual proof for sequences $\{a_n\}$, $\{b_n\}$ works in this abstract setting, since it just uses the positivity of the inner product [Rud74, §4.2].

2.3.2 The duality principle for linear operators

Let V, W be normed vector spaces, that is, vector spaces each endowed with a norm, and with the metric and topology the norm induces. A linear operator is just a linear function $A : V \to W$. We say that A is bounded if there is a constant C such that $|Av| \leq C|v|$ for all $v \in V$. It is easy to see that an operator A is continuous if and only if it is bounded. We define the *operator norm* $|A|$ of A to be $\sup_{v\in V:\, v\neq 0} |Av|/|v|$.

[2] Feel free to add "Bunyakovsky". Cauchy proved the inequality $|\sum_n a_n b_n| \leq \sqrt{\sum_n |a_n|^2} \cdot \sqrt{\sum_n |b_n|^2}$, and Bunyakovsky (1859) and Schwarz (1885) generalized it to integrals.

Now let V, W be Hilbert spaces. Let $A : V \to W$ be a bounded operator, where V and W are endowed with the norms $\sqrt{\langle \cdot, \cdot \rangle}$ induced by their inner products. Then, for any $w \in W$, the map $v \to \langle w, Av \rangle$ from V to $\mathbb{C}$ is continuous and linear. Now, for any continuous linear map $L : V \to \mathbb{C}$ on a Hilbert space V, there is a unique $x \in V$ such that $\langle x, v \rangle = Lv$ for all $v \in V$ (*Riesz representation theorem*,[3] [Rud74, Thm. 4.12]). We define the *dual operator* of A to be the operator $A^* : W \to V$ such that

$$\langle A^* w, v \rangle = \langle w, Av \rangle$$

for all $v \in V$, $w \in W$. If V and W are finite-dimensional and A is written as a matrix, then A^* is just the conjugate transpose of A.

Lemma 2.3 (Duality principle). *Let V, W be Hilbert spaces and $A : V \to W$ a bounded operator. Then*

$$|A| = |A^*|,$$

where $| \cdot |$ denotes the operator norm.

Proof. By Cauchy-Schwarz, $|Av| = \sup_{w \in W : w \neq 0} \langle w, Av \rangle / |w|$ for any $v \in V$. Hence

$$\sup_{\substack{v \in V \\ v \neq 0}} \frac{|Av|}{|v|} = \sup_{\substack{v \in V \\ v \neq 0}} \sup_{\substack{w \in W \\ w \neq 0}} \frac{\langle w, Av \rangle}{|v||w|} = \sup_{\substack{w \in W \\ w \neq 0}} \sup_{\substack{v \in V \\ v \neq 0}} \frac{\langle w, Av \rangle}{|v||w|}.$$

Now, by definition, $\langle w, Av \rangle = \langle A^* w, v \rangle$. Hence

$$\sup_{\substack{w \in W \\ w \neq 0}} \sup_{\substack{v \in V \\ v \neq 0}} \frac{\langle w, Av \rangle}{|v||w|} = \sup_{\substack{w \in W \\ w \neq 0}} \sup_{\substack{v \in V \\ v \neq 0}} \frac{\langle A^* w, v \rangle}{|w||v|} = \sup_{\substack{w \in W \\ w \neq 0}} \frac{|A^* w|}{|w|} = |A^*|.$$

□

We could also consider more general norms on V. The norm $| \cdot |'$ dual to a norm $| \cdot |$ would be defined by $|w|' = \sup_{v \in V : v \neq 0} |\langle v, w \rangle| / |v|$. Thus, for instance, Hölder's inequality amounts to the statement that the norm dual to the L^r-norm $| \cdot |_r$ on a space of functions is the $L^{r'}$ norm $| \cdot |_{r'}$, where r' is the real such that $1/r + 1/r' = 1$. We can define the operator norm $| \cdot |_{r,s}$ by $\sup_{v \in V : v \neq 0} |Av|_s / |v|_r$, and then we have that $|A|_{r,s} = |A^*|_{s',r'}$, where s' is such that $1/s + 1/s' = 1$.

We will not need to work in such generality. We will just have to work with weighted ℓ^2-norms $|\{a_n\}| = \sqrt{\sum_n |a_n|^2 \eta(n)}$, which, in an abstract framework, such as ours, are just ℓ^2-norms, being defined by an inner product $\sum_n a_n b_n \eta(n)$. Weighted ℓ^2-norms are thus covered by the above; with respect to the inner product we have just described, the norm dual to $|\{a_n\}| = \sqrt{\sum_n |a_n|^2 \eta(n)}$ is itself.

[3] Also called *Fréchet-Riesz representation theorem*.

2.3.3 A few words on derivatives and integrability

Most of the time, we will work with rather general smoothing functions η. We would like to be able to give bounds that depend on η only in so far as they depend on, say, the L^1 norms of η and its first few derivatives – and we would also like not to make unnecessarily strong assumptions on η.

Let us give an example. If $f : \mathbb{R} \to \mathbb{C}$ is continuously differentiable everywhere, then, as we will see in §2.4.1, its Fourier transform $\widehat{f}(t)$ (defined as in (2.8)) obeys the following bound for all $t \neq 0$:

$$\left|\widehat{f}(t)\right| \leq \frac{|f'|_1}{2\pi t}. \tag{2.5}$$

We need not really assume that $|f'|_1$ be finite (or that $t \neq 0$), as then the bound is not false, just void.

What is less obvious is that the class of functions f for which bound (2.5) makes sense and holds is broader than the class C^1 of continuous differentiable functions. Here it is crucial to know to give the right meaning to f' and $|f'|_1$.

For instance, say f is the characteristic function $1_{[0,1]}$. Then $f'(t) = 0$ for all $t \neq 0, 1$, and undefined at $t = 0, 1$. Yet we can prove (2.5) for this choice of f just as for $f \in C^1$, viz., by integration by parts, provided that we understand $|f'|_1$ to be 2.

Why would $|f'|_1$ be 2? A first, intuitive take would be to see f' as being $\delta(x) - \delta(x-1)$, where δ is the Dirac delta function (that is, the "function" $\delta(x)$ that is 0 for $x \neq 0$, and has integral 1). This approach generally gives the right answer, but simply reduces the matter of making sense of f' to that of making sense of the Dirac delta function.

There are several, essentially equivalent ways to truly make sense of $|f'|_1 = 2$ (and of the Dirac delta function). All of them broaden the class of functions f for which the fundamental theorem of calculus (FTC)

$$f(b) - f(a) = \int_a^b f'(t)dt \tag{2.6}$$

(which is often stated just for C^1) holds, provided that $f'(t)dt$ is suitably understood. For such functions f, (2.5) will hold as well.

One approach – perhaps the most familiar to most readers nowadays – is by means of measures and Lebesgue integration. Assume f is right-continuous, that is,

$$\lim_{y \to x^+} f(y) = f(x)$$

for all $x \in \mathbb{R}$. (The same procedure will work if f is left-continuous.) The *Lebesgue-Stieltjes measure* df is defined by $df(I) = f(b) - f(a)$ on intervals $I = (a, b]$, and extends uniquely to Borel sets (see, e.g., [SG77, Ch. 5], or [Hal74, §15]). Then the following version of FTC is a tautology:

$$f(b) - f(a) = \int_a^b df.$$

Now, for any measure μ on a space X, we can define its *total variation* $\|\mu\|$ to be

$$\|\mu\| = \sup \sum_{n=1}^{\infty} |\mu(A_n)|$$

where the supremum is taken over all partitions $\{A_n\}$. It is simple to show that the following version of (2.5) holds:

$$\left|\widehat{f}(t)\right| \leq \frac{\|df\|}{2\pi t}. \tag{2.7}$$

We say that f is of *bounded variation* if $\|df\| < \infty$. If f is C^1, then it is easy to see that $df = f'(t)dt$, and hence $\|df\| = |f'|_1$.

Another approach is based on distribution theory. We can see any locally integrable function $f : \mathbb{R} \to \mathbb{R}$ as a *distribution*, meaning a linear functional from the space of "test functions" (meaning: smooth, compactly supported functions) to $\mathbb{R}$: the distribution given by f sends g to $\int_{-\infty}^{\infty} f(t)g(t)dt$. We define f' to be the functional sending g to $-\int_{-\infty}^{\infty} f(t)g'(t)dt$. In other words, f' is defined so that integration by parts works. Thus, it is unsurprising that (2.5) holds exactly as stated. (Define the L^1 norm of a distribution by taking its supremum over test functions g with $|g|_\infty \leq 1$.) For f right-continuous, the L^1 norm $|f'|_1$ of the distribution f' equals $\|df\|$.

A third, more elementary approach is to use *Riemann-Stieltjes integrals*. That is the approach followed in [MV07]. (See [MV07, Appendix A], or [SG77, Ch. 4].)

Readers may choose their favorite rigorization, or use delta functions and refuse to worry. Simply for convenience, we will often state $|f'|_1 < \infty$ as a condition (meaning: f is of bounded variation) and give bounds in terms of $|f'|_1$, saying (or implying) "in the sense of distributions", but the reader will be fully justified in taking "in the sense of distributions" to mean just "read $\|df\|$ instead of $|f'|_1$".

We would also need to know for which functions f the fundamental theorem of calculus holds exactly as stated in (2.6), with f' interpreted as a function, not in the sense of distributions. Requiring that f be in C^1 would be stronger than necessary, but asking that f be differentiable almost everywhere (that is, outside a set of measure zero) would not be sufficient, as Cantor's "devil's staircase" [Rud74, 7.16(b)] shows. For FTC to make sense and hold as stated in (2.6), it turns out to be necessary and sufficient ([Rud74, Thm. 7.20], [Nie97, Thm. 20.8]) that f be *absolutely continuous*. A function is said to be absolutely continuous if, for every $\epsilon > 0$, there is a $\delta > 0$ such that

$$\sum_{i=1}^{n} |f(y_i) - f(x_i)| < \epsilon$$

for any finite collection of disjoint intervals (x_i, y_i), $1 \leq i \leq n$, such that $\sum_{i=1}^{n}(y_i - x_i) < \delta$.

It is easy to see that, if f is continuous, piecewise C^1 (i.e., continuously differentiable outside a discrete set S of points) and of bounded variation, then f is absolutely continuous.

For the sake of clarity, let us walk through a set of conditions we will see several times. We will often say that we will work with a continuous function f that is piecewise C^1, and such that $f', f'' \in L^1$. Let us see what these conditions entail, given that we are to understand f' and f'' "in the sense of distributions". Here $f' \in L^1$ means that f is of bounded variation, and $f'' \in L^1$ means that f' is of bounded variation. We can assume without loss of generality that f' is right continuous: if it had no limit as $x \to x_0^+$, $x_0 \in S$, then f' could not be of bounded variation even on the interval (x_0, x_1), where x_1 is the first point in S to the right of x_0. Thus, FTC holds for f (in the conventional sense, that is, with f' seen as a function), and it also holds for f' (with f'' understood in the sense of distributions). We write $|f'|_1$ to mean $\|df\|$, and $|f''|_1$ to mean $\|d(f')\|$.

In the end, we will choose to work most of the time with functions in C^∞, or at any event with functions in C^1 with absolutely continuous derivatives, and so all of the above becomes less important: there is then no conceivable ambiguity as to what $|f|_1$, $|f'|_1$, $|f''|_1$ could mean. Still, it is better to make the proper level of generality clear. For instance, the smoothing function in [Tao14] is not smooth, but it does satisfy the conditions we just discussed (continuous, piecewise C^1, and $f', f'' \in L^1$ in the sense of distributions). It is good to know that our work in much of Part III is also valid for such a function.

2.4 THE FOURIER TRANSFORM

2.4.1 The Fourier transform: definition and basic properties

We define the Fourier transform on $\mathbb{R}$ as follows: given a function $f : \mathbb{R} \to \mathbb{C}$ in L^1, its Fourier transform $\widehat{f} : \mathbb{R} \to \mathbb{C}$ is given by

$$\widehat{f}(t) = \int_{-\infty}^{\infty} f(x) e(-xt) dx. \tag{2.8}$$

(As is notorious, different definitions in the literature differ by factors of 2π and the like; obviously, this is purely a matter of convention.)

Convolutions. By Fubini's theorem, for $f, g : \mathbb{R} \to \mathbb{C}$ in L^1,

$$\widehat{f * g} = \widehat{f} \cdot \widehat{g}, \tag{2.9}$$

where $f * g : \mathbb{R} \to \mathbb{C}$ is the (additive) convolution

$$(f * g)(x) = \int_{\mathbb{R}} f(y) g(x - y) dy.$$

Again by Fubini, if $f, g \in L^1$, then $f * g$ is a.e. defined and in L^1.

Fourier inversion formula. Given a function $f : \mathbb{R} \to \mathbb{C}$ such that f and $\widehat{f}$ are both in L^1, then

$$f(x) = \int_{-\infty}^{\infty} \widehat{f}(t) e(xt) dt \quad \text{(Fourier inversion formula)} \tag{2.10}$$

almost everywhere [Rud74, Thm. 9.11]. In particular, if $\widehat{f}$ is identically 0, then $f(x) = 0$ almost everywhere.

It is possible to change the conditions for the Fourier inversion formula somewhat, with corresponding changes in the conclusion. For instance, if $f \in L^1$ is of bounded variation (or even just of bounded variation in some neighborhood of every point), then

$$\frac{f(x^+) + f(x^-)}{2} = \lim_{T\to\infty} \int_{-T}^{T} \widehat{f}(t) e(xt) dt \tag{2.11}$$

for *every* $x \in \mathbb{R}$, with no assumptions needed on $\widehat{f}$ ([Tit48, Ch. I, Thm. 23], [Boc59, Ch. III, Thm. 11]). It is clear that, since f is of bounded variation, $f(x)$ equals the left side of (2.11) for all but a countable number of x_0.

Plancherel's theorem. The Fourier transform is an L^2-isometry: $|\widehat{f}|_2 = |f|_2$ for $f \in L^1 \cap L^2$ [Rud74, Thm. 9.13]. Since one can express an inner product as a linear combination of squares of ℓ^2-norms, it follows easily that, for $f, g \in L^1 \cap L^2$, $\langle \widehat{f}, \widehat{g}\rangle = \langle f, g\rangle$. We apply the name *Plancherel's theorem* both to $|\widehat{f}|_2 = |f|_2$ and to $\langle \widehat{f}, \widehat{g}\rangle = \langle f, g\rangle$. For a discussion on nomenclature, see the end of the section.

The Fourier transform in L^2. Since $L^1 \cap L^2$ is a dense subset of L^2, the fact that the Fourier transform is bounded as an L^2 operator (more than that: an isometry) implies that there is a unique extension of the Fourier transform to L^2 ([Rud74, Thm. 9.13(a)]). This extension is still an isometry. For all $f \in L^2$,

$$\widehat{f}(x) = \lim_{T\to\infty} \int_{-T}^{T} f(t) e(-xt) dt$$

almost everywhere, and, moreover, a Fourier inversion formula

$$f(x) = \lim_{T\to\infty} \int_{-T}^{T} \widehat{f}(t) e(xt) dt$$

also holds almost everywhere [Rud74, Thm. 9.13(d)]. Otherwise put, $\hat{\hat{f}}(x) = f(-x)$ still holds a.e. in L^2.

Fourier inversion and convolution. It follows immediately from (2.9) and (2.10) that, for $f, g : \mathbb{R} \to \mathbb{C}$ in L^1 such that $f \cdot g$, $\widehat{f}$, $\widehat{g}$ are also in L^1,

$$\widehat{f \cdot g} = \widehat{f} * \widehat{g}. \tag{2.12}$$

Just as the conditions for the Fourier transform to hold can be softened or modified, so can the conditions here. For instance, if $f, \widehat{f} \in L^1$ and $g \in L^2$, then, by [SW71, Thm. 2.6], $\widehat{f} * \widehat{g}$ and $f \cdot g$ lie in L^2, and (2.12) still holds almost everywhere.

Affine transformations. Derivatives. Let us go over a couple of properties that we will use later. For $f \in L^1$ and $a, u \in \mathbb{R}$, $u \neq 0$, the Fourier transform of the map

$x \mapsto f(ux+a)$ is

$$\begin{aligned}\int_{-\infty}^{\infty} f(ux+a)e(-xt)dx &= e\left(\frac{a}{u}t\right)\int_{-\infty}^{\infty} f(ux+a)e\left(-\frac{(ux+a)t}{u}\right)dx\\ &= e\left(\frac{a}{u}t\right)\cdot\frac{1}{u}\int_{-\infty}^{\infty} f(y)e\left(-\frac{yt}{u}\right)dy = e\left(\frac{a}{u}t\right)\cdot\frac{\widehat{f}(t/u)}{u}.\end{aligned} \tag{2.13}$$

If $f(x)$ and $xf(x)$ are in L^1, then

$$\widehat{xf}(t) = -\frac{1}{2\pi i}\widehat{f}'(t), \tag{2.14}$$

since, as can be easily verified, we can exchange the order of integration on x and differentiation on t here, just using that $f, xf \in L^1$.

For $f \in L^1$, of bounded variation and right-continuous (or left-continuous),

$$\widehat{f}(t) = \frac{\widehat{f'}(t)}{2\pi i t} \tag{2.15}$$

by integration by parts, where we interpret f' in the sense of distributions. (Use the fact that, since f is in L^1 and of bounded variation, $\lim_{t\to\infty} f(t) = \lim_{t\to-\infty} f(t) = 0$.)

Norms. One can give bounds for the size of $\widehat{f}$ with respect to different norms. The trivial bound is $|\widehat{f}|_\infty \leq |f|_1$. Let $k \geq 1$. Iterating integration by parts, we can show that, provided that f is piecewise C^k and (if $k \geq 2$) in C^{k-2}, and f, f', …, $f^{(k)}$ are all in L^1, then

$$\widehat{f}(t) = O^*\left(\frac{|\widehat{f^{(k)}}|_\infty}{(2\pi t)^k}\right) = O^*\left(\frac{|f^{(k)}|_1}{(2\pi t)^k}\right). \tag{2.16}$$

To be precise: we apply integration by parts $(k-1)$ times in the conventional sense, understanding $f, f', \ldots, f^{k-1}$ as functions, and then apply integration by parts one last time understanding $f^{(k)}$ as a distribution.

Incidentally, (2.16) implies that, if f, f' and f'' are all in L^1, then $\widehat{f}$ is in L^1, and so

$$\frac{f(x^+)+f(x^-)}{2} = \int_{-\infty}^{\infty} \widehat{f}(t)e(xt)dt$$

for every $x \in \mathbb{R}$, with the integral converging absolutely.

Self-duality of the Gaussian. The function $e^{-\pi x^2}$ (the *Gaussian*, in one of its standard normalizations) is its own Fourier transform, and this is true even under a complex shift, after rescaling. This fact can be easily proved as follows: for $z \in \mathbb{C}$ and $f_z(x) = e^{-\pi(x-z)^2}$,

$$\begin{aligned}e^{\pi z^2}\widehat{f}_z(t) &= e^{\pi z^2}\int_{-\infty}^{\infty} e^{-\pi(x-z)^2}e(-xt)dx = e^{-\pi(t+iz)^2}\int_{-\infty}^{\infty} e^{-\pi(x+(it-z))^2}dx\\ &= e^{-\pi(t+iz)^2}\int_{(it-z)-\infty}^{(it-z)+\infty} e^{-\pi u^2}du = e^{-\pi(t+iz)^2}\int_{-\infty}^{\infty} e^{-\pi u^2}du.\end{aligned} \tag{2.17}$$

The last integral is the Gaussian integral, whose value is well-known to be 1. Hence

$$e^{\pi z^2}\widehat{f_z}(t) = e^{-\pi(t+iz)^2} = f_{-iz}(t). \tag{2.18}$$

We obtain as a special case that $\widehat{f}(t) = f(t)$ for $f(t) = f_0(t) = e^{-\pi t^2}$, i.e., the Gaussian is self-dual. We can also use a multiplicative rescaling: using (2.13), we see that, for $f_{\cdot u}(t) = f(\sqrt{u}t) = e^{-\pi u t^2}$ and any $u > 0$,

$$\widehat{f_{\cdot u}} = \frac{f(t/\sqrt{u})}{\sqrt{u}} = \frac{f_{\cdot 1/u}(t)}{\sqrt{u}} = \frac{e^{-\pi t^2/u}}{\sqrt{u}}. \tag{2.19}$$

2.4.2 The Fourier transform in general

We may, of course, define the Fourier transform of a function f defined not over $\mathbb{R}$, but rather over $\mathbb{Z}$ or $\mathbb{R}/\mathbb{Z}$ or over a finite abelian group, say. In the introduction, we discussed functions defined over $\mathbb{Z}$, and how their transforms are functions on $\mathbb{R}/\mathbb{Z}$:

$$\widehat{f}(\alpha) = \sum_{n\in\mathbb{Z}} f(n)e(-\alpha n).$$

The Fourier inversion theorem and Plancherel's theorem are both true in this context, and easier to prove than over $\mathbb{R}$. The Fourier inversion theorem

$$f(n) = \int_{\mathbb{R}/\mathbb{Z}} \widehat{f}(\alpha)e(\alpha n)d\alpha$$

holds whenever f is in ℓ^1.

In general, given a function $f : G \to \mathbb{C}$ over a locally compact abelian group, we may define its Fourier transform $\widehat{f}$ as a function from $\widehat{G}$ to $\mathbb{C}$, where the group $\widehat{G}$ (the *(Pontryagin) dual* of G) is defined to be the space of continuous group homomorphisms from G to $\mathbb{R}/\mathbb{Z}$, endowed with the compact-open topology and a corresponding Haar measure. For instance, $\widehat{\mathbb{Z}} \sim \mathbb{R}/\mathbb{Z}$, $\widehat{\mathbb{R}/\mathbb{Z}} \sim \mathbb{Z}$, $\widehat{\mathbb{R}} \sim \mathbb{R}$, and the dual of a finite abelian group G is isomorphic to G itself (though its total measure may be different, depending on the normalization chosen).

We will not need to work over general locally compact abelian groups. Suffice it to say that some familiar rules hold in full generality. For instance, for any locally compact abelian group G, the transform of a convolution is a product of transforms:

$$\widehat{f * g}(\alpha) = \widehat{f}(\alpha)\cdot\widehat{g}(\alpha), \tag{2.20}$$

provided of course that f and g are in L^1 and that the measure is normalized properly. If G is finite, then every function on G is in L^1, and (2.20) becomes very easy to check: a convolution is then just a finite sum

$$(f * g)(x) = \sum_{\substack{y_1,y_2\in G\\ y_1y_2=x}} f(y_1)g(y_2),$$

and we can write the Fourier transform also as a finite sum, multiplicatively:

$$\widehat{f}(\chi) = \sum_{x\in G} f(x)\overline{\chi(x)}, \tag{2.21}$$

where $\chi : G \to \mathbb{C}^*$ is a group homomorphism. "Multiplicatively" here refers simply to the fact that we write the group law in G as $\cdot$, rather than $+$, and so we work with a multiplicative character χ, rather than an additive character $e(\alpha n)$. We see that

$$\begin{aligned}\widehat{f * g}(\chi) &= \sum_{x\in G}\left(\sum_{\substack{y_1,y_2\in G\\ y_1y_2=x}} f(y_1)g(y_2)\right)\overline{\chi(x)}\\ &= \sum_{y_1\in G} f(y_1)\overline{\chi(y_1)}\cdot\sum_{y_2\in G} f(y_2)\overline{\chi(y_2)} = \widehat{f}(\chi)\cdot\widehat{g}(\chi).\end{aligned} \tag{2.22}$$

The Fourier inversion formula is also easy for G finite. In general, Fourier-based work over finite groups can be more straightforward than over $\mathbb{R}$ or $\mathbb{Z}$, in that one need not bother about convergence problems. We should, however, be a little careful with normalization. Since we define the Fourier transform as in (2.21), the Fourier inversion formula reads

$$f(x) = \frac{1}{|G|}\sum_{\chi\in\widehat{G}}\widehat{f}(\chi)\chi(x),$$

and Plancherel's theorem reads

$$|G|\cdot\sum_{x\in G}|f(x)|^2 = \sum_{\chi\in\widehat{G}}\left|\widehat{f}(\chi)\right|^2. \tag{2.23}$$

Proving either statement is an easy exercise, involving little other than replacing $\widehat{f}$ by its definition. At any rate, we will work almost solely over $\mathbb{Z}$, $\mathbb{R}/\mathbb{Z}$ and $\mathbb{R}$, since it is what our problem naturally demands. Occasionally, we will work over $G = \mathbb{Z}/q\mathbb{Z}$; in that case, characters $\chi : G \to \mathbb{C}$ are maps of the form $a \mapsto e(ab/q)$ with $b \in \mathbb{Z}/q\mathbb{Z}$.

For functions f over $\mathbb{R}/\mathbb{Z}$, Plancherel's theorem is known as *Parseval's theorem*. Parseval worked more than a century before Plancherel, and, in particular, predated Fourier and his inversion formula. Unsurprisingly, Parseval really proved $\langle f, g\rangle = \sum_n a_n\overline{b_n}$ for series $f = \sum_n a_n e(\alpha n)$, $g = \sum_n b_n e(\alpha n)$, $a_n, b_n \in \mathbb{R}$, rather than for general functions $f, g : \mathbb{R}/\mathbb{Z} \to \mathbb{C}$ in L^2; what one can say is that, strictly speaking, he proved Plancherel's theorem for functions defined over $\mathbb{Z}$.

The name "Parseval's theorem" is sometimes applied to Plancherel's theorem over $\mathbb{R}$ or over general G, though the result over $\mathbb{R}$ is due to Plancherel, and the generalization to arbitrary G is due to Pontryagin (and to van Kampen and Weil, who removed an unnecessary assumption). Alternatively, one may apply the name *Plancherel's theorem* to the statement that the Fourier transform is an isometry, and *Parseval's identity* to the algebraic statement $|v|_2^2 = \sum_n\langle v, e_n\rangle$ in a vector space V with an inner product $\langle\cdot,\cdot\rangle$ and an orthonormal basis $\{e_n\}$ if V is complete with respect to the norm $|\cdot|_2$ induced by the inner product. Plancherel's theorem is then proved by Parseval's identity plus some analysis (or just the orthogonality of characters, if the group is finite). Some yet attach the name of Parseval or Plancherel to $|\widehat{f}|_2 = |f|_2$ and that of Plancherel or Parseval to $\langle f, g\rangle = \langle \widehat{f}, \widehat{g}\rangle$, though either statement is an immediate consequence of the other. The entire naming issue can be avoided by identifying Plancherel and Parseval, as in [Bor49], or by metempsychosis. This latter option would only be appropriate, given that Pythagoras's theorem is an application of Parseval's theorem, though one involving a little geometry rather than analysis.

2.5 THE MELLIN TRANSFORM

2.5.1 Basic properties

The *Mellin transform* of a function $\phi : (0, \infty) \to \mathbb{C}$ is

$$M\phi(s) := \int_0^\infty \phi(x) x^{s-1} dx \tag{2.24}$$

whenever the integral makes sense. If $\phi(x)x^{\sigma-1}$ is in L^1 with respect to dx (i.e., $\int_0^\infty |\phi(x)|x^{\sigma-1}dx < \infty$), then the Mellin transform is defined on the line $\sigma + i\mathbb{R}$. Moreover, if $\phi(x)x^{\sigma-1}$ is in L^1 for $\sigma = \sigma_1$ and for $\sigma = \sigma_2$, where $\sigma_2 > \sigma_1$, then it is easy to see that it is also in L^1 for all $\sigma \in (\sigma_1, \sigma_2)$, and that the Mellin transform is holomorphic on $\{s : \sigma_1 < \Re s < \sigma_2\}$ and continuous on $\{s : \sigma_1 \leq \Re s \leq \sigma_2\}$. We then say that $\{s : \sigma_1 < \Re s < \sigma_2\}$ is a *strip of holomorphy* for the Mellin transform. We do say that $\{s : \sigma_1 < \Re s < \sigma_2\}$ is a strip of holomorphy if $M\phi$ is holomorphic on it, even if $M\phi$ does not extend continuously to the closure of the strip.

The Mellin transform becomes a Fourier transform (of $f(v) = \phi(e^v)e^{\sigma v}$) by means of the change of variables $x = e^v$: for $s = \sigma - 2\pi i\tau$,

$$\begin{aligned} M\phi(s) &= \int_0^\infty \phi(x)x^{s-1}dx = \int_{-\infty}^\infty \phi\left(e^v\right) e^{v(s-1)} e^v dv \\ &= \int_{-\infty}^\infty \phi\left(e^v\right) e^{v\sigma} e(-v\tau) dv = \int_{-\infty}^\infty f(v) e(-v\tau) dv = \widehat{f}(\tau). \end{aligned} \tag{2.25}$$

Mellin inversion formula. As a consequence, we obtain from (2.10) that, for ϕ such that $\phi(x)x^{\sigma-1}$ is in L^1 and $t \mapsto (M\phi)(\sigma + it)$ is also in L^1,

$$\phi(x) = \frac{1}{2\pi i} \int_{\sigma - i\infty}^{\sigma + i\infty} (M\phi)(s) x^{-s} ds \qquad \text{(Mellin inversion formula)} \tag{2.26}$$

almost everywhere. We also obtain from (2.11) that, if $\phi(x)x^{\sigma-1}$ is in L^1 and $\phi(x)x^\sigma$ has bounded variation in a neighborhood of every point,

$$\frac{\phi(x^+) + \phi(x^-)}{2} = \frac{1}{2\pi i} \cdot \lim_{T\to\infty} \int_{\sigma - iT}^{\sigma + iT} (M\phi)(s) x^{-s} ds \tag{2.27}$$

for every $x_0 > 0$. It is clear that, for $\phi \in L^1$, the function $\phi(x)x^\sigma$ has bounded variation in a neighborhood of every point if and only if $\phi(x)$ does.

Isometry. We also obtain that the Mellin transform is an isometry, in the sense that

$$\int_0^\infty |\phi(x)|^2 x^{2\sigma} \frac{dx}{x} = \frac{1}{2\pi} \int_{-\infty}^\infty |M\phi(\sigma+it)|^2 dt. \tag{2.28}$$

Recall that, in the case of the Fourier transform, for $|\widehat{\phi}|_2 = |\phi|_2$ to hold, it is enough that ϕ be in $L^1 \cap L^2$. Thus, for (2.28) to hold, it is enough that $\phi(x)x^{\sigma-1}$ be in L^1 and $\phi(x)x^{\sigma-1/2}$ be in L^2 (again, with respect to dx, in both cases). We will refer to (2.28) as “Plancherel”, just as for the Fourier transform.

Convolutions. We write $f *_M g$ for the multiplicative, or Mellin, convolution of f and g:

$$(f *_M g)(x) = \int_0^\infty f(w) g\left(\frac{x}{w}\right) \frac{dw}{w}. \tag{2.29}$$

By a change of variables the convolution rule (2.9) for the Fourier transform,

$$M(f *_M g) = Mf \cdot Mg \tag{2.30}$$

on the intersection of the regions on which Mf and Mg are defined. Moreover,

$$M(f \cdot g)(s) = \frac{1}{2\pi i} \int_{\sigma-i\infty}^{\sigma+i\infty} Mf(z) Mg(s-z) dz \tag{2.31}$$

for $\Re s = \sigma + \sigma'$, provided that $f(x)x^{\sigma-1}$, $g(x)x^{\sigma'-1}$ and $f(x)g(x)x^{\sigma+\sigma'-1}$ are in $L^1((0,\infty))$, Mf is in L^1 on $\sigma + i\mathbb{R}$ and Mg and in L^1 on $\sigma' + i\mathbb{R}$, so that the conditions for the Mellin inversion formula are fulfilled.

In fact, (2.31) follows from the following general statement together with Mellin inversion. Let $F_1 \in L^1(\sigma_1 + i\mathbb{R})$, $F_2 \in L^1(\sigma_2 + i\mathbb{R})$. Define F_3 as the convolution

$$F_3(s) = \frac{1}{2\pi i} \int_{\sigma_1 - i\infty}^{\sigma_1 + i\infty} F_1(z) F_2(s-z) dz$$

for $\Re s = \sigma_3 = \sigma_1 + \sigma_2$. Let $f_1, f_2, f_3 : (0,\infty) \to \mathbb{C}$ be defined as inverse Mellin transforms

$$f_i = \frac{1}{2\pi i} \int_{\sigma_i - i\infty}^{\sigma_i + i\infty} (MF_i)(s) x^{-s} ds.$$

Then, by the convolution rule (2.9) for the Fourier transform,

$$f_3 = f_1 \cdot f_2. \tag{2.32}$$

Transformation rules. Examples of Mellin transforms. We also have several useful transformation rules, just as for the Fourier transform. For example,

$$\begin{aligned}
&M(f(ux))(s) = u^{-s} Mf(s) \qquad \text{for } u > 0 \text{ real},\\
&M(xf(x))(s) = Mf(s+1)\\
&M(xf'(x))(s) = -s \cdot Mf(s),\\
&M((\log x) f(x))(s) = (Mf)'(s),\\
&M\left(x^{s_0} f(x)\right)(s) = (Mf)(s+s_0),
\end{aligned} \tag{2.33}$$

as is easy to check from the definition. The third rule here deserves some further comment. As usual, $f'(x)$ is to be understood in the sense of distributions. For this rule to hold for s with $\Re s = \sigma$, it is enough that both sides of the equation be well-defined, that is, it is enough that $f'(x)x^\sigma$ and $f(x)x^{\sigma-1}$ be in L^1: it is an exercise to show that, if $f(x)x^{\sigma-1} \in L^1$ holds, $f(x)x^\sigma \to 0$ for some sequence of x tending to 0 and also for some sequence tending to ∞, and then the rule follows by integration by parts.

Here are a few other useful Mellin transforms, all of them easy to derive.

function f	Mellin transform Mf	parameters	strip of holomorphy
$1_{[0,u]}$	u^s/s	$u > 0$	$\Re s > 0$
$\frac{1}{k!}\left(\log^+ \frac{u}{x}\right)^k$	u^s/s^{k+1}	$u > 0$	$\Re s > 0$
$\frac{(u-x)^k}{k!}1_{[0,u]}$	$\frac{u^{s+k}}{s(s+1)\cdots(s+k)}$	$u > 0$	$\Re s > 0$

For further examples, see, e.g., [BBO10, Table 11.3]. The Mellin transform of e^{-x} is the *Gamma function* $\Gamma(s)$, which we shall study in §3.4.

2.5.2 The Mellin transform and Dirichlet series

The Mellin transform is useful in number theory in part because it establishes a relation between a sum over the positive integers $\sum_n a_n f(n)$, $a_n \in \mathbb{C}$, where $f : [0,\infty) \to \mathbb{C}$ is a weight, and a *Dirichlet series* $\alpha(s) = \sum_n a_n n^{-s}$. Assume a_n, f and σ are such that $x \mapsto \sum_n |a_n||f(nx)|x^{\sigma-1}$ is in L^1 with respect to dx. Then, by the first rule in (2.33), the Mellin transform of $x \mapsto \sum_n a_n f(nx)$ is

$$\sum_n a_n M(f(nx)) = \sum_n a_n n^{-s} \cdot Mf(s) = \alpha(s) \cdot Mf(s).$$

Let us see two examples.

- Define $A(y) = \sum_{n \le y} a_n$. Clearly, $A(1/x) = \sum_n a_n 1_{[0,1/n]}(x)$. Since the Mellin transform of $1_{[0,u]}$ is $s \mapsto u^s/s$, it follows that the Mellin transform of $A(1/x)$ is $\alpha(s)/s$. The transform is defined for every $s = \sigma + it$ with $\sigma > 0$ large enough for $\sum_n |a_n| n^{-\sigma}$ to converge.

  Since $t \mapsto x^{-(\sigma+it)}/(\sigma+it)$ is not in L^1 for any σ, the conditions for the Mellin inversion formula in the form (2.26) do not hold; however, those for (2.27) do. Hence, for $\sigma > 0$ is large enough for $\sum_n |a_n| n^{-\sigma}$ to converge, and any $x > 0$,

  $$\sum_{n \le 1/x}{}' a_n = \lim_{T \to \infty} \frac{1}{2\pi i} \int_{\sigma - iT}^{\sigma + iT} \frac{\alpha(s)}{s} x^{-s} ds, \qquad \text{(Perron's formula)}$$

  where

  $$\sum_{n \le y}{}' a_n := \begin{cases} \frac{a_n}{2} + \sum_{n<y} a_n & \text{if } y \in \mathbb{Z}^+, \\ \sum_{n<y} a_n & \text{otherwise.} \end{cases}$$

- Continuous weights f make matters more pleasant. For instance, let $R_k(x) = (1/k!) \sum_n a_n (\log^+(x/n))^k$, $k \geq 1$. Since the Mellin transform of
$$x \mapsto \frac{1}{k!}\left(\log^+ \frac{u}{x}\right)^k$$
is $s \mapsto u^s/s^{k+1}$, the Mellin transform of $x \mapsto R_k(1/x)$ is $s \mapsto \alpha(s)/s^{k+1}$. Now, for $k \geq 1$, $t \mapsto (1/(\sigma + it)^{k+1})$ *is* in L^1 for $\sigma > 0$ fixed, and so (2.26) does apply:
$$R_k(1/x) = \frac{1}{2\pi i}\int_{\sigma - i\infty}^{\sigma + i\infty} \frac{\alpha(s)}{s^{k+1}} x^{-s} ds \tag{2.34}$$
for $\sigma > 0$ larger than the abscissa of absolute convergence, i.e., for $\sigma > 0$ large enough for $\sum_n |a_n| n^{-\sigma}$ to converge.

2.6 COMPLEX-ANALYTIC TOOLS

We will take the commonalities of basic courses in complex analysis (analytic continuation, Cauchy's theorem, the bounded-modulus principle, Liouville's theorem, and so forth) as read. Let us quickly go over some material that is specially useful in analytic number theory.

2.6.1 Writing a function in terms of its zeros

2.6.1.1 Estimates based on Jensen's formula

Let $f(s)$ be analytic on an open domain containing the closed disk $|s| \leq R$. Let $s_1, s_2, \ldots, s_n$ be the zeros of $f(z)$ on the same closed disk. Then
$$\log|f(0)| = -\sum_{i=1}^{n} \log \frac{R}{|s_1|} + \frac{1}{2\pi R}\int_{|s|=R} \log|f(s)|\,|ds| \quad \text{(Jensen's formula)} \tag{2.35}$$
See the proof in, e.g., [Ahl78, §5.3.1]. (The main idea is that, if $f(s)$ is free of zeros on the closed disk, then $\Re \log f(s) = \log|f(s)|$ is harmonic, and so (2.35) holds. If $f(s)$ does have zeros, then it can first be multiplied by a factor (the inverse of a *Blaschke product*) so that the zeros disappear but the value of $|f(s)|$ for $|s| = R$ does not change.) The following immediate consequence goes by[4] "Jensen's inequality" in [MV07, §6.1].

Corollary 2.4. *Let $f(s)$ be analytic on an open domain containing the disk $|s| \leq R$. Assume $|f(0)| \neq 0$. Then, for any $r < R$, the number of zeros of $f(s)$ with $|s| \leq r$ is at most*
$$\frac{\log \max_{|s|=R} |f(s)|/|f(0)|}{\log R/r}.$$

[4] There might be a risk of confusion with Jensen's inequality on norms – but then "Cauchy's theorem" and "Liouville's theorem" can also mean other things in other fields.

Corollary 2.4 allows us to derive bounds on the number of zeros of a function f from bounds on the size of its values $f(s)$. We will also need bounds on the size of $f'(s)/f(s)$ based on information on the number and location of the zeros of f. Lemma 2.6, part of an approach introduced by Landau (see [Tit86, §III.3.9]), will give us such bounds.

We first need a standard lemma that will also be useful elsewhere. By contour integration, it is easy to tell the value of $f(s)$, $f'(s)$, etc., inside a circle C from the values of $f(s)$ on the circle C. As it turns out, one can bound $|f(s)|$, $|f'(s)|$, etc., inside C given only a bound on the *real* part $\Re f(s)$ of $f(s)$ on C.

Lemma 2.5 (Borel-Carathéodory)**.** *Let f be analytic on an open domain containing the disk $|s| \leq R$, with $f(0) = 0$. Assume $\Re f(s) \leq A$ for all s with $|s| = R$. Then, for $k \geq 0$ and $|s| < R$,*

$$|f^{(k)}(s)| \leq \frac{2k!AR}{(R - |s|)^{k+1}}$$

Proof. See [Ten15, Thm. II.3.16] or [MV07, Lem. 6.2]. In brief: write $f(s)$ as a Taylor series $\sum_{n=1}^{\infty} a_n s^n$, and show that

$$|a_n|R^n = \frac{1}{\pi}\int_0^{2\pi} (1 + \cos(n\theta + \theta_n)) \cdot \Re f(Re^{i\theta}) d\theta, \tag{2.36}$$

where $\theta_n = \arg a_n$. The right side of (2.36) is at most $2A$. Finish by applying $\sum_{n\geq k}\binom{n}{k}\alpha^{n-k} = 1/(1-\alpha)^{k+1}$ with $\alpha = |s|/R$. □

Lemma 2.6. *Let f be analytic on the disk $|s| \leq 2R$, and non-zero at $s = 0$. Let M be the maximum of $|f(s)/f(0)|$ on the disk $|s| \leq 2R$, and let $s_1, s_2, \dots, s_n$ be the zeros of f on the disk $|s| \leq R$. Then, for all s with $|s| < R$,*

$$\frac{f'(s)}{f(s)} = \sum_{i=1}^{n} \frac{1}{s - s_i} + O^*\left(\frac{2R\log M}{(R - |s|)^2}\right) \tag{2.37}$$

for all s with $|s| \leq r$.

Proof. Let $g(s) = (f(s)/f(0)) \cdot \prod_{i=1}^{n}(1 - s/s_i)^{-1}$. Then $|1 - s/s_i| \geq 1$ for $|s| = 2R$ and all i, so $|g(s)| \leq M$ for $|s| = 2R$, and thus $|g(s)| \leq M$ for $|s| < 2R$ as well. Since $g(s)$ is free of zeros for $|s| \leq R$, the function $\log g(s)$ is single-valued for $|s| \leq R$. Apply Lemma 2.5 (with $k = 1$) to $\log g(s)$ instead of $f(s)$, with $A = \log M$. □

2.6.1.2 In the manner of Weierstrass

In general it is possible to give an exact expression for an analytic function $f(s)$ or for $f'(s)/f(s)$ in terms of the zeros of $f(s)$, provided that we have a bound of a certain strength on f, possibly together with some additional conditions.

The simplest case may be that of polynomials.

Exercise 2.7. Let f be an entire function such that $|f(s)| \leq |s|^C$ for some C. Show $f(s)$ must be a polynomial of degree $\leq C$. (Hint: use the fact that every bounded entire function is constant.) Conclude that thus $f(s)$ is of the form $K\prod_\rho(s-\rho)$, where ρ goes over the zeros of f, with multiplicity.

In order to prove more general results, we first show that an entire function can be forced to be constant by growth conditions much milder than being constant, if some other conditions are met.

Lemma 2.8. *Let f be an entire function such that $f(s+1) = f(s)$. Suppose that $|f(s)| \ll e^{(2-\epsilon)\pi|s|}$ for some $\epsilon > 0$. Then f is constant.*

Proof. Since $f(s+1) = f(s)$, we can write $f(s)$ in the form $f(s) = g(z)$, where $z = e^{2\pi i s}$ and g is analytic on its domain of definition $\mathbb{C}\setminus 0$. By $|f(s)| \ll e^{(2-\epsilon)\pi|s|}$, we know that, as $z \to 0$ and as $z \to \infty$, $g(z)$ is bounded by $1/|z|^{1-\epsilon}$ and by $\sqrt{z}$, respectively. Hence, the singularities at 0 and ∞ are both removable, and so $f(z)$ can be extended to a bounded entire function. A bounded entire function must be constant, and so we are done. □

We typically use Lemma 2.8 to show that two functions that have poles at the same places to the same order are the same function, up to a constant factor.

Exercise 2.9. (Euler[5]) Prove that

$$\pi\cot\pi s = \frac{1}{s} + \sum_{n=1}^{\infty}\left(\frac{1}{n+s} - \frac{1}{n-s}\right). \tag{2.38}$$

(Hint: multiply both sides by $\sin\pi s$ first.)

Lemma 2.10. *Let $f(s)$ be a nowhere vanishing entire function. Assume that there is a $\rho < 2$ such that, for every large enough r, there is an $R \in [r, 2r]$ such that $|f(s)| \le e^{R^\rho}$ for all s with $|s| = R$. Then $f(s) = e^{A+Bs}$ for some $A, B \in \mathbb{C}$.*

Proof. The function $\log f(s)$ is a well-defined (single-valued) entire function. Given any s_0, we choose R as in the statement with $r = 2|s_0|$. Then $\Re\log f(s) = \log|f(s)| \ll |R|^\rho$ for $|s| = R$. By Lemma 2.5 (Borel-Carathéodory), it follows that $|\log f(s_0)| \ll R^\rho \ll |s_0|^\rho$. Since $\log f(s_0)$ is entire, it follows that $\log f(s_0)$ is a linear polynomial on s_0. □

Lemma 2.11 (Hadamard)**.** *Let $f(s)$ be an entire function such that, for some $\rho < 2$, $|f(s)| \ll e^{|s|^\rho}$ as $|s| \to \infty$. Assume $f(0) \neq 0$. Let $s_1, s_2, \dots$ be the zeros of $f(s)$. Then*

$$f(s) = e^{A+Bs}\prod_{i=1}^{\infty}\left(1 - \frac{s}{s_i}\right)e^{s/s_i} \tag{2.39}$$

for some $A, B \in \mathbb{C}$. The product here converges uniformly on compact sets.

[5] Of course, we live after Weierstrass and Euler did not. He often proceeded less than rigorously to derive what turned out to be correct statements, which were then justified in other ways – sometimes by others (e.g. N. Bernoulli), sometimes by himself. See [Var06, §3.3 and 3.5].

Proof. By Cor. 2.4, the number of zeros s of f with $|s| \leq R$ is $\ll R^\rho$. Hence $\sum_{i=1}^n |s_i|^{-2}$ converges, and thus the infinite product

$$P(s) = \prod_{i=1}^{\infty} \left(1 - \frac{s}{s_i}\right) e^{s/s_i}$$

converges as well. Hence $F(s) = f(s)/P(s)$ is a nowhere vanishing entire function.

The rest is given in detail in [MV07, proof of Lem. 10.11] or [Ten15, proof of Thm. 3.20]. One first shows that, for any r, there is an $R \in [r, 2r]$ not too close to any root (pigeonhole) and then that $|P(s)| \gg e^{-O(R^\rho \log R)}$ for $|s| = R$. Then one applies Lemma 2.10 (with any $\rho' \in (\rho, 2)$ instead of ρ), and obtains that $F(s) = e^{A+Bs}$. □

It follows immediately from (2.39) that

$$\frac{f'(s)}{f(s)} = B + \sum_{i=1}^{\infty} \left(\frac{1}{s - s_i} + \frac{1}{s_i}\right). \tag{2.40}$$

2.6.2 Phragmén-Lindelöf. Hadamard's three-line theorem.

We recall that the maximum modulus principle states that, if a function f is holomorphic inside a connected bounded open domain Ω and continuous on its closure $\overline{\Omega}$, and satisfies $|f(z)| \leq C$ on the boundary $\partial\Omega$, then $|f(z)| \leq C$ for all $z \in \Omega$. If Ω is unbounded, then the conclusion may not be true. Let, for instance, $\Omega = (-\pi/2, \pi/2) + i\mathbb{R}$. Then $f(z) = \exp(\exp(iz))$ satisfies $|f(z)| = 1$ on $\partial\Omega$, but it is unbounded inside Ω.

The idea of the Phragmén-Lindelöf principle is that one can in fact bound $f(z)$ in an unbounded region Ω given its values on $\partial\Omega$, provided that we are also given a growth condition. There are many variants; the following is a basic statement for vertical strips.

Lemma 2.12 (Phragmén-Lindelöf)**.** *Let $\Omega = (-\pi/2, \pi/2) + i\mathbb{R}$. Let f be holomorphic on Ω and continuous on $\overline{\Omega}$. Assume that $|f(z)| \leq 1$ for all $z \in \partial\Omega$. Assume as well that there are constants $0 \leq \mathscr{r} < 1$ and c such that*

$$|f(z)| < \exp(c \exp(\mathscr{r} \Im z)) \tag{2.41}$$

for all $z \in \Omega$. Then $|f(z)| \leq 1$ for all $z \in \Omega$.

The proof is a staple of texts in complex analysis; see, e.g., [Rud74, Thm. 12.9].

Proof. Let $\beta \in (\mathscr{r}, 1)$. For $\epsilon > 0$, define an auxiliary function $h_\epsilon(z) = e^{-\epsilon \cos(\beta z)}$. Then $|h_\epsilon(z)| < 1$ for $z \in \Omega$, and it is also easy to check that, thanks to condition (2.41), $f(z)h_\epsilon(z) \to 0$ within Ω as $|\Im z| \to \infty$. We can then choose y_0 such that $|f(z)h_\epsilon(z)| \leq 1$ for $z \in \Omega$ with $|\Im z| \geq y_0$, and then apply the maximum modulus principle to the domain $\{z \in \Omega : |\Im z| < y_0\}$ to conclude that $|f(z)h_\epsilon(z)| \leq 1$ there as well. Thus $|f(z)h_\epsilon(z)| \leq 1$ for all $z \in \Omega$. Since, for any z, $h_\epsilon(z) \to 1$ as $\epsilon \to 0$, we are done. □

We can give a variant for sectors, as in, e.g., [Tit39, §5.61] or [GS68, v. 2, §IV.7.2].

Lemma 2.13 (Phragmén-Lindelöf for sectors)**.** *Let* $\Omega = \{z \in \mathbb{C} : \alpha < \arg z < \beta\}$ *for some* $\alpha, \beta \in \mathbb{R}$ *with* $0 < \beta - \alpha < 2\pi$*. Let* f *be holomorphic on* Ω *and continuous on* $\overline{\Omega}$*. Assume that* $|f(z)| \leq C$ *for all* $z \in \partial\Omega$*. Assume as well that there are constants* $0 \leq \mathscr{r} < \pi/(\beta - \alpha)$ *and* c *such that* $|f(z)| \leq C \exp(c|z|^{\mathscr{r}})$ *for all* $z \in \Omega$*. Then* $|f(z)| \leq C$ *for all* $z \in \Omega$*.*

Exercise 2.14. Proceeding as in the proof of Lemma 2.12, prove Lemma 2.13.

We can bound $f(z)$ within a strip even if the function is not bounded on its boundary, provided that we have some control on the growth of $f(z)$ on the boundary. The following statement is used frequently in analytic number theory; it is referred to as a *convexity bound*, since, as we will see, it states that a certain exponent is a convex function of the abscissa x. We will see subconvexity bounds in the context of zeta functions in §3.5.4.

Lemma 2.15 (Hadamard's three-line theorem, generalized)**.** *Let* $\Omega = (a, b) + i\mathbb{R}$*. Let* f *be holomorphic on* Ω *and continuous on* $\overline{\Omega}$*. Assume that* $|f(a + it)| \leq C(|t| + 1)^{\alpha}$ *and* $|f(b+it)| \leq C(|t|+1)^{\beta}$ *for all* $t \in \mathbb{R}$*. Assume as well that condition (2.41) holds for some* c *and* $0 \leq \mathscr{r} < \pi/(b - a)$*. Then, for any* $0 \leq u \leq 1$ *and* $x_u = ua + (1 - u)b$*,*

$$|f(x_u + it)| \ll C(|t| + 1)^{u\alpha+(1-u)\beta}$$

for all $t \in \mathbb{R}$*.*

The proof and the statement are essentially as in [Tit39, §5.65], except we use the more relaxed growth condition (2.41), as in [Rud74, Thm. 12.9].

Proof. We can assume without loss of generality that $a = -\pi/2$ and $b = \pi/2$; the rescaling involved gives us $0 \leq \mathscr{r} < 1$. For $\Im z \geq 1$, let

$$g(z) = (-iz)^{\alpha \frac{b-z}{b-a} + \beta \frac{z-a}{b-a}}, \tag{2.42}$$

where $s_1^{s_2} = e^{s_2 \log s_1}$ is defined by the principal branch of log. Proceed as in Lemma 2.12, using the function $f(z)/g(z)$ instead of $f(z)$, and the region $(-\pi/2, \pi/2) + i[1, \infty)$ instead of $(-\pi/2, \pi/2) + i\mathbb{R}$.

For $\Im \leq -1$, do the same, only with iz in (2.42) instead of $-iz$. □

Chapter Three

Series and summation

We shall go over summation techniques and some simple kinds of sums. We will also quickly review the essentials on the Riemann zeta function and Dirichlet functions.

There are a couple of places where we go beyond what can be readily found in other texts. In §3.2, we derive bounds on sums of the form $\sum_n f(n)e(\alpha n)$, f continuous; oddly enough, some of them (Lemma 3.4 and part of what follows) seem to be new, though the main tool is due to Euler. Explicit L^2 bounds on the decay of $\zeta(s)$ such as those we will cover in §3.5.4 are a little hard to find. Other than that, much of the material in the chapter should be familiar to most readers.

3.1 SUMMATION FORMULAE

3.1.1 Sweet memories of childhood

If $f : (0, \infty) \to \mathbb{R}$ is non-increasing, then

$$\sum_{n\leq x} f(n) \leq \int_0^x f(t)dt \quad \text{and} \quad \sum_{n=1}^{\infty} f(n) \leq \int_0^{\infty} f(t)dt$$

for any $x \geq 0$, in each case assuming that the integral converges. If $f : [1, \infty) \to \mathbb{R}$ is non-decreasing, then

$$\sum_{n\leq x} f(n) \leq \int_1^x f(t)dt + f(x).$$

If $f : (0, \infty) \to \mathbb{R}$ is convex, then

$$\sum_{\substack{n\in\mathbb{Z}^+\\ n \text{ odd}}} f(n) \leq \frac{1}{2}\int_0^{\infty} f(t)dt \tag{3.1}$$

provided that the integral on the right converges, since

$$f(n) \leq \frac{1}{2}\int_{n-1}^{n+1} f(s)ds.$$

Moreover, if $f(t)$ is convex, non-increasing and non-negative for $0 \leq t \leq x$, then

$$\sum_{\substack{n\leq x\\ n \text{ odd}}} f(n) \leq \frac{1}{2}\int_0^x f(t)dt + \frac{f(x)}{2} \tag{3.2}$$

for any $x \geq 0$, since

$$\sum_{\substack{n\leq x\\ n \text{ odd}}} f(n) \leq \frac{1}{2}\int_0^{2\lfloor (x+1)/2\rfloor} f(t)dt$$

and

$$\int_x^{2\lfloor (x+1)/2\rfloor} f(t)dt \leq \int_x^{2\lfloor (x+1)/2\rfloor} f(x)dt \leq 1\cdot f(x) \leq f(x).$$

3.1.2 The Euler-Maclaurin formula

Let $f : \mathbb{R}\to\mathbb{C}$ be of bounded variation. We are all familiar with the simplest way to approximate the integral of f: we can approximate it by a sum – over the integers, say. We can then ask ourselves what the error term might be. Here is a simple bound: by FTC,

$$\begin{aligned}\int_{n-1/2}^{n+1/2} f(x)dx &= \int_{n-1/2}^{n+1/2}\left(f(n)+\int_n^x f'(t)dt\right)dx\\ &= f(n)+\int_n^{n+1/2} f'(t)\left(\int_t^{n+1/2} dx\right)dt - \int_{n-1/2}^{n} f'(t)\left(\int_{n-1/2}^{t} dx\right)dt\\ &= f(n) - \int_{n-1/2}^{n+1/2} f'(t)B_1(\{t\})\,dt,\end{aligned}$$

where $B_1(x) = x-1/2$, we see that, for a, b integers with $a\leq b$,

$$\int_{a-1/2}^{b+1/2} f(x)dx = \sum_{n=a}^{b} f(n) - \int_{a-1/2}^{b+1/2} B_1(\{x\})f'(x)dx. \tag{3.3}$$

There are several ways in which we can generalize (3.3). We may let $a\to-\infty$ and $b\to\infty$, and obtain

$$\begin{aligned}\sum_{n\in\mathbb{Z}} f(n) &= \int_{\mathbb{R}} f(x)\,dx + \int_{\mathbb{R}} B_1(\{x\})f'(x)dx\\ &= \int_{\mathbb{R}} f(x)dx + O^*\left(\frac{|f'|_1}{2}\right)\end{aligned} \tag{3.4}$$

for any function $f:\mathbb{R}\to\mathbb{C}$ such that f, f' are in L^1.

We can also iterate the procedure, provided that we make enough assumptions on f for higher derivatives of f to make sense. Most of the time, we will have enough with one iteration. What we obtain is the following.

Lemma 3.1 (Euler-Maclaurin formula, second order)**.** *Let $f:\mathbb{R}\to\mathbb{C}$ be a continuous, piecewise C^1 function such that f, f', f'' are in L^1. Then*

$$\sum_{n=-\infty}^{\infty} f(n) = \int_{-\infty}^{\infty} f(x)dx - \frac{1}{2}\int_{-\infty}^{\infty} B_2(\{x\})f''(x)dx, \tag{3.5}$$

where $B_2(x) = x^2 - x + 1/6$. *Moreover,*

$$\sum_{n=-\infty}^{\infty} f(n) = \int_{-\infty}^{\infty} f(x)dx + O^*\left(\frac{1}{16}|f''|_1\right). \tag{3.6}$$

A thoughtless application of (3.5) would give $1/12$ instead of $1/16$ in (3.6).

As we discussed in §2.3.3, when we say that f' and f'' are in L^1, we mean they are in L^1 as distributions, that is to say: the functions f and f' are of bounded variation. The integral in (3.5), just like the L^1 norm in (3.6) is to be understood as the integral of a distribution, with contributions from the singularities of f'' (i.e., the discontinuities of f'). Otherwise put, $f''(x)dx$ means df', and $|f''|_1$ denotes the total variation of f.

Proof. First of all, notice that, because f' is in L^1, $f(t)$ tends to a limit as $t \to \infty$ or $t \to -\infty$. Since f is in L^1, that limit is 0. By the same reasoning, using the facts that f' and f'' are in L^1, we also obtain that $\lim_{t\to-\infty} f'(t) = \lim_{t\to\infty} f'(t) = 0$.

We can assume without loss of generality that f' is continuous at the integers, simply by shifting f by a tiny amount; doing so changes both sides of (3.5) and (3.6) by something tiny.

Now let $F(t)$ be such that $F'(t) = B_1(t)$ everywhere. Then, by integration by parts,

$$\begin{aligned}\int_n^{n+1} B_1(\{x\})f'(x)dx &= \int_0^1 B_1(t)f'(n+t)dt \\ &= F(1)f'(n+1) - F(0)f'(n) - \int_0^1 F(t)f''(n+t)dt.\end{aligned}$$

Since $\int_0^1 B_1(x) = 0$, we have $F(0) = F(1)$. Hence

$$\begin{aligned}\sum_{n=-\infty}^{\infty} f(n) &= \int_{-\infty}^{\infty} f(x)dx + F(0) \lim_{\substack{a\to-\infty \\ b\to\infty}} (f'(b) - f'(a)) - \int_{-\infty}^{\infty} F(\{x\})f''(x)dx \\ &= \int_{-\infty}^{\infty} f(x)dx - \int_{-\infty}^{\infty} F(\{x\})f''(x)dx.\end{aligned}$$

We obtain (3.5) and (3.6) by noting that $F(t) = B_2(t)/2$ and $F(t) = (x^2 - x + 1/8)/2$ are valid choices (i.e., $F'(t) = B_1(t)$ in either case), and that the maximum of $|(x^2 - x + 1/8)/2|$ for $0 \le x \le 1$ is $1/16$. □

In general, we define the *kth Bernoulli polynomial* $B_k(t)$ by

$$B_0(t) = 1, \qquad B'_{j+1}(t) = (j+1)B_j(t), \qquad \int_0^1 B_j(t)dt = 0.$$

We can now state the Euler-Maclaurin formula for arbitrary order. The proof is just like that of Lemma 3.1, with induction on k; see, e.g., [MV07, App. B].

Lemma 3.2 (Euler-Maclaurin formula, arbitrary order)**.** *Let $k \geq 1$, $a, b \in \mathbb{R}$, $a < b$. Let $f : [a, b] \to \mathbb{C}$ be C^{k-2} (for $k > 1$), piecewise C^{k-1}, and such that $f, f', \ldots, f^{(k)}$ are in L^1 and $f^{(k-1)}$ is well-defined at a and b. Then*

$$\sum_{a<n\leq b} f(n) = \int_a^b f(x)dx + \sum_{j=1}^{k} \frac{(-1)^j}{j!} \Big(B_j(\{b\})f^{(j-1)}(b) - B_j(\{a\})f^{(j-1)}(a)\Big) + \frac{(-1)^{k+1}}{k!} \int_a^b B_k(\{x\})f^{(k)}(x)dx. \tag{3.7}$$

Yet again: when we say that $f^{(k)}$ (say) is in L^1, we mean that this is so even when $f^{(k)}$ is seen as a distribution, i.e., the function $f^{(k-1)}$ is of bounded variation. On another note – the same trick would work here as in Lemma 3.1: we may replace B_k (and only B_k) by any other function $F(t)$ such that $F'(t) = kB_{k-1}(t)$.

Of course one may send $a \to -\infty$ or $b \to \infty$ in (3.7). For a an integer, letting $b \to \infty$, and assuming that $\lim_{b\to\infty} f^{(i)}(b) = 0$ for every $0 \leq i \leq k-1$, we obtain that

$$\sum_{n>a} f(n) = \int_a^\infty f(x)dx - \sum_{j=1}^{k} \frac{(-1)^j B_j}{j!} f^{(j-1)}(a) + \frac{(-1)^{k+1}}{k!} \int_a^\infty B_k(\{x\})f^{(k)}(x)dx,$$

where we define $B_j = B_j(0)$. The constants B_j are called *Bernoulli numbers*. It is easy to show that $B_1 = -1/2$, and that $B_j = 0$ for $j > 2$ odd. Thus, for a an integer and $k \geq 1$, we may also write

$$\sum_{n\geq a} f(n) = \int_a^\infty f(x)dx - \sum_{j=1}^{k} \frac{B_j}{j!} f^{(j-1)}(a) + \frac{(-1)^{k+1}}{k!} \int_a^\infty B_k(\{x\})f^{(k)}(x)dx. \tag{3.8}$$

We already know that $B_2 = 1/6$. The next values are $B_4 = -1/30$ and $B_6 = 1/42$. By (say) [MV07, Cor. B.3 and B.4], if k is even, then $|B_k(x)| \leq |B_k|$ for all $0 \leq x < 1$, and, for every $k \geq 2$,

$$\max_{x\in[0,1]} |B_k(x)| = \frac{2\zeta(k)}{(2\pi)^k} k! \tag{3.9}$$

We can also give versions of all of these formulae with summation over the odd integers only, thanks to the fact that odd numbers are an arithmetic progression. By (3.4), for any $f : \mathbb{R} \to \mathbb{C}$ such that $f, f' \in L^1$,

$$\begin{aligned} \sum_{n \text{ odd}} f(n) = \sum_n g(n) &= \int_{-\infty}^{\infty} g(x)dx + O^*(|g'(x)|_1/2) \\ &= \int_{-\infty}^{\infty} f(2x+1)dx + O^*(|f'(2x+1)|_1/2) \\ &= \frac{1}{2}\int_{-\infty}^{\infty} f(x)dx + O^*\left(\frac{|f'|_1}{2}\right), \end{aligned} \tag{3.10}$$

where we define $g : \mathbb{R} \to \mathbb{C}$ by $g(x) = f(2x+1)$. This estimate is [Tao14, (3.1)]; it is used throughout [Tao14].

Similarly, by (3.6), for any continuous and piecewise continuously differentiable $f : \mathbb{R} \to \mathbb{C}$ such that f, f', f'' are in L^1,

$$\sum_{n \text{ odd}} f(n) = \frac{1}{2}\int_{-\infty}^{\infty} f(x)dx + O^*\left(\frac{1}{8}|f''|_1\right). \tag{3.11}$$

While we will most often use these formulae to estimate sums in terms of integrals (as Euler did), they can also be used to estimate integrals in terms of sums (as Maclaurin did). In particular, (3.10) should be easily recognizable as the simplest estimate on the error term of the midpoint rule for numerical integration, whereas (3.11) is a variant of the well-known better estimate on the same error term when f'' is in L^1. See §4.1.3 and in particular (4.1). The fact that we have a factor of $1/8$ rather than $1/6$ in (3.11) is due to our improvement in the proof of Lemma 3.1. If we were taking an integral on a compact interval not containing the support of f, we could not, of course, make that improvement; if we could, it would mean that everybody up to now has been working with a suboptimal estimate of the error term in the midpoint rule.

For much more on the Euler-Maclaurin formula, and on Bernoulli numbers, in part from a computational vantage point, see [Coh07, §9.1–9.3].

3.1.3 Summation by parts and Abel summation

Summation by parts is so common that it barely needs to be mentioned: for any integer N, $\{a_n\}_{1\le n\le N}$, $\{b_n\}_{1\le n\le N}$, $a_n, b_n \in \mathbb{C}$,

$$\begin{aligned}\sum_{1\le n\le N} a_n b_n &= \sum_{1\le n\le N} (A(n) - A(n-1))\cdot b_n \\ &= A(N)\cdot b_N - \sum_{1\le n\le N-1} A(n)\cdot(b_{n+1} - b_n),\end{aligned}$$

where $A(x) = \sum_{1\le n\le x} a_n$. As we see, this is just a rearrangement of terms.

Abel's summation formula is the following variant. Let $\{a_n\}_{n\ge r}$, $a_n \in \mathbb{C}$, $r \in \mathbb{R}$. Let $F : [r,\infty) \to \mathbb{C}$ be of bounded variation. Define $A_r(x) = \sum_{r<n\le x} a_n$. Since $A_r(u)$ depends only on $\lfloor u \rfloor$,

$$\int_r^x A_r(u)F'(u)du = \sum_{r<n\le x} A_r(n)\int_n^{\min(n+1,x)} F'(u)du.$$

Here, as usual, we choose to see F' as a distribution. Now,

$$\begin{aligned}\sum_{r<n\leq x} A_r(n)\int_n^{\min(n+1,x)} F'(u)du &= \sum_{r<n\leq x} A_r(n)\cdot(F(\min(n+1,x))-F(n))\\ &= A_r(\lfloor x\rfloor)F(x) - \sum_{r<n\leq x}(A_r(n)-A_r(n-1))F(n)\\ &= A_r(x)F(x) - \sum_{r<n\leq x} a_n F(n),\end{aligned}$$

and so

$$\sum_{r<n\leq x} a_n F(n) = A_r(x)F(x) - \int_r^x A_r(u)F'(u)du. \quad \text{(Abel summation)} \qquad (3.12)$$

Summation by parts and Abel summation are just special cases of integration by parts with respect to the Lebesgue integral – that is, the Lebesgue integral with respect to the usual measures on $\mathbb{Z}$ and $\mathbb{R}$.

3.1.4 The Poisson summation formula

The Poisson summation formula states that, if (a) $f:\mathbb{R}\to\mathbb{C}$ is continuous, (b) f is in in L^1, (c) f is of bounded variation, (d) the restriction of $\widehat{f}$ to the integers is in ℓ^1, then

$$\sum_{n\in\mathbb{Z}} f(n) = \sum_{n\in\mathbb{Z}} \widehat{f}(n). \qquad (3.13)$$

See, e.g., [MV07, Thm. D.3].

It is easy to see conditions (a)–(d) hold automatically if the following conditions (which we often need at any rate) are assumed: (a') f is continuous, (b') f is piecewise continuously differentiable, (c') f, f' and f'' are in L^1, where f'' is understood as a distribution. Conditions (a) and (b) are immediate from (a') and (c'), condition (c) follows easily from (b') and (c'), and condition (d) follows from the fact that $\widehat{f}$ decays at least quadratically (by (2.16) and the assumption that f'' is in L^1).

Under the same conditions on f (namely, conditions (a)-(d), or (a')–(c')), and for any $q\in\mathbb{Z}^+$, $a\in\mathbb{Z}$,

$$\sum_{\substack{n\in\mathbb{Z}\\ n\equiv a \bmod q}} f(n) = \frac{1}{q}\sum_{b=0}^{q-1} e\left(\frac{ab}{q}\right)\sum_{n\in\mathbb{Z}} \widehat{f}\left(n+\frac{b}{q}\right). \qquad (3.14)$$

We can easily deduce this equality – which we shall also call Poisson summation –

from (3.13): letting $g(x) = f(a + qx)$, we see, by (3.13) and (2.13), that

$$\sum_{\substack{n\in\mathbb{Z}\\ n\equiv a \bmod q}} f(n) = \sum_{n\in\mathbb{Z}} g(n) = \sum_{n\in\mathbb{Z}} \widehat{g}(n) = \frac{1}{q}\sum_{n\in\mathbb{Z}} e\left(\frac{an}{q}\right)\widehat{f}\left(\frac{n}{q}\right)$$
$$= \frac{1}{q}\sum_{b=0}^{q-1} e\left(\frac{ab}{q}\right)\sum_{n\in\mathbb{Z}} \widehat{f}\left(n+\frac{b}{q}\right).$$

3.2 EXPONENTIAL SUMS

The simplest estimates for an exponential sum $\sum_n f(n)e(\alpha n)$ are of course those derived from the trivial bound $|\sum_n f(n)e(\alpha n)| \leq \sum_n |f(n)|$, followed by a bound on $\sum_n |f(n)|$ as in (3.4), (3.6), (3.10), or (3.11) (used for the function $h(x) = |f(x)|$ instead of $f(x)$).

One can often do better by obtaining cancellation, as follows.

3.2.1 A first arbitrary-order estimate

Lemma 3.3. *Let $k \geq 1$ and $f : \mathbb{R} \to \mathbb{C}$ be continuous. If $k \geq 2$, assume that f is $(k-2)$ times continuously differentiable, and that $f^{(k-2)}$ is piecewise continuously differentiable. In any event, assume that $f, f', \ldots, f^{(k)}$ are in L^1. Then*

$$\left|\sum_{n\in\mathbb{Z}} f(n)e(\alpha n)\right| \leq \frac{\left|f^{(k)}\right|_1}{|2\sin(\pi\alpha)|^k} \tag{3.15}$$

for every $\alpha \in \mathbb{R}$.

This is the last part of [Tao14, Lem. 3.1] (though there f is required to be smooth and compactly supported; these assumptions are then relaxed). Vinogradov worked with (3.15) for $k = 1$ and the brutal truncation $f = 1_{[0,1]}$, ([Vin54, Ch. I, Lem. 6]; proof by geometric series).

Proof. By summation by parts,

$$\sum_{n\in\mathbb{Z}} (f(n+1) - f(n))e(\alpha n) = \sum_{n\in\mathbb{Z}} f(n)(e(\alpha(n-1)) - e(\alpha n))$$
$$= (e(-\alpha) - 1)\sum_{n\in\mathbb{Z}} f(n)e(\alpha n).$$

Hence,

$$\sum_{n\in\mathbb{Z}} f(n)e(\alpha n) = \frac{1}{e(-\alpha) - 1}\int_0^1 \left(\sum_{n\in\mathbb{Z}} f'(n+t)e(\alpha n)\right) dt. \tag{3.16}$$

Note here that $|e(-\alpha) - 1| = 2|\sin \pi\alpha|$. Finally,

$$\int_0^1 \left(\sum_{n\in\mathbb{Z}} f'(n+t)e(\alpha n) \right) dt = \int_{-\infty}^{\infty} f'(x)e(\alpha\lfloor x\rfloor)dx = O^*(|f'|_1).$$

Therefore, (3.15) holds for $k = 1$. To prove (3.15) for $k \geq 2$, apply (3.15) (with $k-1$ instead of k and $f'(n+t)$ instead of f) to the sum within the integral in (3.16). □

3.2.2 Vinogradov meets Poisson and Euler

We will use Lemma 3.3 only for $k = 1$. For $k \geq 2$, we can replace it by the following lemma. It is never worse than Lemma 3.3, and as we will discuss soon, it is usually strictly stronger than it.

Lemma 3.4. *Let $k \geq 2$ and $f : \mathbb{R} \to \mathbb{C}$ be continuous. Assume that f is $(k-2)$ times continuously differentiable, that $f^{(k-2)}$ is piecewise continuously differentiable, and that $f, f', \dots, f^{(k)}$ are in L^1.*

Then

$$\left|\sum_{n\in\mathbb{Z}} f(n)e(\alpha n)\right| \leq |\widehat{f^{(k)}}|_\infty \cdot \begin{cases} \frac{-\cot^{(k-1)}(\pi\alpha)}{2^k(k-1)!} & \text{if } k \text{ is even,} \\ \frac{-\cot^{(k-2)}(\pi\alpha)}{2^k(k-2)!|\sin\pi\alpha|} & \text{if } k \text{ is odd} \end{cases} \tag{3.17}$$

for every $\alpha \in \mathbb{R}$.

In particular, for $k = 2$,

$$\left|\sum_{n\in\mathbb{Z}} f(n)e(\alpha n)\right| \leq \frac{|\widehat{f''}|_\infty}{|2\sin\pi\alpha|^2}. \tag{3.18}$$

Note that $|\widehat{f^{(k)}}|_\infty \leq |f^{(k)}|_1$. Comparing the left side of (3.20), raised to the $(k/2)$th power, to the right side of (3.21), we see that

$$\frac{-\cot^{(k-1)}(\pi\alpha)}{(k-1)!} \leq \frac{1}{(\sin\pi\alpha)^k}$$

for α real and $k \geq 2$ even, with strict inequality when $\alpha \notin \mathbb{Z}$ and $k > 2$. Hence the right side of (3.17) is bounded above by the right side of (3.15) for every $k \geq 2$. For instance, $-\cot^{(3)}(\pi\alpha)/3! = (1 - 2(\sin\pi\alpha)^2/3)/(\sin\pi\alpha)^4 \leq 1/(\sin\pi\alpha)^4$.

Proof. By the Poisson summation formula,

$$\sum_{n\in\mathbb{Z}} f(n)e(\alpha n) = \sum_{n\in\mathbb{Z}} \widehat{f}(n-\alpha). \tag{3.19}$$

Since $\widehat{f}(t) = \widehat{f'}(t)/(2\pi i t)$,

$$\sum_{n\in\mathbb{Z}} \widehat{f}(n-\alpha) = \sum_{n\in\mathbb{Z}} \frac{\widehat{f'}(n-\alpha)}{2\pi i(n-\alpha)} = \sum_{n\in\mathbb{Z}} \frac{\widehat{f''}(n-\alpha)}{(2\pi i(n-\alpha))^2} = \dots = \sum_{n\in\mathbb{Z}} \frac{\widehat{f^{(k)}}(n-\alpha)}{(2\pi i(n-\alpha))^k}.$$

Differentiating Euler's formula (2.38), we see that

$$\sum_{n\in\mathbb{Z}} \frac{1}{(n+s)^2} = -(\pi\cot\pi s)' = \frac{\pi^2}{(\sin\pi s)^2}, \tag{3.20}$$

and, in general,

$$\sum_{n\in\mathbb{Z}} \frac{1}{(n+s)^k} = \frac{-(\pi\cot\pi s)^{(k-1)}}{(k-1)!} = -\frac{\pi^k\cot^{(k-1)}\pi s}{(k-1)!}. \tag{3.21}$$

Hence

$$\left|\sum_{n\in\mathbb{Z}} \widehat{f}(n-\alpha)\right| \leq |\widehat{f^{(k)}}|_\infty \sum_{n\in\mathbb{Z}} \frac{1}{(2\pi(n-\alpha))^k} = |\widehat{f^{(k)}}|_\infty \cdot \frac{-\cot^{(k-1)}\pi\alpha}{2^k(k-1)!}$$

for k even, $k\geq 2$.

In order to prove (3.17) for $k\geq 3$ odd, apply (3.17) with $f'(n+t)$ instead of f and $k-1$ instead of k to bound the sum within the integral in (3.16). □

As in §3.1.2 (or [Tao14]), we want versions of these results over the odd numbers. Clearly

$$\sum_{n \text{ odd}} f(n)e(\alpha n) = e(\alpha)\sum_n g(n)e(2\alpha n), \tag{3.22}$$

for $g:\mathbb{R}\to\mathbb{C}$ defined by $g(x) = f(2x+1)$. We now apply Lemma 3.3 (for $k=1$) and Lemma 3.4 (for $k=2$) to the sum on the right. with $t\mapsto f(2t+1)$ instead of f. We obtain that, for any continuous, piecewise continuously differentiable $f:\mathbb{R}\to\mathbb{C}$ such that $f, f', f''\in L^1$,

$$\sum_{n \text{ odd}} f(n)e(\alpha n) = O^*\left(\min\left(\frac{|f'|_1}{|2\sin 2\pi\alpha|}, \frac{|\widehat{f''}|_\infty}{2|\sin 2\pi\alpha|^2}\right)\right). \tag{3.23}$$

This is the bound we will usually employ. When α is very close to an integer, we will use (3.11) instead, with $|f|$ instead of f.

* * *

Is it possible to improve Lemma 3.3 for $k=1$? In the light of Lemma 3.4, one might be tempted to guess that $|\sum_{n\in\mathbb{Z}} f(n)e(\alpha n)| \leq |\widehat{f'}|_\infty/|2\sin\pi\alpha|$. That, however, is wrong. Indeed, one cannot in general bound $|\sum_{n\in\mathbb{Z}} f(n)e(\alpha n)|$ by $|\widehat{f'}|_\infty$ times a function of α. The following example – kindly contributed by W. Sawin – shows this impossibility.

Let g be a smooth function of compact support with the following properties: $g(x) = 1/|x|$ for $x\in[-N,-1]$ and $x\in[1,N]$; inside $[-1,1]$, it is defined arbitrarily (as long as it stays smooth), while outside $[-N,N]$, it decays rapidly. Let f be given by $f(t) = \widehat{g}(-t)$, so that, by the Fourier inversion formula, $\widehat{f} = g$. Since g is smooth and compactly supported, f is smooth, and, by (2.16), f and all of its derivatives decay more rapidly than any power $1/|t|^k$ as $|t| \to \infty$. Thus, the Poisson summation formula holds for $h(x) = f(x)e(\alpha x)$, and so

$$\sum_{n\in\mathbb{Z}} f(n)e(\alpha n) = \sum_{n\in\mathbb{Z}} h(n) = \sum_{n\in\mathbb{Z}} \widehat{h}(n) = \sum_{n\in\mathbb{Z}} \widehat{f}(n-\alpha).$$

At the same time, $\sum_{n\in\mathbb{Z}} \widehat{f}(n-\alpha) = \sum_{n\in\mathbb{Z}} g(n-\alpha) \gg \log N$, whereas $|\widehat{f}'|_\infty = |tg(t)|_\infty \ll 1$. Hence, we cannot bound $\sum_{n\in\mathbb{Z}} f(n)e(\alpha n)$ by a multiple of $|\widehat{f}'|_\infty$.

3.3 HARMONIC SUMS

It is easy to see that

$$\sum_{n\leq x} \frac{1}{n} - \log x \tag{3.24}$$

converges to a limit as $x \to \infty$: the areas lying above the hyperbola and under the horizontal line $y = 1/n$ for $n \leq x \leq n+1$ can be stacked on top of each other, and so their union has a well-defined area less than 1. The limit γ of (3.24) as $x \to \infty$ is called the *Euler-Mascheroni constant*, or simply *Euler's constant*. Its value is $\gamma = 0.57721567\ldots$

By the same geometrical argument,

$$\sum_{n\leq x} \frac{1}{n} = \log x + \gamma + O^*(1/x) \tag{3.25}$$

for $x > 0$. Observing that $x \mapsto 1/x$ is convex, we see that in fact

$$\sum_{n\leq x} \frac{1}{n} < \log x + \gamma + \frac{1}{2x} \tag{3.26}$$

for $x \geq 1$. As we shall see in a moment, one can improve the lower bound in (3.25) as well.

By the same geometrical argument, for any $x > 0$,

$$\sum_{\substack{n\leq x \\ n \text{ odd}}} \frac{1}{n} = \frac{1}{2}\log x + \frac{\gamma + \log 2}{2} + O^*\left(\frac{1}{x}\right). \tag{3.27}$$

The constant $\log 2$ is here because $\log 2 = 1 - 1/2 + 1/3 - 1/4 + 1/5 - 1/6 + \ldots$; this is the total area of the rectangles of the form $[2m, 2m+1] \times [1/2m, 1/(2m-1)]$ for $m \geq 1$. Again by convexity,

$$\sum_{\substack{n\leq x \\ n \text{ odd}}} \frac{1}{n} < \frac{1}{2}\log x + \frac{\gamma + \log 2}{2} + \frac{1}{2x} \tag{3.28}$$

for $x > 0$.

Somewhat cruder bounds can also be useful. For instance, for $x \geq 1$,

$$\sum_{n \leq \frac{x-1}{2}} \frac{1}{n} \leq \int_{1/2}^{x/2} \frac{dx}{x} = \log x, \tag{3.29}$$

yet again by convexity.

Comparing (3.25) and Lemma 3.2 (Euler-Maclaurin), we see that

$$\log x + \gamma + O\left(\frac{1}{x}\right) = \sum_{n \leq x} \frac{1}{n} = \log x + \frac{1}{2} - \frac{B_1(\{x\})}{x} + \int_1^x \frac{2B_1(\{t\})}{t^3} dt$$

and so, letting $x \to \infty$, we obtain

$$\gamma = \frac{1}{2} + \int_1^\infty \frac{2B_1(\{t\})}{t^3} dt.$$

Hence

$$\begin{aligned} \sum_{n \leq x} \frac{1}{n} &= \log x + \gamma - \frac{B_1(\{x\})}{x} - \int_x^\infty \frac{2B_1(\{t\})}{t^3} dt \\ &= \log x + \gamma - \frac{B_1(\{x\})}{x} + O^*\left(\frac{1}{x^2}\right). \end{aligned} \tag{3.30}$$

Since $B_1(x) = x - 1/2$, it follows that $\sum_{n \leq x} 1/n = \log x + \gamma + O^*(0.54/x)$ for $x \geq 25$, and thus, as noted in [RA17, Lem. 2.1],

$$\sum_{n \leq x} \frac{1}{n} = \log x + \gamma + O^*\left(\frac{c}{x}\right) \tag{3.31}$$

for $x \geq 1$, with $c = 2(\log 2 + \gamma - 1) = 0.54072\ldots$, and

$$\sum_{n \leq x} \frac{1}{n} = \log x + \gamma + O^*\left(\frac{\gamma}{x}\right) \tag{3.32}$$

for $x > 0$. We check either equation by finding the extrema of the difference $(\sum_{n \leq x} 1/n) - (\log x + \gamma)$ in each interval $[n, n+1]$, $0 \leq n \leq 24$.

Similarly, by (3.27) and Lemma 3.2 applied to $x \mapsto 1/(2x-1)$,

$$\frac{\log x}{2} + \frac{\gamma + \log 2}{2} + o(1) = \sum_{n \leq \frac{x}{2}} \frac{1}{n} = \log x + \frac{1}{2} - \frac{B_1(\{x/2 + 1/2\})}{x} + \int_1^x \frac{4B_1(\{t\})}{t^3} dt$$

and so

$$\begin{aligned} \sum_{\substack{n \leq x \\ n \text{ odd}}} \frac{1}{n} &= \frac{\log x}{2} + \frac{\gamma + \log 2}{2} - \frac{B_1(\{x/2 + 1/2\})}{x} - \int_x^\infty \frac{4B_1(\{t\})}{t^3} dt \\ &= \frac{\log x}{2} + \frac{\gamma + \log 2}{2} - \frac{B_1(\{x/2 + 1/2\})}{x} + O^*\left(\frac{2}{x^2}\right). \end{aligned} \tag{3.33}$$

3.4 THE GAMMA FUNCTION

We define the *Gamma function* $\Gamma(s)$ for $\Re s > 0$ as the Mellin transform of $x \to e^{-x}$:

$$\Gamma(s) = \int_0^\infty e^{-x} x^{s-1} dx. \tag{3.34}$$

It is easy to see that the integral converges for $\Re s > 0$. In the domain $\Re s > 0$, the identity $\Gamma(s+1) = s\Gamma(s)$ follows from (3.34) by integration by parts. It is then a simple exercise to show that Γ can be extended to a meromorphic function on $\mathbb{C}$ with simple poles at $s = 0, -1, -2, \dots$, obeying $\Gamma(s+1) = s\Gamma(s)$ at all other s. We define $\Gamma(s)$ on the complex plane in this way.

Since $\Gamma(1) = 1$ and $\Gamma(s+1) = s\Gamma(s)$, we see that $\Gamma(n) = (n-1)!$ for all $n \in \mathbb{Z}^+$.

3.4.1 Basic theory

The study of $\Gamma(s)$ goes back to Euler, at least for s real. The statements and proofs below, all very well-known, have points in common with Euler's work, as can be seen in [Var06, §3.6], though of course standards of rigor in analysis were different in Euler's day.

We follow in part [MV07, App. B] and [Var06, §3.6]. Far more detailed treatments can be found in [AAR99, Ch. 1] and [Rem98, Ch. 2].

Proposition 3.5. *For every $s \in \mathbb{C}$ other than $s = 0, -1, -2, \dots$,*

$$\Gamma(s) = \lim_{N\to\infty} \frac{N^s N!}{s(s+1)\cdots(s+N)} \qquad \text{(Gauss's formula)}. \tag{3.35}$$

Moreover,

$$\Gamma(s) = \frac{e^{-\gamma s}}{s} \prod_{n=1}^{\infty} \frac{e^{s/n}}{1+s/n}, \qquad \text{(Weierstrass product)} \tag{3.36}$$

where γ is Euler's constant.

Proof. Let us first prove Gauss's formula (3.35) for $\Re s > 0$. By induction on N,

$$\frac{N!}{s(s+1)\cdots(s+N)} = \int_0^\infty (1-y)^N y^{s-1} dy,$$

where we use integration by parts N times. We change variables $x = Ny$ and obtain

$$\frac{N^s N!}{s(s+1)\cdots(s+N)} = \int_0^\infty f_N(x) dx,$$

where

$$f_N(x) = \begin{cases} (1-x/N)^N x^{s-1} & \text{for } 0 \le x \le N, \\ 0 & \text{otherwise.} \end{cases}$$

Now, since $1 - y \leq e^{-y}$ for $0 \leq y \leq 1$, we see that $|f_N(x)| \leq f(x)$, where $f(x) = e^x x^{\Re s - 1}$. Since $\Re s > 0$, $\int_0^\infty f(x)dx < \infty$. Therefore, by dominated convergence,

$$\lim_{N\to\infty} \int_0^\infty f_N(x)dx = \int_0^\infty \left(\lim_{N\to\infty} f_N(x)\right) dx = \int_0^\infty e^{-x}x^{s-1}dx,$$

where we use the limit $\lim_{t\to\infty}(1 - 1/t)^t = e^{-1}$. Thus, (3.35) holds for $\Re s > 0$.

By the estimate (3.25) on the harmonic sum $\sum_{n\leq N} 1/n$,

$$\frac{N^s N!}{s(s+1)\cdots(s+N)} = e^{O^*(s/N)} \cdot \frac{e^{-\gamma s}}{s} \prod_{n=1}^{\infty} \frac{e^{s/n}}{1+s/n}.$$

Hence the right sides of (3.35) and (3.36) are equal for any $s \in \mathbb{C} \setminus \{0, -1, -2, \dots\}$. In particular, since the right side of (3.36) is a meromorphic function on $\mathbb{C}$, so is the right side of (3.35). By analytic continuation, we conclude that both (3.35) and (3.36) hold for all $s \in \mathbb{C} \setminus \{0, -1, -2, \dots\}$. □

We could also prove (3.36) in the style of Weierstrass, applying Lemma 2.11 to the function $1/s\Gamma(s)$. We would then need a growth bound on $\Gamma(s)$ that we will prove later, in Corollary 3.10. See [Ahl78, §5.2.5].

Let us prove the next result using an approach from §2.6.1.2.

Lemma 3.6. *On the entire complex plane,*

$$\Gamma(s)\Gamma(1-s) = \frac{\pi}{\sin \pi s}$$

Proof. The function $\Gamma(s)\Gamma(1-s)$ has a simple pole at every integer and is holomorphic elsewhere; moreover, by $s\Gamma(s) = \Gamma(s+1)$, we see that $\Gamma(s)\Gamma(1-s)$ is periodic with period 1. The function $\sin \pi s$ has a zero at every integer and is periodic with period 1. Thus, $\Gamma(s)\Gamma(1-s)\sin \pi s$ is entire and also has period 1. It is clear from the definition that $|\Gamma(s)| \leq \Gamma(\Re s)$. Hence, for $s = \sigma + it$,

$$|\Gamma(s)\Gamma(1-s)\sin \pi s| \ll |\sin \pi s| = \left|\frac{e^{i\pi s} - e^{-i\pi s}}{2}\right| \ll e^{\pi t}. \tag{3.37}$$

We now apply Lemma 2.8 and obtain $\Gamma(s)\Gamma(1-s)\sin \pi s$ equals a constant c.

Since $\Gamma(1) = 1$ and $\Gamma(s) = \Gamma(1+s)/s$, we see that $\Gamma(s)\Gamma(1-s)$ behaves like $1/s$ as $s \to 0$, whereas $\sin \pi s$ behaves like πs. Hence $c = \pi$. □

Since $\Gamma(1/2)$ is, by definition, a positive real, Lem. 3.6 gives us that

$$\Gamma(1/2) = \sqrt{\pi}. \tag{3.38}$$

Lemma 3.7. *On the entire complex plane,*

$$\Gamma(s)\Gamma(s+1/2) = \sqrt{\pi}2^{1-2s}\Gamma(2s) \qquad \textit{(Legendre's duplication formula)} \tag{3.39}$$

Proof. Write $g_N(s)$ for the expression within the limit in Gauss's formula (3.35). A quick check gives

$$2^{2s} g_N(s) g_N(s+1/2) = \frac{2N+1}{2s+2N+1} \cdot 2 g_{2N+1}(s) g_N(1/2).$$

We let $N \to \infty$ and use equation (3.38). □

Lemma 3.8 (Hankel)**.** *For $\epsilon > 0$, let $\mathscr{H} = \mathscr{H}(\epsilon)$ be a* Hankel contour, *going on a straight line from $-\infty - i\epsilon$ to $-i\epsilon$, then on a semicircle counterclockwise from $-i\epsilon$ to $i\epsilon$, then on a straight line from $i\epsilon$ to $-\infty - i\epsilon$. Then, for all $s \in \mathbb{C}$,*

$$\frac{1}{2\pi i} \int_{\mathscr{H}} e^z z^{-s} dz = \frac{1}{\Gamma(s)}, \tag{3.40}$$

where we define z^{-s} by the principal branch of the logarithm $\log z$*.*

Proof. Since both sides of the equation (3.40) are entire, it is enough to prove the equation for $\Re s < 1$.

We can let $\epsilon \to 0^+$ without changing the value of the integral. By $\Re s < 1$, the integral over the semicircle tends to 0. We are left with $1/2\pi i$ times

$$\begin{aligned}\lim_{\epsilon\to 0^+} \left(\int_{-\infty - i\epsilon}^{-i\epsilon} e^z z^{-s} dz - \int_{-\infty + i\epsilon}^{i\epsilon} e^z z^{-s} dz \right) &= \left(e^{\pi i s} - e^{-\pi i s}\right) \int_0^\infty e^{-x} x^{-s} dx \\ &= (2i \sin \pi s) \cdot \Gamma(1-s).\end{aligned}$$

We finish by Lemma 3.6. □

We will need *Stirling's formula* with explicit error terms.

Lemma 3.9 (Stirling's formula)**.** *For $s \in \mathbb{C}$ with $\Re s \geq 0$, $s \neq 0$,*

$$\log \Gamma(s) = \left(s - \frac{1}{2}\right) \log s - s + \frac{1}{2} \log 2\pi + \frac{1}{12s} + O^*\left(\frac{1 + 3\pi/4}{360|s|^3}\right). \tag{3.41}$$

The condition $\Re s \geq 0$ can be relaxed and the error term improved [GR94, (8.344)].

Proof. By Euler-Maclaurin (Lemma 3.2) to order $k = 1$ for $f(x) = \log x$,

$$\begin{aligned}\sum_{n=1}^{N} \log n &= \int_1^N (\log x) dx + (-1) B_1(0)(f(N) - f(1)) + \int_1^N B_1(\{x\}) f'(x) dx \\ &= N(\log N - 1) + \frac{1}{2} \log N + \int_1^N \frac{\{x\} - 1/2}{x} dx.\end{aligned} \tag{3.42}$$

Again by Euler-Maclaurin, this time to order $k = 4$ for $f(x) = \log(s + x)$,

$$\sum_{n=1}^{N} \log(s+n) = (s+N)(\log(s+N)-1) - s(\log s - 1) + \frac{1}{2}(\log(s+N) - \log(s)) + \frac{1/6}{2!}\left(\frac{1}{s+N} - \frac{1}{s}\right) - \frac{1/30}{4!}\left(\frac{2}{(s+N)^3} - \frac{2}{s^3}\right) + I_4(s,N), \tag{3.43}$$

where

$$I_4(z,N) = \frac{(-1)^5}{4!}\int_0^N B_4(\{x\}) f^{(4)}(x)dx = -\frac{1}{4}\int_0^N \frac{B_4(\{x\})}{(s+x)^4}dx.$$

By Gauss's formula (3.35),

$$\log\Gamma(s) = \lim_{N\to\infty}\left(s\log N - \log s + \sum_{n=1}^{N}\log n - \sum_{n=1}^{N}\log(s+n)\right).$$

The limit as $N \to \infty$ of

$$s\log N + N(\log N - 1) + \frac{1}{2}\log N - N(\log(s+N)-1) - s\log(s+N) - \frac{1}{2}\log(s+N)$$

is simply $\lim_{N\to\infty} N\log N - N\log(s+N) = \lim_{N\to\infty} -N\log(1+s/N) = -s$. Hence, by (3.42) and (3.43),

$$\log\Gamma(s) = \left(s - \frac{1}{2}\right)\log s - s + \int_1^\infty \frac{\{x\} - 1/2}{x}dx + \frac{1}{12s} - \frac{1}{360s^3} - \lim_{N\to\infty} I_4(s,N),$$

Here $|\lim_{N\to\infty} I_4(s,N)| < \frac{|B_4|}{4}\int_0^\infty \frac{dx}{|s+x|^4} \leq \frac{1}{120}\int_0^\infty \frac{dx}{(|s|^2+x^2)^2}$ and

$$\int_0^\infty \frac{dx}{(|s|^2+x^2)^2} = \left.\frac{x/2}{|s|^4 + |s|^2x^2}\right|_0^\infty + \left.\frac{\arctan(x/|s|)}{2|s|^3}\right|_0^\infty = \frac{\pi/2}{2|s|^3}.$$

In particular, $\log\Gamma(s)$ asymptotes to $(s - 1/2)\log s - s + c$ as $|s| \to \infty$, where $c = \int_1^\infty (\{x\} - 1/2)dx/x$. To obtain that $c = \log\sqrt{2\pi}$, compare both sides of Legendre's duplication formula (3.39) for $s = it$. □

Let us work out what Stirling's formula implies on the behavior of $\Gamma(s)$ for $s = \sigma + it$, σ fixed, $|t| \to \infty$.

Corollary 3.10. *For $s = \sigma + it$, $\sigma \geq 0$, $t \neq 0$ with $|t| \geq \max(6\sigma^2, 4\sigma^3, 4)$,*

$$|\Gamma(\sigma+it)| = \left(1 + O^*\left(\frac{2}{9t}\right)\right)\cdot\sqrt{2\pi}|t|^{\sigma-1/2}e^{-\frac{\pi|t|}{2}}, \tag{3.44}$$

Proof. We start from (3.41). Clearly

$$\Re\left(\left(s - \frac{1}{2}\right)\log s - s\right) = \left(\sigma - \frac{1}{2}\right)\log|s| - t\arg(s) - \sigma.$$

Since $|\Gamma(s)| = |\Gamma(\overline{s})|$, we can assume without loss of generality that $t > 0$. Then

$$\arg(s) = \frac{\pi}{2} - \arctan\frac{\sigma}{t} \in \frac{\pi}{2} - \frac{\sigma}{t} + \left[0, \frac{(\sigma/t)^3}{3}\right]$$

and

$$\log|s| = \log t + \log\sqrt{1 + \frac{\sigma^2}{t^2}} \in \log t + \left[0, \frac{\sigma^2}{2t^2}\right],$$

with the consequence that

$$\left(\sigma - \frac{1}{2}\right)\log|s| - t\arg(s) - \sigma = \left(\sigma - \frac{1}{2}\right)\log t - \frac{\pi}{2}t + O^*\left(\max\left(\frac{\sigma^3}{2t^2}, \frac{\sigma^3}{3t^2} + \frac{\sigma^2}{4t^2}\right)\right).$$

Thus

$$\log|\Gamma(s)| = \left(\sigma - \frac{1}{2}\right)\log t - \frac{\pi}{2}t + \frac{1}{2}\log 2\pi + O^*\left(\max\left(\frac{\sigma^3}{2t^2}, \frac{\sigma^3}{3t^2} + \frac{\sigma^2}{4t^2}\right) + \frac{1}{12t} + \frac{1 + 3\pi/4}{360t^3}\right).$$

Under the assumption $t \geq \max(6\sigma^2, 4\sigma^3, 4)$, the error term here is bounded by c_0/t, where $c_0 = \frac{5}{24} + 0.00058\ldots$. Since $(e^{c_0/t} - 1)t$ is a decreasing function of t, and $e^{-c_0/t} > 1 - c_0/t$, it is clear that, for $t \geq 4$, $e^{O^*(c_0/t)} = 1 + O^*(c_1/t)$ with $c_1 = (e^{c_0/4} - 1) \cdot 4 = 0.21446\ldots > c_0$. □

3.4.2 The digamma function

We define the *digamma function* $\digamma(s)$ by

$$\digamma(s) = (\log\Gamma(s))' = \frac{\Gamma'(s)}{\Gamma(s)}. \tag{3.45}$$

It follows immediately from Lemmas 3.6 and 3.7 that

$$\digamma(1 - s) - \digamma(s) = \pi\cot\pi s \tag{3.46}$$

and

$$\digamma(s) + \digamma(s + 1/2) = 2(\digamma(2s) - \log 2). \tag{3.47}$$

Looking at the Weierstrass product (3.36), we see that

$$\digamma(s) = -\gamma - \frac{1}{s} - \sum_{n=1}^{\infty}\left(\frac{1}{s+n} - \frac{1}{n}\right) \tag{3.48}$$

Thus we obtain, for instance,

$$\digamma(1) = -\gamma, \qquad \digamma(1/2) = -\gamma - 2\log 2, \tag{3.49}$$

where we use $\log(1+t) = \sum_{n=0}^{\infty}(-t)^n/n$ with $t \to 1^-$. It is obvious from (3.48) that, for s real, $\digamma(s)$ increases as s increases.

Lemma 3.11. *For* $\Re s \geq 0$,

$$\digamma(s) = \log s - \frac{1}{2s} + O^*\left(\frac{1}{4|s|^2}\right),$$

where log *denotes the principal branch of the logarithm.*

The condition $\Re s \geq 0$ can again be relaxed.

Proof. By Euler-Maclaurin (Lemma 3.2) to order $k = 2$ for $f(x) = 1/(s+x)$,

$$\sum_{n=1}^{N} \frac{1}{s+n} = \int_0^N \frac{dx}{s+x} + \sum_{j=1}^{k} \frac{(-1)^j B_j(0)}{j!} (f^{(j-1)}(N) - f^{(j-1)}(0)) + I_4(s,N)$$
$$= \log(s+N) - \log s + \frac{1}{2}\left(\frac{1}{s+N} - \frac{1}{s}\right) - \frac{1}{12}\left(\frac{1}{(s+N)^2} - \frac{1}{s^2}\right) + I_2(s,N),$$

where

$$I_2(s,N) = \frac{(-1)^{k+1}}{k!} \int_0^N B_k(\{x\}) f^{(k)}(x) dx = \int_0^N \frac{B_2(\{x\})}{(s+x)^3} dx.$$

Hence, by (3.48) and (3.25),

$$\digamma(s) = -\gamma - \frac{1}{s} + \log N + \gamma + O^*(1/N)$$
$$- \left(\log(s+N) - \log s + \frac{1}{2}\left(\frac{1}{s+N} - \frac{1}{s}\right) - \frac{1}{12}\left(\frac{1}{(s+N)^2} - \frac{1}{s^2}\right) - I(N)\right),$$

and so, letting $N \to \infty$, we see that

$$\digamma(s) = \log s - \frac{1}{2s} + \frac{1}{12s^2} + \int_0^\infty \frac{B_2(\{x\})}{(s+x)^3} dx.$$

The statement follows by $|B_2(\{x\})| \leq 1/6$ and

$$\int_0^\infty \frac{1}{|s+x|^3} dx \leq \int_0^\infty \frac{dx}{(|s|^2 + x^2)^{\frac{3}{2}}} = \left.\frac{x}{|s|^2\sqrt{|s|^2 + x^2}}\right|_0^\infty = \frac{1}{|s|^2}.$$

□

3.5 THE RIEMANN ZETA FUNCTION

The Riemann zeta function $\zeta(s)$ is one of the most basic objects in analytic number theory, and the central matter of a typical first course. This section is meant to be a minimal review. A far more thorough treatment can be found in [Tit86, Ch. I–II].

3.5.1 Definition. Analytic continuation and functional equation.

The Riemann zeta function $\zeta(s)$ is defined by

$$\zeta(s) = \sum_{n=1}^{\infty} n^{-s} \tag{3.50}$$

for $\Re s > 1$. On the rest of the complex plane, it is defined by analytic continuation; it has a pole at $s = 1$, and is holomorphic elsewhere. The fact that an analytic continuation exists is of course not a priori obvious; let us see how to obtain it, since we will use the main ideas later.

Analytic continuation. For $\Re s > 1$,

$$\sum_{n=1}^{\infty} \left(\frac{1}{n^s} - \int_n^{n+1} \frac{dt}{t^s} \right) = \sum_{n=1}^{\infty} \frac{1}{n^s} - \int_1^{\infty} \frac{dt}{t^s} = \zeta(s) - \frac{1}{s-1}. \tag{3.51}$$

The nth term of the sum on the left side is of size $O(|s| n^{-s-1})$, as we can quickly ascertain taking the derivative of $1/t^s$. Hence, the sum converges absolutely for $\Re s > 0$. Therefore, $\zeta(s) - 1/(s-1)$ can be continued analytically to the region $\Re s > 0$.

We can also apply Euler-Maclaurin to arbitrary order $k \geq 1$, with $f(x) = x^{-s}$, $a = 1$ and $b \to \infty$, as in (3.8). We obtain that, for $\Re s > 1$,

$$\begin{aligned} \zeta(s) - \frac{1}{s-1} &= \sum_{n=1}^{\infty} n^{-s} - \int_1^{\infty} \frac{dt}{t^s} \\ &= -\sum_{j=1}^{k} \frac{B_j}{j!} f^{(j-1)}(1) + \frac{(-1)^{k+1}}{k!} \int_1^{\infty} B_k(\{x\}) f^{(k)}(x) dx. \end{aligned} \tag{3.52}$$

Here $f^{(r)}(1) = (-1)^r s(s+1)\cdots(s+r-1)$. The expression on the last line of (3.52) is well-defined for $\Re s > 1 - k$, as the integral in it is bounded by a constant times the integral of $\left|f^{(k)}(s)\right| = \left|s(s+1)\ldots(s+k-1)x^{-s-k}\right|$ for x from 1 to ∞. Hence, $\zeta(s) - 1/(s-1)$ can be extended to all of $\mathbb{C}$. The Taylor expansion of $\zeta(s)$ around $s = 1$ begins starts as follows:

$$\zeta(s) = \frac{1}{s-1} + \gamma + \ldots, \tag{3.53}$$

where γ is Euler's constant.

Functional equation. The zeta function has a line of symmetry at $\Re s = 1/2$, in the following sense.

Proposition 3.12 (Functional equation)**.** *For all $s \in \mathbb{C}$,*

$$\pi^{-\frac{s}{2}} \Gamma\left(\frac{s}{2}\right) \zeta(s) = \pi^{-\frac{1-s}{2}} \Gamma\left(\frac{1-s}{2}\right) \zeta(1-s). \tag{3.54}$$

Proof. By the definition (3.34) of $\Gamma(s)$ and a change of variables,

$$\pi^{-\frac{s}{2}} \Gamma\left(\frac{s}{2}\right) n^{-s} = \int_0^{\infty} e^{-\pi n^2 x} x^{\frac{s}{2}-1} dx.$$

Hence, for $\Re s > 1$,

$$\begin{aligned} \pi^{-\frac{s}{2}}\Gamma\left(\frac{s}{2}\right)\zeta(s) &= \int_0^\infty \left(\sum_{n=1}^\infty e^{-\pi n^2 x}\right) x^{\frac{s}{2}-1}dx \\ &= \int_0^\infty \left(\frac{\theta(ix/2)-1}{2}\right) x^{\frac{s}{2}-1}dx, \end{aligned} \tag{3.55}$$

where θ is the *theta function*: [1]

$$\theta(x) = \sum_{n\in\mathbb{Z}} e(n^2 x). \tag{3.56}$$

In other words, the left side of (3.54) equals the value of the Mellin transform of $x \mapsto (\theta(ix/2)-1)/2$ at $s/2$.

Applying the Poisson summation formula (3.13) to the rescaled Gaussian $f(t) = e^{-\pi x t^2}$, we obtain

$$\theta(ix/2) = \sum_{n\in\mathbb{Z}} \widehat{f}(n) = \frac{1}{\sqrt{x}}\sum_{n\in\mathbb{Z}} e^{-\pi n^2/x} = \frac{\theta(i/2x)}{\sqrt{x}},$$

where we use (2.19) (self-duality of the Gaussian). Thus, $(\theta(ix/2)-1)/2$ equals $(\theta(i/2x)/\sqrt{x}-1)/2 = ((\theta(i/2x)-1)/2)/\sqrt{x} + 1/2\sqrt{x} - 1/2$, and so, splitting the integral on the right side of (3.55) at $x = 1$, we obtain that, for $\Re s > 1$,

$$\begin{aligned} \pi^{-\frac{s}{2}}\Gamma\left(\frac{s}{2}\right)\zeta(s) &= \int_0^1 \left(\frac{\theta(i/2x)-1}{2\sqrt{x}} + \frac{1}{2\sqrt{x}} - \frac{1}{2}\right) x^{\frac{s}{2}}\frac{dx}{x} \\ &\quad + \int_1^\infty \left(\frac{\theta(ix/2)-1}{2}\right) x^{\frac{s}{2}}\frac{dx}{x} \\ &= -\frac{1}{1-s} - \frac{1}{s} + \int_1^\infty \left(\frac{\theta(ix/2)-1}{2}\right)\left(x^{\frac{1-s}{2}} + x^{\frac{s}{2}}\right)\frac{dx}{x}, \end{aligned} \tag{3.57}$$

where we are doing a change of variables $x \mapsto 1/x$ on the first integral. The integral in the last line of (3.57) defines a function of s that is (a) holomorphic on all of $\mathbb{C}$ (because of the fast decay of $\theta(ix/2)-1$) and (b) invariant under $s \mapsto 1-s$. We obtain that the functional equation (3.54) holds for all s. □

Remarks. The proof of the functional equation above gives us the analytic continuation of the left side of (3.57) (and thus of $\zeta(s)$) to all of $\mathbb{C}$ for free. Some other proofs

[1] There are several theta functions, and notation for them is not completely standardized. There is also no universal agreement on which one should be called *the* theta function. Both [IK04] and [MV07] define $\theta(x)$ to be $\sum_{n\in\mathbb{Z}} e^{-\pi n^2 x}$. That choice would simplify notation slightly in the proof here, but many feel strongly that θ should be a modular form on the upper half plane. Here we follow [Shi73, Ch. I, (2.26)], [Iwa97, Ch. 2.72] and [Gol06, p. 2]. In terms of the Jacobi theta function $\vartheta_3(z,q) = \sum_{n\in\mathbb{Z}} q^{n^2}\cos 2nz$ (to follow the notation in [AS64, §16.27]), the function $\theta(x)$ in (3.56) equals $\vartheta_3(0,q)$ for $q = e^{\pi i\tau}$.

establish the functional equation first for $0 < \Re s < 1$; then it follows for all $s \in \mathbb{C}$ by the analytic continuation of $\zeta(s)$ to $\Re s > 0$, which we proved at the outset. See [Tit86, Ch. II]) for a collection of proofs of the functional equation.

As has often been said, the functional equation can be "proved" if one applies the Poisson summation formula directly to $f(x) = |x|^{-s}$, in a highly non-rigorous (and anachronistic) 18th-century way – that is, blithely ignoring conditions necessary for convergence. (The function $f(x) = |x|^{-s}$ is never in L^1.) The proof above can thus be seen as making the "proof" correct by the introduction of a Gaussian weight.

3.5.2 Approximating $\zeta(s)$.

Let us discuss now how to determine the value of $\zeta(s)$ for an arbitrary $s \in \mathbb{C}$. We can apply to Euler-Maclaurin with $f(x) = x^{-s}\eta(x/N)$, where $\eta : \mathbb{R} \to [0, 1]$ is a smooth function with $\eta(t) = 0$ for $t \leq 1$ and $\eta(t) = 1$ for $t \geq 2$, say. Then, for $N > 1$, setting $a = 1$, $b \to \infty$, we obtain

$$\begin{aligned}\zeta(s) &= \sum_{n=1}^{\infty}(n^{-s} - f(n)) + \sum_{n=1}^{\infty} f(n)\\ &= \sum_{n=1}^{2N}(n^{-s} - f(n)) + \int_1^{\infty} f(x)dx + \frac{(-1)^{2k+1}}{2k!}\int_N^{\infty} B_{2k}(\{x\})f^{(2k)}(x)dx\end{aligned} \tag{3.58}$$

for $k \geq 1$ and $\Re s > 1$. Integrating by parts, we see that $\int_1^{\infty} f(x)dx = c_\eta \frac{N^{1-s}}{s-1}$ for $c_\eta = \int_1^2 t^{1-s}\eta'(t)dt$, and thus we obtain again a meromorphic continuation to $\Re s > 1 - 2k$, together with an approximation to $\zeta(s)$ valid in the same region:

$$\zeta(s) = \sum_{n=1}^{2N}(n^{-s} - f(n)) + c_\eta \frac{N^{1-s}}{1-s} + O^*\left(\frac{|B_{2k}|}{(2k)!}\int_N^{\infty}\left|f^{(2k)}(x)\right|dx\right). \tag{3.59}$$

For $s = \sigma + it$ and $k \lll |t|$, the dominant term in $f^{(2k)}(x)$ is generally

$$(x^{-s})^{(2k)}\eta(x/N) = s(s+1)\ldots(s+2k-1)x^{-s-2k}\eta(x/N),$$

which is bounded by about $O((|t|/N)^{2k})$. Taking $N = c|t|$ with c a constant, and choosing k proportional to $\log T$, we obtain, using (3.9), that the error term in (3.59) is of size $O\left(|t|^{-C}\right)$, with C arbitrary.

For $\Re s < 0$, it often makes more sense to approximate $\zeta(1-s)$ and then use the functional equation. In fact, even for $0 < \Re s < 1$, it is often preferable to use a method based on the functional equation, called the *approximate functional equation*, or the *Riemann-Siegel formula*, which is a variant thereof. The approximate functional equation involves $O(\sqrt{|t|})$ terms, whereas the sum (3.59) has $2N \gg |t|$ terms. In §4.3.2, we shall discuss these matters further in a computational context for the more general case of Dirichlet L-functions $L(s, \chi)$.

3.5.3 Zeros of $\zeta(s)$.

3.5.3.1 Completed zeta function. Trivial and non-trivial zeros. Critical line and critical strip.

It is an immediate consequence of the functional equation (3.54) that the *completed zeta function* $\xi(s)$, defined by

$$\xi(s) = \frac{1}{2}s(s-1)\pi^{-\frac{s}{2}}\Gamma\left(\frac{s}{2}\right)\zeta(s), \tag{3.60}$$

satisfies

$$\xi(s) = \xi(1-s) \tag{3.61}$$

and is entire. Since $\Gamma(s)$ has simple poles at $s = 0, -1, -2, \ldots$ and vanishes nowhere, it follows that $\zeta(s)$ has zeros at $s = -2, -4, -6, \ldots$ (called *trivial zeros*). We know that $\zeta(s) \neq 0$ for $\Re s > 1$, because the series for $1/\zeta(s)$ converges in that region; hence, by the functional equation, $\zeta(s)$ has no zeros with $\Re s < 0$ other than the trivial zeros. All other zeros (called, naturally, *non-trivial zeros*) must lie in the strip $0 \leq \Re s \leq 1$, called the *critical strip*. Yet again by the functional equation, the set of non-trivial zeros of $\zeta(s)$ is invariant under $s \mapsto 1-s$. It is clear that it is also invariant under $s \mapsto \overline{s}$.

The *critical line* is the line $\Re s = 1/2$ in the complex plane, i.e., the axis of symmetry of $\xi(s)$. The *Riemann Hypothesis* (RH) states that every non-trivial zero of $\zeta(s)$ lies on the critical line. As the reader of course knows, the Riemann Hypothesis is still unproved as of the time of writing.

3.5.3.2 Zero-free regions.

Even proving that $\zeta(s)$ has no zeros with $\Re s = 1$ is far from trivial; it was the key step in Hadamard's and de la Vallée-Poussin's proof of the prime number theorem. As was shown by de la Vallée Poussin (see, e.g., [Dav67, §13]), the same method yields that $\zeta(s)$ has no zeros in a region of the form

$$\left\{\sigma + it : \sigma \geq 1 - \frac{1}{C\log|t|},\ |t| \geq 2\right\},$$

with $C > 0$ a constant. Such a region is called a *classical zero-free region*. There are explicit values for C, starting with de la Vallée-Poussin's own work; see [MT15, Table 1] for the history of successive improvements. The smallest value for C known to date is $C = 5.573412$ ([MT15], building on [Kad05]).

We will make only indirect use of zero-free regions (with older, larger values of C, as it happens), in that some of the bounds we will quote in §5.3.1 rely on the bound (5.13) [Dus16, Thm. 1.3], whose proof uses such a region.

There are asymptotically wider zero-free regions for $\zeta(s)$, due to Vinogradov and Korobov [Kor58], [Vin58]. Such regions have been made explicit [Che00], [For02b], but the explicit versions are narrower than known classical zero-free regions for all but very large t. We will make no use of Vinogradov-Korobov-type regions. There are also *zero-density estimates*, showing that there cannot be many non-trivial zeros with

$\Re s \neq 1/2$ in certain ranges. Such estimates can be crucial for some applications (in particular, for proofs of results on primes in short intervals), but, again, we shall not use them; see, however, the remarks on [KL14] in §4.6.

3.5.3.3 *Counting zeros of* $\zeta(s)$.

We will need some standard estimates on the number $N(T)$ of non-trivial zeros of $\zeta(s)$ with $0 \leq \Im s \leq T$, that is, all zeros of $\zeta(s)$ with $0 \leq \Im s \leq T$ and $0 < \Re s < 1$. We count any zero with $\Im s = 0, T$ and $0 < \Re s < 1$ as half a zero. (There are actually no zeros with $\Im s = 0$ and $0 < \Re s < 1$. We should add that all zeros are to be counted with multiplicity, though it is a standard conjecture that $\zeta(s)$ has no zeros of multiplicity > 1.)

We will use the following simple bound only to ensure convergence, and thus will not need an explicit constant, though we could easily obtain it.

Lemma 3.13. *For $T \geq 3$,*

$$N(T+1) - N(T) \ll \log T.$$

Proof. Let $f(z) = \zeta(2 + i(T+1/2) + z)$. We know that

$$|f(0)| = \left|\prod_p \left(1 - p^{-(2+i(T+1/2))}\right)\right| \geq \prod_p (1 - p^{-2}) = \frac{6}{\pi^2} > 0.$$

By (3.59), for any $\sigma < 1$, $\zeta(s) \ll_\sigma |\Im s|^{1-\sigma}$ in the region $\Re s \geq \sigma$, $|\Im s| \geq 1$. Apply Corollary 2.4 (Jensen) with $R = 3$, $r = \sqrt{5}$, say. □

It is possible to give rather precise estimates for $N(T)$; indeed it is classical work, due to Riemann and von Mangoldt. Assume for simplicity that there is no zero s with $\Im s = T$. Thanks to the functional equation (3.61), we know that s is a zero of $\xi(s)$ (that is, a non-trivial zero of $\zeta(s)$) if and only if $1 - s$ is a zero of $\xi(s)$. Hence, $N(T)$ equals half the number of zeros within the rectangular contour $\mathscr{R}$ going from (say) $-1 - iT$ to $2 - iT$, then to $2 + iT$, then to $-1 + iT$, then back down to $-1 - iT$. Then, by Cauchy's theorem,

$$\begin{aligned} N(T) &= \frac{1}{2} \cdot \frac{1}{2\pi} \int_{\mathscr{R}} \frac{\xi'(s)}{\xi(s)} ds = \frac{1}{2} \cdot \frac{1}{2\pi} \int_{\mathscr{R}} \log' \xi(s) ds \\ &= \frac{1}{2\pi} \int_{\mathscr{C}} \log' \xi(s) ds = \frac{1}{2\pi} \Delta_{\mathscr{C}} \arg \xi(s), \end{aligned}$$

where $\mathscr{C}$ is the half of $\mathscr{R}$ going from $1/2 - iT$ to $1/2 + iT$ on half-plane to the right of the line $\Re s = 1/2$, and $\Delta_{\mathscr{C}} \arg \xi(s)$ means $\arg \xi(1/2 + iT) - \arg \xi(1/2 + iT)$ for the branch of arg continuous on $\mathscr{C}$.

By the definition (3.60) of $\xi(s)$,

$$\arg \xi(s) = \arg s(s-1) + \arg \pi^{-\frac{s}{2}} + \arg \Gamma\left(\frac{s}{2}\right) + \arg \zeta(s).$$

It is clear that $\Delta_{\mathscr{C}} \arg s(s-1) = 2\pi$ and $\Delta_{\mathscr{C}} \pi^{-s/2} = -T\log\pi$. By Stirling's formula (3.41), we know that

$$\begin{aligned}\Delta_{\mathscr{C}} \arg\Gamma\left(\frac{s}{2}\right) &= \Im\log\Gamma\left(\frac{iT}{2}+\frac{1}{4}\right) - \Im\log\Gamma\left(-\frac{iT}{2}+\frac{1}{4}\right) = 2\Im\log\Gamma\left(\frac{iT}{2}+\frac{1}{4}\right)\\ &= T\log\left|\frac{iT}{2}+\frac{1}{4}\right| - \frac{1}{2}\arg\left(\frac{iT}{2}+\frac{1}{4}\right) - T + \Im\frac{1}{6s} + O^*\left(\frac{1+3\pi/4}{360|s|^3}\right)\\ &= T\log\frac{T}{2} + \frac{1}{8T} - \frac{1}{2}\left(\frac{\pi}{2}-\frac{1}{2T}\right) - T - \frac{1}{6T} + O^*\left(\frac{0.426}{T^3}\right)\\ &= T\log\frac{T}{2} - T - \frac{\pi}{4} + \frac{5}{24T} + O^*\left(\frac{0.426}{T^3}\right).\end{aligned}$$

Hence

$$N(T) = \frac{T}{2\pi}\log\frac{T}{2\pi e} + S(T) + \frac{7}{8} + \frac{5/48}{\pi T} + O^*\left(\frac{0.068}{T^3}\right), \qquad (3.62)$$

where

$$S(T) = \frac{1}{2\pi}\Delta_{\mathscr{C}}\arg\zeta(s) = \frac{1}{2\pi}\Delta_{\mathscr{C}}\Im\log\zeta(s) = \frac{1}{2\pi}\Im\int_{\mathscr{C}}\frac{\zeta'(s)}{\zeta(s)}ds.$$

Given Lemmas 2.6 and 3.13, and the bound $|\log\zeta(\sigma+iT)| \ll \log T$ (for $\sigma \geq 1/2$, $T \geq 2$), it is an exercise to establish that $S(T) = O(\log T)$. Getting as good an explicit bound as one can takes more work. There is already a good explicit bound in von Mangoldt's work [vM05]. See [Tru12, Table 1] for a list of results since then. The best results to date are [Tru12, Thm. 1] and [Tru14, Thm. 1]: for $T \geq e$,

$$\begin{aligned}|S(T)| &\leq 0.17\log T + 1.998\\ |S(T)| &\leq 0.112\log T + 0.278\log\log T + 2.51.\end{aligned} \qquad (3.63)$$

3.5.4 Growth of $\zeta(s)$

3.5.4.1 Trivial bounds and convexity bounds

It is easy to see that, for any $s = \sigma + it$, $\sigma > 1$, we have $|\zeta(s)| \leq \zeta(\sigma) \ll 1/(\sigma-1)$. It is also straightforward to show, using (3.59), that $|\zeta(s)| \ll \log|t|$ for $\sigma \geq 1$, $|t| \geq 2$, say, and $\zeta(s) \ll_\sigma |t|^{1-\sigma}$ for $\sigma < 1$, $|t| \geq 1$.[2] Let us see how to do better for $\sigma < 1$.

The *Lindelöf hypothesis* states that $|\zeta(1/2+it)| \ll_\epsilon |t|^\epsilon$ for every $\epsilon > 0$. It is not hard to show, using the functional equation, the bound $|\zeta(1+it)| \ll \log|t|$ and Stirling's formula, that $|\zeta(it)| \ll_\epsilon |t|^{1/2+\epsilon}$. It would then follow, by means of Hadamard's three-line theorem that

$$|\zeta(\sigma+it)| \ll_\epsilon \begin{cases}|t|^{1/2-\sigma+\epsilon} & \text{for } 0<\sigma<1/2,\\ |t|^\epsilon & \text{for } \sigma > 1/2.\end{cases}$$

[2] The assumption that $|t|$ is larger than a constant is actually implicit in our use of asymptotic notation $\ll$, provided that we make clear that we are speaking of the case $|t| \to \infty$. We will make that assumption in what follows.

Of course, the Lindelöf hypothesis remains unproved. One can apply the three-line theorem (§2.6.2) using only trivial inputs, namely, the bounds $|\zeta(\sigma+it)| \ll_\sigma 1$ for $\sigma > 1$ and $|\zeta(\sigma+it)| \ll_\sigma |t|^{1/2-\sigma}$ for $\sigma < 0$ (a consequence of the previous bound, by the functional equation). One then obtains that $|\zeta(\sigma+it)| \ll_{\sigma,\epsilon} |t|^{(1-\sigma)/2+\epsilon}$ for $0 < \sigma \leq 1$. In particular, $|\zeta(1/2+it)| \ll_\epsilon |t|^{1/4+\epsilon}$. This bound is called the *convexity bound*. Bounds with lower exponents are called *subconvex bounds*. The record at the time of writing is $|\zeta(1/2+it)| \ll_\epsilon |t|^{\frac{13}{84}+\epsilon}$ [Bou17].

There are explicit versions of the convexity bound ([Bac18], [Leh70]) and of at least one subconvex bound [CG04, Cor. to Thm. 3]. We will not actually make use of them. For our purposes, it will be enough to use L^2 bounds on the tails, rather than L^∞ bounds.

3.5.4.2 L^2 bounds on $\zeta(s)$

Non-explicit bounds on the L^2 norm of $\zeta(\sigma+it)$ (σ fixed, t restricted to an interval) are classical ([Lan09b, 806–819, 905–906], [HL17], [HL22b], [Lit24]; see the introduction to [Ing27] for an exposition). Unlike known L^∞ bounds, they are essentially tight. It should not be surprising that much more is known about $\zeta(\sigma+it)$ in the L^2 norm than in the L^∞ norm, since the Mellin transform is an L^2-isometry.

The following is an explicit L^2 bound on $\zeta(\sigma+it)$ of the form we need. It is a subset of [DHA19, Thm. 1.1].

Proposition 3.14. *Let $0 < \sigma < 1$. Then*

$$\int_{\sigma+iT}^{\sigma+i\infty} \frac{|\zeta(s)|^2}{|s|^2} ds \leq \begin{cases} \frac{\zeta(2\sigma)}{T} + \frac{12.13}{\left(\sigma-\frac{1}{2}\right)(1-\sigma)} \cdot \frac{1}{T^{2\sigma}} & \text{if } \sigma > 1/2, \\ \frac{3\pi}{5}\frac{\log T}{T} + \frac{7.72}{T} & \text{if } \sigma = 1/2, \\ \frac{\zeta(2-2\sigma)}{2\sigma(2\pi)^{1-2\sigma}} \cdot \frac{1}{T^{2\sigma}} + \frac{15.49}{\sigma^2\left(\frac{1}{2}-\sigma\right)} \cdot \frac{1}{T} & \text{if } \sigma < 1/2, \end{cases} \tag{3.64}$$

assuming $T \geq 4$ for $\sigma \neq 1/2$, and $T \geq 200$ for $\sigma = 1/2$.

Two distinct methods are used to establish these bounds. One of them is traditional: we bound

$$\int_{\sigma-iT}^{\sigma+iT} |\zeta(s)|^2 ds$$

for $\sigma > 1/2$ by first approximating $\zeta(\sigma+it)$ by a finite sum $S(t)$ (via Euler-Maclaurin), and then use mean-value theorems to bound the L^2 norm of $S(t)$ on $[-T, T]$. Lastly, we use the functional equation to obtain bounds for $\sigma < 1/2$ from bounds for $\sigma > 1/2$.

The same argument works for $\sigma = 1/2$, but the following alternative method gives better bounds for $\sigma = 1/2$ and a broad range of T. (In general, the alternative method can be better for small T, whereas the traditional method is better for T large.) By what we saw in §2.5.2, for $\Re s > 1$, $\zeta(s)/s$ is the Mellin transform of $x \mapsto \sum_n 1_{[0,1/n]}(x) = \lfloor 1/x \rfloor$. We choose a continuous approximation g to $1_{[0,1]}$, and denote its Mellin transform by G. Then $G(s)\zeta(s)$ is the Mellin transform of $\sum_n g(nx)$. The difference $\zeta(s)/s - G(s)\zeta(s)$ extends holomorphically to $\Re s > 0$.

Its L^2 norm in that region can thus be determined using the fact that the Mellin transform is an L^2-isometry. If g is identical to $1_{[0,1]}$ outside a small neighborhood of 1, then $\zeta(s)/s - G(s)\zeta(s)$ is in fact the Mellin transform of a function whose L^2 norm is easy to determine. We finish by choosing a good approximation g to $1_{[0,1]}$ such that $G(s)$ is small compared to $1/s$ for $s = \sigma + it$, $|t| \geq T$, so that $\zeta(s)/s - G(s)\zeta(s)$ is a good approximation to $\zeta(s)/s$ for those s. See [DHA19, §3.1–3.3] for a full account.

Yet a third possibility is to follow an approach based on the approximate functional equation (see (4.17)). This is the route followed by [Sim]. It would also seem feasible to give explicit estimates on all terms occurring in Atkinson's formula [Atk49] for $\sigma = 1/2$, or its generalization in [MM93] to $1/2 \leq \sigma \leq 1$. One could even go further and work towards an explicit version of [Bal78]. It seems clear that using the approximate functional equation is asymptotically the best of the three approaches. We will make do with (3.64), as we simply need some tail bounds for computations, and what we have is enough.

3.5.4.3 Bounding $\zeta'(s)/\zeta(s)$ and $1/\zeta(s)$

Some words on bounding $\zeta'(s)/\zeta(s)$ and $1/\zeta(s)$ are in order. It is clear that these functions can be bounded only within a zero-free region, as both functions have poles where $\zeta(s)$ has zeros. To bound $\zeta'(s)/\zeta(s)$, we can use Lemma 2.6. To bound $1/\zeta(s)$, we can bound $\log 1/|\zeta(s)| = -\Re \log \zeta(s)$ by bounding the derivative $(\log \zeta(s))' = \zeta'(s)/\zeta(s)$; see [Tit86, §3.10].

In this way, [Tru15a] proves the following explicit bounds [Tru15a, Table 2] for $s = \sigma + it$:

$$\begin{aligned} \left|\frac{\zeta'(s)}{\zeta(s)}\right| &\leq 80.38 \log t, & \left|\frac{1}{\zeta(s)}\right| &\leq 3.1 \cdot 10^6 \log t \quad \text{for } t \geq 51,\ \sigma \geq 1 - \frac{1}{8 \log t}, \\ \left|\frac{\zeta'(s)}{\zeta(s)}\right| &\leq 40.14 \log t, & \left|\frac{1}{\zeta(s)}\right| &\leq 1900 \log t \quad \text{for } t \geq 133,\ \sigma \geq 1 - \frac{1}{12 \log t}. \end{aligned} \tag{3.65}$$

It is clear that the constants in the bounds for $1/\zeta(s)$ are an issue, and the constants in the bounds for $\zeta'(s)/\zeta(s)$ can be one. It is then no wonder that, as we shall see in §5.3, serious work has been put into how to avoid working with $1/\zeta(s)$ directly.

3.5.5 Writing $\zeta(s)$ in terms of its zeros. Special values.

We can now write $\zeta(s)$ (or $\xi(s)$) in terms of its zeros, as Hadamard did.[3] By Stirling's formula (3.41) and an easy bound on $\zeta(s)$ for $\sigma \geq 1/2$ (as at the beginning of §3.5.4), the completed zeta function $\xi(s)$ (defined in (3.60)) obeys

$$|\xi(s)| \ll e^{|s|^{1+\epsilon}}$$

[3] For historical remarks on much of the basic theory, see [Lan09a, K. 1–2].

as $|s| \to \infty$, for any $\epsilon > 0$. Hence, we can apply Lemma 2.11, and obtain that

$$\xi(s) = e^{A+Bs} \prod_{\rho} \left(1 - \frac{s}{\rho}\right) e^{-s/\rho}$$

for some $A, B \in \mathbb{C}$, where the product is over the zeros of $\xi(s)$, i.e., the non-trivial zeros of $\zeta(s)$; by (2.40),

$$\frac{\xi'(s)}{\xi(s)} = B + \sum_{\rho} \left(\frac{1}{s-\rho} + \frac{1}{\rho}\right), \tag{3.66}$$

where ρ ranges over the non-trivial zeros of $\zeta(s)$.

By definition (3.60),

$$\frac{\xi'(s)}{\xi(s)} = \frac{1}{s} + \frac{1}{s-1} - \frac{1}{2}\log\pi + \frac{1}{2}\digamma(s/2) + \frac{\zeta'(s)}{\zeta(s)}. \tag{3.67}$$

It is clear from (3.66) that

$$B = \frac{\xi'(0)}{\xi(0)} = -\frac{\xi'(1)}{\xi(1)}.$$

By (3.49), (3.53) and (3.67),

$$\frac{\xi'(1)}{\xi(1)} = 1 - \frac{1}{2}\log\pi + \frac{1}{2}(-\gamma - 2\log 2) + \gamma = 1 - \log 2 + \frac{\gamma}{2} - \frac{\log\pi}{2}.$$

Hence $B = -1 - \gamma/2 + (\log 4\pi)/2$.

We also obtain from (3.48) and (3.67) that

$$\frac{\zeta'(0)}{\zeta(0)} = \frac{\gamma}{2} + \frac{\log\pi}{2} + 1 + \frac{\xi'(0)}{\xi(0)} = \log 2\pi. \tag{3.68}$$

3.6 DIRICHLET CHARACTERS

A *Dirichlet character* $\chi : \mathbb{Z} \to \mathbb{C}$ of modulus q is a character χ of $(\mathbb{Z}/q\mathbb{Z})^*$ lifted to $\mathbb{Z}$ with the convention that $\chi(n) = 0$ when $(n,q) \neq 1$. (In other words: χ is completely multiplicative and periodic modulo q, and vanishes on integers not coprime to q.) Again by convention, there is a Dirichlet character of modulus $q = 1$, namely, the *trivial character* $\chi_T : \mathbb{Z} \to \mathbb{C}$ defined by $\chi_T(n) = 1$ for every $n \in \mathbb{Z}$.

3.6.1 Orthogonality. Primitive and imprimitive characters.

It is simple to see that, for any abelian finite group G,

$$\frac{1}{|G|}\sum_{\chi \in \hat{G}} \chi(g) = \begin{cases} 1 & \text{if } g = 1, \\ 0 & \text{otherwise,} \end{cases} \qquad \frac{1}{|G|}\sum_{g \in G} \chi(g) = \begin{cases} 1 & \text{if } \chi = \chi_0, \\ 0 & \text{otherwise,} \end{cases} \tag{3.69}$$

where $\hat{G}$ is the group of characters of G and χ_0 denotes the identity element of $\widehat{G}$. As follows easily from (3.69), the elements of $\widehat{G}$ are an orthonormal basis for the vector space of functions $G \to \mathbb{C}$ endowed with the inner product

$$\langle v, w\rangle = \sum_{g\in G} \overline{v(g)} w(g).$$

(Here and in §3.6.2, we will find it convenient to work with inner products that correspond to the uniform probability measure on finite sets such as G, rather than to the counting measure.)

It follows that, for any $q \geq 1$ and $n \in \mathbb{Z}$,

$$\frac{1}{q} \sum_{\chi \bmod q} \chi(n) = \begin{cases} 1 & \text{if } n \equiv 1 \bmod q, \\ 0 & \text{otherwise,} \end{cases}$$

where χ ranges over all Dirichlet characters $\bmod q$. Moreover, Dirichlet characters $\bmod q$ form a basis for the vector space of functions $\mathbb{Z} \to \mathbb{C}$ of period q supported on integers coprime to q, and that basis is orthonormal with respect to the inner product $\langle v, w\rangle = (1/\phi(q)) \sum_{a \bmod q} \overline{v(n)} w(n)$.

If χ is a character modulo q and χ' is a character modulo $q'|q$ such that $\chi(n) = \chi'(n)$ for all n coprime to q, we say that χ' *induces* χ. A character is *primitive* if it is not induced by any character of smaller modulus. Characters that are not primitive are called *imprimitive*.

Given a character χ, we write χ^* for the (uniquely defined) primitive character inducing χ. (The uniqueness of the primitive character inducing χ is an elementary but non-obvious fact; see, e.g., [Spi70] or [MV07, §9.1].) If χ^* is of modulus q, we say q is the *conductor* of χ.

The identity χ_0 of the group of characters $\bmod q$ is called the *principal* character $\bmod q$. It satisfies $\chi_0(n) = 1$ when $(n, q) = 1$ and $\chi_0(n) = 0$ when $(n, q) \neq 1$. A principal character is induced by the trivial character χ_T, and thus has conductor 1.

3.6.2 Multiplicative and additive characters

We will often need to pass from additive characters $e(\alpha n)$ to Dirichlet characters and vice versa. We can already establish some useful lemmas by means of basic identities, as we are about to see. More detailed work requires studying the Fourier coefficients of Dirichlet characters, called *Ramanujan sums* (3.6.3).

The most basic ℓ^2 lemma is just a matter of renormalization.

Lemma 3.15. *Let $\{a_n\}_{n=1}^{\infty}$, $a_n \in \mathbb{C}$, be in ℓ^1. Assume that $a_n = 0$ for n not coprime to q. Then*

$$\sum_{a \bmod q} \left| \sum_n a_n e\left(\frac{an}{q}\right) \right|^2 = \frac{1}{\phi(q)} \sum_{\chi \bmod q} \left| \sum_n a_n \chi(n) \right|^2. \tag{3.70}$$

Proof. The maps $n \to e(an/q)$, $a \in \mathbb{Z}/q\mathbb{Z}$, form an orthonormal basis of the space of functions $\mathbb{Z}/q\mathbb{Z} \to \mathbb{C}$ with the inner product $\langle f, g\rangle = (1/q) \sum_{m\in\mathbb{Z}/q\mathbb{Z}} \overline{f(m)} g(m)$.

Applying Parseval's identity to the function $f : \mathbb{Z}/q\mathbb{Z} \to \mathbb{C}$ defined by $f(m) = |\sum_{n\equiv m} a_n|^2$, we obtain

$$\frac{1}{q}\sum_{a \bmod q}\left|S\left(\frac{a}{q}\right)\right|^2 = \sum_{m\in\mathbb{Z}/q\mathbb{Z}} |f(m)|^2,$$

where $S(\alpha) = \sum_n e(\alpha n)$.

At the same time, the maps $n \to \chi(n)$, χ a character of $(\mathbb{Z}/q\mathbb{Z})^*$, are an orthonormal basis of the space of functions $(\mathbb{Z}/q\mathbb{Z})^* \to \mathbb{C}$ with the inner product $\langle f, g\rangle = (1/\phi(q))\sum_{m\in\mathbb{Z}/q\mathbb{Z}} \overline{f(m)}g(m)$. Hence, again by Parseval's identity,

$$\frac{1}{\phi(q)}\sum_{\chi \bmod q}\left|\sum_n \chi(n)a_n\right|^2 = \sum_{m\in(\mathbb{Z}/q\mathbb{Z})^*} |f(m)|^2.$$

Since a_n is supported on integers coprime to q,

$$\sum_{m\in(\mathbb{Z}/q\mathbb{Z})^*} |f(m)|^2 = \sum_{m\in\mathbb{Z}/q\mathbb{Z}} |f(m)|^2,$$

and so we obtain (3.70). □

We can also ask ourselves what happens when we restrict our attention to characters $n \mapsto e(an/q)$ with a coprime to q, and to primitive characters $\chi \bmod q$. The following lemma will prove particularly useful in the large sieve and related contexts; it can already be found in, say, [Bom74, pp. 24–25].

Lemma 3.16. *Let $\{a_n\}_{n=1}^{\infty}$, $a_n \in \mathbb{C}$, be in ℓ^1. Assume that $a_n = 0$ for n not coprime to q. Then*

$$\sum_{\substack{a \bmod q\\(a,q)=1}}\left|\sum_n a_n e\left(\frac{an}{q}\right)\right|^2 = \sum_{\substack{q^*|q\\(q^*,q/q^*)=1\\ \mu^2(q/q^*)=1}} \frac{q^*}{\phi(q)}\cdot \sideset{}{^*}\sum_{\chi \bmod q^*}\left|\sum_n a_n\chi(n)\right|^2, \qquad (3.71)$$

where $\sum_\chi^*$ denotes a sum taken over primitive characters only.

Proof. Write $S(\alpha)$ for $\sum_n a_n e(\alpha n)$. By Lemma 3.15 and inclusion-exclusion,

$$\begin{aligned}\sum_{\substack{a \bmod q\\(a,q)=1}}\left|S\left(\frac{a}{q}\right)\right|^2 &= \sum_{q'|q}\mu\left(\frac{q}{q'}\right)\sum_{a \bmod q'}\left|S\left(\frac{a}{q'}\right)\right|^2\\ &= \sum_{q'|q}\mu\left(\frac{q}{q'}\right)\frac{q'}{\phi(q')}\sum_{\chi \bmod q'}\left|\sum_n a_n\chi(n)\right|^2\\ &= \sum_{q'|q}\mu\left(\frac{q}{q'}\right)\frac{q'}{\phi(q')}\sum_{q^*|q'}\sideset{}{^*}\sum_{\chi \bmod q^*}\left|\sum_n a_n\chi(n)\right|^2.\end{aligned}$$

It is easy to check that, for any $q^*|q$,

$$\sum_{q':q^*|q'|q} \mu\left(\frac{q}{q'}\right)\frac{q'}{\phi(q')} = \begin{cases} q^*/\phi(q) & \text{if } q/q^* \text{ is square-free and coprime to } q^* \\ 0 & \text{otherwise,} \end{cases}$$

and so we are done. □

3.6.3 Ramanujan sums

Given a Dirichlet character χ of modulus q and an element $b \in \mathbb{Z}/q\mathbb{Z}$, we write $\tau(\chi, b)$ for the *Gauss sum*

$$\tau(\chi, b) = \sum_{a \bmod q} \chi(a)e(ab/q) \tag{3.72}$$

associated to a Dirichlet character χ with modulus q. The fact that Gauss sums appear frequently in number theory should be unsurprising, as $b \mapsto \tau(\chi, -b)$ is simply the Fourier transform of χ, seen as a function from $\mathbb{Z}/q\mathbb{Z}$ to $\mathbb{C}$.

We let

$$\tau(\chi) = \tau(\chi, 1) = \sum_{a \bmod q} \chi(a)e(a/q). \tag{3.73}$$

If $(b, q) = 1$, then it is easy to see that $\tau(\chi, b) = \chi(b^{-1})\tau(\chi)$.

For χ principal, $\tau(\chi, b)$ equals the *Ramanujan sum*

$$c_q(b) = \sum_{\substack{a \bmod q \\ (a,q)=1}} e(ab/q). \tag{3.74}$$

By Möbius inversion in the form (2.1),

$$c_q(b) = \sum_{a \bmod q}\left(\sum_{d|(a,q)} \mu(d)\right) e(ab/q) = \sum_{d|q} \mu(d) \sum_{a \bmod q/d} e\left(\frac{ab}{q/d}\right) = \sum_{\substack{d|q \\ \frac{q}{d}|b}} \mu(d)\frac{q}{d},$$

and so

$$c_q(b) = \sum_{d|(b,q)} \mu\left(\frac{q}{d}\right) d. \tag{3.75}$$

In particular,

$$c_q(b) = \mu\left(\frac{q}{(b,q)}\right)\phi((b,q)) \qquad \text{for } q \text{ square-free,}$$

$$c_q(b) = \mu(q) \qquad \text{if } (b,q) = 1. \tag{3.76}$$

Let us bound the Fourier transform of a primitive character. The following bound is completely standard; it is essentially due to Gauss for q prime.

Lemma 3.17. *Let χ be a primitive Dirichlet character of conductor q. Then*

$$|\tau(\chi, b)| = \begin{cases} \sqrt{q} & \text{if } b \text{ is coprime to } q, \\ 0 & \text{otherwise.} \end{cases}$$

Proof. We know that $\tau(\chi, b) = \chi(b^{-1})\tau(\chi)$ for b coprime to q. Since χ is supported on integers coprime to q,

$$|\tau(\chi)|^2 = \sum_{\substack{a,a' \bmod q \\ (a,q)=1 \\ (a',q)=1}} \chi(a')\overline{\chi(a)} e\left(\frac{a'-a}{q}\right) = \sum_{b \bmod q} \chi(b) \sum_{\substack{a \bmod q \\ (a,q)=1}} e\left(\frac{(b-1)a}{q}\right).$$

The inner sum here is a Ramanujan sum. We apply (3.75), and obtain

$$|\tau(\chi)|^2 = \sum_{\substack{b \bmod q \\ (b,q)=1}} \chi(b) \sum_{d|(b-1,q)} \mu\left(\frac{q}{d}\right) d = \sum_{d|q} \mu\left(\frac{q}{d}\right) d \sum_{\substack{b \bmod q \\ b \equiv 1 \bmod d}} \chi(b). \tag{3.77}$$

For $d = q$, the innermost sum equals 1. For $d|q$ with $d \neq q$, the map $r \mapsto r \bmod q/d$ is an isomomorphism from $\{b \in (\mathbb{Z}/q\mathbb{Z})^* : b \equiv 1 \bmod d\}$ to $(\mathbb{Z}/(q/d)\mathbb{Z})^*$. In this way, χ defines a character $\chi_d \bmod q/d$, and

$$\sum_{\substack{b \bmod q \\ b \equiv 1 \bmod d}} \chi(b) = \sum_{x \bmod q/d} \chi_d(x) \tag{3.78}$$

is non-zero only if χ_d is trivial. If χ_d is trivial, then $\chi(b) = 1$ for all $b \equiv 1 \bmod d$, and so χ is periodic $\bmod d$, meaning it is not principal; contradiction. We conclude that the sum on the left of (3.78) is 0 for all $d|q$ with $d \neq q$, and so, by (3.77), $|\tau(\chi)|^2 = q$. Thus, $|\tau(\chi, b)|^2 = q$ for b coprime to q. At the same time, by Plancherel,

$$\begin{aligned} \phi(q) = \sum_{x \bmod q} |\chi(x)|^2 = \frac{1}{q} \sum_{b \bmod q} |\tau(\chi, b)|^2 &= \frac{1}{q}\left(\sum_{\substack{b \bmod q \\ (b,q)=1}} q + \sum_{\substack{b \bmod q \\ (b,q)\neq 1}} |\tau(\chi, b)|^2\right) \\ = \phi(q) + \frac{1}{q} \sum_{\substack{b \bmod q \\ (b,q)\neq 1}} |\tau(\chi, b)|^2 . \end{aligned}$$

Hence, $\tau(\chi, b) = 0$ for every b not coprime to q. □

The uses of Lemma 3.17 are many. For instance, given Lemmas 3.3 and 3.17, the Pólya-Vinogradov inequality becomes an exercise.

Exercise 3.18. (Pólya-Vinogradov inequality) Let χ be a primitive Dirichlet character of conductor $q > 1$. Then, for any interval $I \subset \mathbb{R}$,

$$\left|\sum_{n \in I \cap \mathbb{Z}} \chi(n)\right| \leq \sqrt{q} \log q. \tag{3.79}$$

Remark. The factor of $\log q$ in (3.79) is there because the sum on the left side of (3.79) is not smoothed. It is not any harder to obtain an estimate without a factor of $\log q$ for smoothed sums $\sum_n \chi(n) f(n)$. (We will find ourselves in an analogous situation in §11.2, when we consider trigonometric sums arising in Vinogradov's approach to the ternary Goldbach problem.) There are also (many) versions of the Pólya-Vinogradov inequality that are sharper than (3.79), even for non-smoothed sums. See, e.g., [FS13] or [Pom11].

3.7 DIRICHLET L FUNCTIONS

All we are about to say can be found in, say, [Dav67], or in nearly any other introductory textbook on analytic number theory. See [MV07, Ch. 9–10] for a far more detailed exposition and [IK04, Ch. 5] for both more detail and much greater generality. We already went over the case of the trivial character in §3.5: the Riemann zeta function is the L-function corresponding to the trivial character.

3.7.1 Definition. Analytic continuation and functional equation.

Let χ be a Dirichlet character. The *Dirichlet L-function* $L(s, \chi)$ is defined by

$$L(s,\chi) = \sum_{n=1}^{\infty} \chi(n) n^{-s} \tag{3.80}$$

for $\Re s > 1$, and by analytic continuation on the rest of the complex plane. The function $L(s, \chi)$ has a pole at $s = 1$ if χ is principal, and no poles otherwise. As before, continuing the function analytically to $\Re s > 0$ is easy, and the existence of an analytic continuation to all of $\mathbb{C}$ will follow from the proof of the functional equation that we are about to state.

The *functional equation* for Dirichlet L-functions tells us that, for χ primitive,

$$\Lambda(s,\chi) = \varepsilon(\chi)\Lambda(1-s,\overline{\chi}), \tag{3.81}$$

where

$$\Lambda(s,\chi) = \left(\frac{q}{\pi}\right)^{\frac{s+\kappa}{2}} \Gamma\left(\frac{s+\kappa}{2}\right) L(s,\chi), \tag{3.82}$$

$$\varepsilon(\chi) = \frac{\tau(\chi)}{i^{\kappa}\sqrt{q}}, \qquad \kappa = \begin{cases} 0 & \text{if } \chi(-1) = 1, \\ 1 & \text{if } \chi(-1) = -1. \end{cases} \tag{3.83}$$

The function $\Lambda(s, \chi)$ is called a *completed L-function*. It is entire, unless χ is trivial, in which case $L(s, \chi)$ equals $\zeta(s)$, which has a simple pole at $s = 1$, and so then $\Lambda(s, \chi)$ also has a simple pole at $s = 1$.

The functional equation can be proved much as we proved it in the case of χ trivial (§3.5.1). (What follows is a summary; for a full account, see [IK04, §4.6], or simply work out the details using §3.5.1 as a model.) Let χ be primitive. Then

$$\Lambda(s,\chi) = \frac{1}{2}\int_0^\infty \theta_\chi\left(\frac{ix}{2q}\right) x^{\frac{s+\kappa}{2}-1} dx,$$

where θ_χ is the theta function corresponding to χ:

$$\theta_\chi(z) = \sum_{n\in\mathbb{Z}} \chi(n) n^\kappa e(n^2 z), \tag{3.84}$$

with κ as in (3.83). Much as in the case of χ trivial, the functional equation (3.81) – a symmetry law for $L(s,\chi)$ – reduces to a symmetry law for $\theta_\chi(x)$:

$$\theta_\chi(ix/2q) = \epsilon(\chi) x^{-\kappa-\frac{1}{2}} \theta_{\overline{\chi}}(i/2xq). \tag{3.85}$$

We prove this law by first decomposing $\theta_\chi(z) = \sum_{a \bmod q} \chi(a)\theta(z;q,a)$, where

$$\theta(z;q,a) = \sum_{\substack{n\in\mathbb{Z} \\ n\equiv a \bmod q}} \chi(n) n^\kappa e(n^2 z).$$

Then we apply the Poisson summation formula (in the form (3.14)), using the fact that the Gaussian is self-dual, or, if $\kappa = 1$, using the fact $\widehat{g}(t) = g(t)/i$ for $g(t) = te^{-\pi t^2}$, as follows readily from the self-duality of the Gaussian together with rule (2.14). We thus obtain the functional equation (3.81).

3.7.2 Approximating $L(s,\chi)$.

We may write

$$L(s,\chi) = q^{-s} \sum_{a=1}^{q} \chi(a)\zeta(s, a/q), \tag{3.86}$$

where $\zeta(s,\alpha)$ is a *Hurwitz zeta function*:

$$\zeta(s,\alpha) = \sum_{n=0}^{\infty} (n+\alpha)^{-s} \tag{3.87}$$

for $\Re s > 1$.

To compute $\zeta(s,\alpha)$, we may use Euler-Maclaurin, much as in §3.5.2. We obtain that, for $\sigma > 1-2k$,

$$\zeta(s,\alpha) = \sum_{n=1}^{2N} ((n+\alpha)^{-s} - f(n)) + c_{\eta,\alpha}\frac{N^{1-s}}{1-s} + O^*\left(\frac{|B_{2k}|}{(2k)!}\int_N^\infty \left|f^{(2k)}(x)\right| dx\right), \tag{3.88}$$

where $f(x) = \eta(x/N)/(x+\alpha)^s$, $\eta : \mathbb{R} \to [0,1]$ is a smooth function of our choice (with $\eta(x) = 0$ for $x \leq 1$ and $\eta(x) = 1$ for $x \geq 2$, say) and $c_{\eta,\alpha}$ is a constant. The challenging case is, as usual, that of $t = \Im s$ large. For N proportional to $|t|$ and k proportional to $\log |t|$, the error term in (3.88) is $O(|t|^{-C})$, where C is arbitrary.

We can evaluate the sum in (3.88) by brute force, in time linear on N, and hence linear on $|t|$. Thus, we can give an arbitrarily good approximation to $L(s,\chi)$ in time $O(q|t|)$. We will discuss later (§4.3.2) how to do better, whether for an individual χ or when we must compute $L(s,\chi)$ for many values of χ.

Just as in the case of $\zeta(s)$, we also obtain an easy bound on $L(s,\chi)$, namely, $|L(s,\chi)| \ll_\sigma q|t|^{1-\sigma}$ for $\sigma < 1$, $\Re s \geq \sigma$ and, say, $|t| \geq 1$. (One can easily give a uniform bound $L(s,\chi) \ll \log qt$ for χ non-principal, $\Re s \geq 1$, $|t| \geq 1$ by partial summation, as in [Ten15, Thm. 8.18], and of course we also have $|L(s,\chi)| \leq \zeta(\sigma)$.) Again, we could do better – in particular, by means of convexity, as in §3.5.4; see [Rad60, Thm. 3] for an explicit bound – but these simple bounds are already useful.

3.7.3 Zeros of $L(s,\chi)$.

3.7.3.1 Trivial and non-trivial zeros. Critical line and critical strip.

Much as for the Riemann zeta function, it is an immediate consequence of the functional equation (3.81) that the zeros of $\Lambda(s,\chi)$ are symmetric, with the line $\Re s = 1/2$ as the axis of symmetry: the zeros of $\Lambda(s,\chi)$ are the same as the zeros of $\Lambda(1-s,\overline{\chi})$, and, since $\Lambda(1-s,\overline{\chi}) = \overline{\Lambda(1-\overline{s},\chi)}$, we obtain that the set of zeros of $\Lambda(s,\chi)$ is invariant under the map $s \mapsto 1-\overline{s}$, i.e., $\sigma + i\tau \mapsto (1-\sigma) + i\tau$. If χ is a real-valued character, then the zeros of $\Lambda(s,\chi)$ are invariant under the map $s \to \overline{s}$ as well.

At this point we must distinguish between the zeros of $L(s,\chi)$ and those of $\Lambda(s,\chi)$. Recall that $\Gamma(s)$ has a pole at every non-positive integer s, while its inverse $1/\Gamma(s)$ is entire. Thus, for χ principal, the set of zeros of $L(s,\chi)$ is the set of zeros of $\Lambda(s,\chi)$, plus, possibly, zeros at non-positive integers. In fact, it follows from (3.81) that $L(s,\chi)$ does have zeros at some non-positive integers: $s = -1,-3,-5,\dots$ if $\kappa = 1$, $s = 0,-2,-4,\dots$ if $\kappa = 0$ and χ is non-trivial, and $s = -2,-4,-6,\dots$ if χ is trivial.

For χ a non-principal character mod q induced by a primitive character χ^* mod d, it follows from the definition of $L(s,\chi)$ that

$$L(s,\chi) = L(s,\chi^*) \cdot \prod_{p|q/d} \left(1 - \frac{\chi^*(p)}{p^s}\right).$$

Hence, $L(s,\chi)$ has zeros at all purely imaginary $s = it$ such that $p^{it} = \chi^*(p)$, besides having zeros at all the zeros of $L(s,\chi^*)$.

A zero of $L(s,\chi)$ is called *non-trivial* if it is a zero of $\Lambda(s,\chi^*)$, for χ^* the primitive character inducing χ. The other zeros of $L(s,\chi)$ are called *trivial*. Just as for the Riemann zeta function, it is easy to show that $L(s,\chi)$ has no zeros with $\Re s > 1$, and thus no zeros with $\Re s < 0$ other than the trivial ones.

In contrast, just as in the special case $L(s,\chi) = \zeta(s)$, the fact that $L(s,\chi)$ has no zeros with $\Re s = 1$ is a non-trivial fact of deep arithmetical significance; for instance, just proving that $L(s,\chi)$ does not have a zero at $s = 1$ enabled Dirichlet to prove that there are infinitely many primes in an arithmetic progression $a + m\mathbb{Z}$, $\gcd(a,m) = 1$.

See, e.g., [Dav67, §1], [MV07, §4.3, §11.1]. It follows, again by (3.81), that $L(s,\chi)$ has no non-trivial zeros with $\Re s = 0$. If χ is primitive, then $L(s,\chi)$ has no zeros with $\Re s = 0$ altogether other than possibly at $s = 0$.

Just as for the Riemann zeta function, the *critical line* is the line $\Re s = 1/2$, i.e., the axis of symmetry of L-functions. The *Generalized Riemann Hypothesis* (GRH) for Dirichlet L-functions reads: for every Dirichlet character χ, all non-trivial zeros of $L(s,\chi)$ lie on the critical line. GRH is of course unproved to date.

For $H > 0$, we say that "GRH(H) holds for $L(s,\chi)$" or "$L(s,\chi)$ satisfies GRH(H)" if every non-trivial zero ρ of $L(s,\chi)$ satisfies either $\Re\rho = 1/2$ or $|\Im\rho| > H$.

3.7.3.2 *Explicit formulae. Zero-free regions.*

The central role played by L-functions in analytic number theory is due in part to their use in estimating the sum of $\chi(p)$ over all primes $p \leq N$, or, more generally, sums of the form

$$\sum_n \Lambda(n)\chi(n)f(n), \tag{3.89}$$

where $\Lambda(n)$ is the von Mangoldt function and $f : \mathbb{R}^+ \to \mathbb{C}$ is a piecewise continuous function of fast enough decay. An expression for (3.89) in terms of $L(s,\chi)$ is called an *explicit formula*. We already discussed explicit formulae briefly in the introduction; we will treat them at much greater length in Chapter 16.

There are zero-free regions and zero-density estimates for L-functions analogous to those for $\zeta(s)$. The most notable difference consists in the difficulty of showing, for χ a non-principal real character, that there is no real zero of $L(s,\chi)$ very close to 1. We do know that, for any Dirichlet character $\chi \bmod q$, $q > 1$, $L(s,\chi)$ has no zeros in the region

$$\left\{\sigma + it : \sigma \geq 1 - \frac{1}{C\log\max(q, q|t|)}\right\}, \tag{3.90}$$

with the possible exception of one real zero when χ is a non-principal character. The lowest value of C known to date is $C = 6.4355$ [Kad02].[4]

A (hypothetical) real zero outside the region (3.90) is called, appropriately, an *exceptional zero*, or, alternatively, *Siegel zero* or *Landau-Siegel zero*. Landau [Lan18a] showed that an exceptional zero could occur for at most one character $\chi \bmod q$ and at most one q for q in a broad range. In Page's formulation [Pag35]: there is a constant C such that, for every $Q \geq 2$, there is at most one real primitive character $\chi \bmod q$, $q \leq Q$, such that $L(s,\chi)$ has a real zero $s = \sigma$ with $\sigma > 1 - c/\log Q$. The constant c is effective; one can easily obtain an explicit value for it from [McC84b] or [Kad02].

Siegel proved [Sie35] that every real zero σ of $L(s,\chi)$ for any Dirichlet character χ satisfies

$$\sigma \leq 1 - \frac{c_\epsilon}{q^\epsilon} \tag{3.91}$$

[4] A further improvement seems to be upcoming. Previously, McCurley [McC84b] gave a zero-free region as in (3.90) with $C = 9.645908801$.

for every $\epsilon > 0$, where the constant $c_\epsilon > 0$ is ineffective. (Landau had proved the weaker bound $\sigma \leq 1 - c'_\epsilon/q^{1/8+\epsilon}$, with c'_ϵ also ineffective, shortly before, in [Lan35].) In fact, as pointed out by Tatuzawa [Tat51], Siegel's work implies that (3.91) holds with an effective implied constant for all characters save possibly by those induced by a single primitive, real-valued character χ.

If one wants an effective bound that holds for all χ, without exceptions, the best bound we can currently give for a real zero σ of $L(s,\chi)$, χ a non-principal real character, is of the form

$$\sigma \leq 1 - \frac{c_A}{\sqrt{q}\log^A q}, \tag{3.92}$$

where $A = 2$ is known since [Pag35] and $1 < A < 2$ is due to Goldfeld and Gross-Zagier (see [Gol85]). For $A = 2$, (3.92) holds with $c_A = 40$ [BMOR18]. The approach for $1 < A < 2$ relies on proving non-trivial lower bounds for *class numbers* of imaginary quadratic fields; there is an explicit version of Goldfeld's argument in [Oes88].

We will not use zero-free regions (whether as in (3.90) or of Vinogradov type) for L-functions other than the zeta function. It goes without saying that we also will not use Siegel's bound (3.91), since our results must be effective. What we will use is partial verifications of the Generalized Riemann Hypothesis, that is, computations that prove that a given L-function satisfies GRH(H) for some constant H. We shall discuss such verifications in §4.3. In particular, they rule out exceptional zeros for $L(s,\chi)$.

3.7.3.3 Counting zeros of $L(s,\chi)$.

Write $N(T,\chi)$ for the number of non-trivial zeros of $L(s,\chi)$ with $0 \leq \Im s \leq T$, that is, the number of zeros of $\zeta(s)$ with $0 < \Re s < 1$ and $0 \leq \Im s \leq T$. As in the case of $\zeta(s)$, we count only half of any zero with $\Im s = 0, T$ and $0 < \Re s < 1$, and do our counting with multiplicity, though the multiplicity of every zero of $L(s,\chi)$ is believed to be 1.

We have again a simple bound.

Lemma 3.19. *For χ a Dirichlet character $\bmod q$ and $T \geq 3$,*

$$N(T+1,\chi) - N(T,\chi) \ll \log qT.$$

Proof. Follow the proof of Lemma 3.19, using the bound $|L(s,\chi)| \ll_\sigma q|\Im s|^{1-\sigma}$ for $\Re s \geq \sigma$. (We can of course also use bounds with a better exponent than $1-\sigma$.) □

We can also proceed just as in §3.5.3.3 to prove, as in [MV07, Cor. 14.6], that, for χ primitive mod q, $q \geq 1$,

$$N(T,\chi) = \frac{T}{2\pi}\log\frac{qT}{2\pi e} + S(T,\chi) - S(0,\chi) - \chi(-1)/8 + O\left(\frac{1}{T}\right), \tag{3.93}$$

for $T \geq 1$, where

$$S(T,\chi) = \frac{1}{\pi}\arg L(1/2+iT,\chi)$$

Here arg is understood so that it varies continuously as s moves on the straight paths from $s = 2$ to $s = 2+iT$ and from $s = 2+iT$ to $s = 1/2+iT$, say, assuming there are no zeros on the latter. Much as before, we can use Lemmas 2.6 and 3.19 together with the bound $L(s,\chi) \ll_\sigma q|\Im s|^{1-\sigma}$ to obtain that, for $T \geq 2$ (say),

$$S(T,\chi) \ll \log qT. \tag{3.94}$$

While making the implied constant in (3.93) into a good explicit constant is easy, getting good constants in (3.94) takes more work. Using a trick of Backlund's, Trudgian [Tru15b, Thm. 1][5] improved on a result by McCurley [McC84a, Thm. 2.1] to obtain

$$N(T,\chi) = \frac{T}{2\pi}\log\frac{qT}{2\pi e} + O^*\left(0.15\log qT + 3.389\right) \tag{3.95}$$

for any χ primitive and non-trivial and $T \geq 1$. By (3.62) and (3.63), and by the fact that the first non-trivial zero of $\zeta(s)$ (that is, the non-trivial zero ρ, or rather the pair of zeros ρ, $\overline{\rho}$, such that $|\Im\rho|$ is minimal) has imaginary part $14.13472\ldots$ ([Gra03], [Edw74, §6.1]), (3.95) also holds for χ trivial and $T \geq 1$.

3.7.4 Writing $L(s,\chi)$ in terms of its zeros. Special values.

We will now write $L(s,\chi)$, or $\Lambda(s,\chi)$, in terms of its zeros, as we did for $\zeta(s)$ and $\xi(s)$ in §3.5.5.[6] Let χ be a primitive Dirichlet character mod q, $q > 1$. By Stirling's formula (3.41) and an easy bound on $L(s,\chi)$ for $\sigma \geq 1/2$ (such as $|L(s,\chi)| \ll q|t|^{1/2}$, in §3.7.2), the completed L-function (defined in (3.82)) satisfies

$$|\Lambda(s,\chi)| \ll e^{|s|^{1+\epsilon}}$$

as $|s| \to \infty$, for any $\epsilon > 0$. Hence, by Lemma 2.11,

$$\Lambda(s,\chi) = e^{A(\chi)+B(\chi)s}\prod_\rho\left(1-\frac{s}{\rho}\right)e^{-s/\rho} \tag{3.96}$$

for some $A(\chi), B(\chi) \in \mathbb{C}$, where (here and in what follows) ρ goes over the zeros of $\Lambda(s,\chi)$, i.e., the non-trivial zeros of $L(s,\chi)$. Hence, just as in (2.40),

$$\frac{\Lambda'(s,\chi)}{\Lambda(s,\chi)} = B(\chi) + \sum_\rho\left(\frac{1}{s-\rho}+\frac{1}{\rho}\right). \tag{3.97}$$

By definition (3.82),

$$\frac{\Lambda'(s,\chi)}{\Lambda(s,\chi)} = \frac{L'(s,\chi)}{L(s,\chi)} + \frac{1}{2}\digamma\left(\frac{s+\kappa}{2}\right) + \frac{1}{2}\log\frac{q}{\pi}, \tag{3.98}$$

[5] Trudgian actually bounds $N(T,\chi) + N(T,\overline{\chi})$, in our notation. However, a look at [Tru15b, §2] suffices to show that the same argument gives a bound on $N(T,\chi)$ equal to half of the bound given there on $N(T,\chi)+N(T,\overline{\chi})$. We choose the constant values corresponding to $\eta = 0.20$ in [Tru15b, Table 1].

[6] Much of what follows is already in [Lan09a, §130]. The material on $B(\chi)$ is also classical, and can be largely found in [Dav67, §12] or [Ten15, §II.8.5]. Equation (3.103) was already familiar to Landau (see, e.g., [Lan18b, p. 216], where he states and uses it in the case $q=4$), as was equation (3.101), proved in greater generality in [Lan19, p. 155]. See also [Coh07, § 10.3] for more on the values of $L(s,\chi)$ and $L'(s,\chi)$ at $s=0$ and at $s=1$.

where $\kappa = [\chi(-1) = -1]$.

We see immediately from (3.98) that

$$B(\chi) = b(\chi) - \frac{\gamma}{2} - \kappa \log 2 + \frac{1}{2} \log \frac{q}{\pi}, \tag{3.99}$$

where $b(\chi)$ is the constant term in the Laurent expansion of $L'(s,\chi)/L(s,\chi)$ around $s = 0$. It is easy and classical to express $\Re B(\chi)$ in terms of the zeros of $\Lambda(s,\chi)$ as follows. By (3.97) and the functional equation (3.81),

$$B(\chi) = \frac{\Lambda'(0,\chi)}{\Lambda(0,\chi)} = -\frac{\Lambda'(1,\overline{\chi})}{\Lambda(1,\overline{\chi})}. \tag{3.100}$$

Applying (3.97) again, this time with $s = 1$, we see that

$$B(\chi) = -B(\overline{\chi}) - \sum_\rho \left(\frac{1}{\rho} + \frac{1}{1-\rho} \right),$$

since, as ρ goes over the zeros of $\Lambda(s,\chi)$, $1-\rho$ goes over the zeros of $\Lambda(s,\overline{\chi})$. As can be told easily from $\overline{\Lambda(s,\chi)} = \Lambda(\overline{s},\overline{\chi})$ and (3.96), $B(\overline{\chi}) = \overline{B(\chi)}$. Hence

$$2\Re B(\chi) = B(\chi) + B(\overline{\chi}) = -\sum_\rho \left(\frac{1}{\rho} + \frac{1}{1-\rho} \right).$$

Now, as ρ goes over the roots of $\Lambda(s,\chi)$, so does $\overline{1-\rho}$. Both $\Re(1/\rho)$ and $\Re(1/(1-\rho))$ are positive, and rearranging positive terms in a sum does not alter its total value. Thus,

$$\frac{1}{2}\sum_\rho \left(\Re\frac{1}{\rho} + \Re\frac{1}{1-\rho} \right) = \frac{1}{2}\sum_\rho \left(\Re\frac{1}{\rho} + \Re\frac{1}{\overline{\rho}} \right) = \sum_\rho \Re\frac{1}{\rho},$$

and so we obtain that

$$\Re B(\chi) = -\sum_\rho \Re\frac{1}{\rho}. \tag{3.101}$$

Thus, by (3.97), for any s,

$$\Re\frac{\Lambda'(s,\chi)}{\Lambda(s,\chi)} = \Re\sum_\rho \frac{1}{s-\rho} = \sum_\rho \frac{\Re s - \Re\rho}{|s-\rho|^2} \tag{3.102}$$

We can also express $B(\chi)$ and $b(\chi)$ as follows. By (3.100) and (3.98),

$$B(\chi) = -\frac{1}{2}\digamma\left(\frac{1+\kappa}{2}\right) - \frac{1}{2}\log\frac{q}{\pi} - \frac{L'(1,\overline{\chi})}{L(1,\overline{\chi})}. \tag{3.103}$$

By (3.49), $\digamma((1+\kappa)/2) = -\gamma - 2(1-\kappa)\log 2$, and so

$$B(\chi) = \frac{\gamma}{2} + (1-\kappa)\log 2 - \frac{1}{2}\log\frac{q}{\pi} - \frac{L'(1,\overline{\chi})}{L(1,\overline{\chi})}. \tag{3.104}$$

Therefore, by (3.99),

$$b(\chi) = \log \frac{2\pi}{q} + \gamma - \frac{L'(1,\overline{\chi})}{L(1,\overline{\chi})}. \tag{3.105}$$

We know from (3.48) that $F(s)$ has a pole $-1/s$ at $s = 0$ and no pole at $s = 1/2$. We conclude that the Laurent expansion of $L'(s,\chi)/L(s,\chi)$ at $s = 0$ is of the form

$$\frac{1-\kappa}{s} + b(\chi) + b_1 s + \dots . \tag{3.106}$$

Bounding $L'(1,\overline{\chi})/L(1,\overline{\chi})$ (and hence $b(\chi)$ and $B(\chi)$) in terms of q is not completely trivial, or rather there is an inherent difficulty due to the possibility of a Siegel zero. See §16.1.5.

Chapter Four

Computational matters

We have to talk about computations – and, in particular, ask ourselves how to carry them out with efficiency and rigor. One might think that the main challenge here is how to write on these matters for other mathematicians without being very dull. It turns out that this is not hard, since there are actual ideas to be discussed – it is not all just bookkeeping.

We will start with an explanation (§4.1) of how to carry out floating-point calculations (that is, computations in $\mathbb{R}$, as opposed to $\mathbb{Z}$ or $\mathbb{Q}$)[1] rigorously. The techniques we use are well-known to specialists (*interval arithmetic*, and its variant, ball arithmetic). We will see how to use it to find maxima and minima and to integrate numerically with rigor. We will also take a brief look at some special functions and how to estimate them rigorously (§4.2).

In §4.3, we will go over explicit results on L-functions, and, in particular, D. Platt's work, which was carried out using interval arithmetic. His verification of GRH up to a bounded height for characters of bounded conductor will be crucial for our work on the major arcs.

We will often want to compute quantities such as $\Lambda(n)$ or $\mu(n)$ for all $n \leq x$, where $x = 10^{12}$, say. Then we need a method that works in time no more than roughly linear on x and space considerably less than x. Segmented sieves (§4.5) are designed for this purpose: a segmented sieve of Eratosthenes works in time essentially linear on x and space proportional to $\sqrt{x}$. We will describe a few easy and well-known improvements. We will also briefly sketch how to do better than $\sqrt{x}$, though we will not need to.

We shall see in §4.6 how to check the ternary Goldbach conjecture for all odd n up to a constant C. Here the approach uses what is not quite pure brute force, but rather a mixture of brute force and known ideas [Sao98], [Pro78]. The computations here are in $\mathbb{Z}$, but we also mention an alternative approach relying on explicit results on the Riemann zeta function [RS03], [KL14].

Lastly, in §4.7, we will say a few words on automated theorem-proving. One technique of that kind was originally used for the proof of a minor lemma, but is no longer needed at all in the current version of the proof. However, automated theorem-proving is intriguing, and very much worth knowing about. The same goes, of course, for computer-aided formal proofs, a subject we will barely touch upon. For a more de-

[1] To be precise: floating-point computer arithmetic deals with numbers represented in the form $n \cdot b^m$, with $n, m \in \mathbb{Z}$ and $b = 2$ or $b = 10$, say, rather than as numbers having a fixed number of digits after the decimal point (fixed-point arithmetic). Floating-point arithmetic is used far more often than fixed-point, as it allows us to work with numbers of very different magnitudes.

tailed discussion, see, e.g., [CFDH+19].

4.1 INTERVAL ARITHMETIC

Interval arithmetic has, as its basic data type, intervals of the form $I = [a/2^\ell, b/2^\ell]$, where $a, b, \ell \in \mathbb{Z}$ and $a \leq b$. Say we have a real number x, and we want to know $\sin(x)$. In general, we cannot represent x in a computer, in part because it may have no finite description. The best we can do is to construct an interval of the form $I = [a/2^\ell, b/2^\ell]$ in which x is contained.

What we ask of a routine in an interval-arithmetic package is to construct an interval $I' = [a'/2^{\ell'}, b'/2^{\ell'}]$ in which $\sin(I)$ is contained. (In practice, the package computes the interval partly in software, by means of polynomial approximations to $\sin$ with precise error terms, and partly in hardware, by means of an efficient usage of rounding conventions.) What we obtain is, in effect, a value for $\sin(x)$ (namely, $(a' + b')/2^{\ell'+1}$) and a bound on the error term (namely, $(b' - a')/2^{\ell'+1}$). The alternative would be to compute all bounds on error terms ourselves; it is much nicer to automate the process, though, of course, the computer then has more to compute.

Interval arithmetic does not eliminate sources of imprecision in computation. It is just that, when a conventional routine might tell us a lie, a routine based on interval arithmetic will yield a less useful truth, that is, a large interval.

A common source of imprecision is cancellation. Say we know $x = 0.003006$ and $y = 0.003005$ to three digits of precision after the leading digit 3. Then we will know only the leading digit of $x - y = 0.000001$, if that. Similarly, in interval arithmetic, if x and y are given as intervals $[x_0, x_1]$, $[y_0, y_1]$, then $y - x$ will be an interval equal to or containing $[y_0 - x_1, y_1 - x_0]$. It is not just that the width δ of the new interval is equal to at least the sum of the widths of the first two intervals, but, above all, that, if $y_0 - x_1$ and $y_1 - x_0$ are much smaller than x_0, x_1, y_0 and y_1, then δ will be large compared to $y_0 - x_1$ and $y_1 - x_0$. This phenomenon is called *catastrophic cancellation*. One must simply devise ways to avoid it (or ask a numerical analyst).

There is a variant of interval arithmetic called *ball arithmetic*. It differs from interval arithmetic in that, instead of storing the end points x_0, x_1 of an interval $I = [x_0, x_1]$, it stores its midpoint $(x_0 + x_1)/2$ and its radius $(x_1 - x_0)/2$. Ball arithmetic can save both time and space, in that the radius, generally quite small, need not to be computed and stored to the same precision as the midpoint. See, e.g., [vdH09] for more details. Since the distinction between interval and ball arithmetic is mostly internal – the user may usually be oblivious to the difference – we will not discuss it further.

Implementation. There are several implementations of interval arithmetic and ball arithmetic available. They differ widely in their speed and in how much work they leave to the user. It might seem reasonable to expect a single package based on interval arithmetic to provide the following:

1. A good collection of functions. This should include not just the most common transcendental functions (such as $\log$, $\exp$ and $\sin$) but also a wider selection of *special*

functions.

2. Rigorous *numerical integration*.
3. Rigorous procedures for finding *roots* of an equation $f(x) = 0$ and *maxima and minima* of a function f on an interval.
4. Routines for *symbolic* or *automatic differentiation*. (See §4.1.2.)
5. Availability as a library of highly optimized routines to be used in compiled programs. Usage should be intuitive and, ideally, correspond to that used for non-interval arithmetic, i.e., $x + y$ should give the sum of x and y.
6. An interactive interface, or incorporation into a system providing such an interface.
7. A "native" mode, that is, the option of using a format or formats supported on hardware (e.g., IEEE double, or, in the future, IEEE quadruple), as well as whatever routines are correctly implemented on hardware.[2]
8. An arbitrary-precision mode – providing arbitrarily small intervals at the cost of being slower than native mode.

One might also add that the package should also be free and, in particular, open-source, for several reasons – not least of them that the code should be verifiable by all, and that non-proprietary code makes the duplication (or octuplication) of labor a little less inevitable.

No package seems to have had all of these capabilities during the time the present work was written (2012–2019). Rather, there is a number of libraries, each with some virtues and not others.

The situation does seem to be improving, particularly as more libraries are developed and incorporated into Sage.[3] Nevertheless, it is still often necessary to cobble things together or implement them oneself. Hence, we must understand how to carry out some basic tasks: quadrature, i.e., numerical integration, and finding roots and extrema. We shall also discuss briefly some special functions.

In the end, for intensive computations, and also for many smaller tasks, I used D. Platt's implementation [Pla11] of double-precision interval arithmetic based on Lambov's [Lam08] ideas. It is a small native-mode library working with IEEE double type on Intel x86/x64 processors, using [DLDDD$^+$10] to implement basic transcen-

[2] "Correctly implemented routines" usually means just the four basic operations and the square root. IEEE standards set out specifications that (say) chip manufacturers must follow if they claim to obey the standard; not everything is specified in the standard, and, on matters that are not specified in the standard, the behavior of different systems may differ from each other, or from common sense and decency. As it happens, the IEEE standard does not require functions other than the four basic operations, the square root and a remainder function to be rounded correctly, or even implemented, even though, in its revised version from 2008, it recommends that several other functions be implemented and rounded correctly [IEE08, §9.2]. The reader should now be psychologically prepared for the fact that some basic functions (such as sin) are implemented with incorrect rounding in the Intel x86/x64 processors that most computers use nowadays. In fact, the error is much larger than just incorrect rounding. Thus, all transcendental functions need to be implemented in software. See [DLDDD$^+$10] for a library that gives fast, correctly rounded versions of basic transcendental functions in IEEE double type.

[3] Sage, also known as SageMath, is a computer algebra system conceived as an open-source alternative to Magma, Maple, Mathematica and other proprietary systems. It works for the most part as an interface to many other open-source mathematics packages. Readers of the future: most contemporaries of the author knew what Sage was, but, surprisingly, some did not.

dental functions correctly. To compute some numerical integrals and expressions involving special functions, and to perform small-scale computations by means of Sage code embedded inside the TeX source of this book, I used the ARB package [Joh13], which implements ball arithmetic. Especially when used as a part of Sage, it seems to be the package that comes closest to fulfilling criteria (1)–(7) above. In earlier versions of the text, I used VNODE-LP [Ned06], which relies on the older PROFIL/BIAS interval arithmetic package [Knü99], to compute a few numerical integrals. Platt's work ([Pla16]; to be addressed in §4.3) used Platt's own routines, supplemented by the MPFI package [RR05] (based on MPFR [FHL$^+$07]) for arbitrary-precision interval arithmetic. ARB uses MPFR as well.

Of course, implementing matters on other interval arithmetic packages will also give correct results; readers are encouraged to write their own programs using whatever libraries they prefer. Speed and convenience are the main things that vary. In practice, if native-mode interval arithmetic slows down matters by a factor of less than 10 relative to ordinary floating-point arithmetic, we are doing well.

For further reference, [Tuc11] provides a brisk introduction with a useful bibliography. The documentation and papers accompanying contemporary implementations of interval arithmetic are often worth skimming.

4.1.1 Finding roots and extrema

The *bisection method* is a particularly simple method for finding maxima and minima of functions, as well as roots. It combines rather nicely with interval arithmetic, which makes the method rigorous. We follow an implementation based on [Tuc11, §5.1.1]. Let us go over the basic ideas.

Let us use the bisection method to find the minima (say) of a function f on a compact interval I_0. (If the interval is non-compact, we generally apply the bisection method to a compact sub-interval and use other tools, e.g., power-series expansions, in the complement.) The method proceeds by splitting an interval into two repeatedly, discarding the halves where the minimum cannot be found. More precisely, if we implement it by interval arithmetic, it proceeds as follows. First, in an optional initial step, we subdivide (if necessary) the interval I_0 into smaller intervals I_k to which the algorithm will actually be applied. For each k, interval arithmetic gives us a lower bound r_k^- and an upper bound r_k^+ on $\{f(x) : x \in I_k\}$. Let m_0 be the minimum of r_k^+ over all k. We can discard all the intervals I_k for which $r_k^- > m_0$. Then we apply the main procedure: starting with $i = 1$, split each surviving interval into two equal halves, recompute the lower and upper bound on each half, define m_i, as before, to be the minimum of all upper bounds, and discard, again, the intervals on which the lower bound is larger than m_i; increase i by 1. We repeat the main procedure as often as needed. In the end, we obtain that the minimum is no smaller than the minimum of the lower bounds (call them $(r^{(i)})_k^-$) on all surviving intervals $I_k^{(i)}$. Of course, we also obtain that the minimum (or minima, if there is more than one) must lie in one of the surviving intervals.

It is easy to see how the same method can be applied (with a trivial modification) to find maxima, or (with very slight changes) to find the roots of a real-valued function on a compact interval. In other words, the bisection method can be used in several ways to replace so-called “proofs by inspection,” such as “the maximum of a function f is less than 4 because I can see it on the screen” or “the plot shows that the function has a root around there”. Of course, the point is that such “proofs” are not actual proofs, whereas the bisection method combined with interval arithmetic is rigorous.

Trivial variant. It is possible to use a very simple version of bisection to check that a function $f : \mathbb{R} \to \mathbb{R}$ is always positive on an interval $[a, b]$, or that, given two functions f_1, f_2, the inequality $f_1(x) > f_2(x)$ holds for all x in $[a, b]$. The second question obviously reduces to the first one: set $f(x) = f_1(x) - f_2(x)$. As for the first question, use interval arithmetic to evaluate f on $[a, b]$. If the result tells us that $f([a, b])$ is contained in $(-\infty, 0]$, then the answer is certainly no; if it tells us that $f([a, b])$ is contained in $(0, \infty)$, the answer is certainly yes. If we are in neither case, we recurse, asking the same question of $[a, (a + b)/2]$ and $[(a + b)/2, b]$; we return the answer “yes” if and only if the result is “yes” on both intervals – we return “no” otherwise. See Algorithm 1.

Algorithm 1 Testing positivity by bisection

```
function BISECPOS($f$,$[a_0, a_1]$)
Ensure: returns whether $f(x) > 0$ for all $x \in [a_0, a_1]$
    $[y_0, y_1] \leftarrow f([a_0, a_1])$
    if $y_1 \leq 0$ then
        return False
    if $y_0 > 0$ then
        return True
    $h \leftarrow (a_0 + a_1)/2$
    return BISECPOS($f$,$[a_0, h]$) and BISECPOS($f$,$[h, a_1]$)
```

Of course, the problem with this simple procedure is that it does not terminate if $f(x)$ equals or gets very close to 0 anywhere on the interval $[a, b]$. Thus, this cheap sort of bisection – trivial to code – is simply something to be applied when a plot already makes it evident that f stays away from the x-axis, or that one of f_1, f_2 is clearly greater than the other throughout the interval.

More sophisticated methods for finding roots and extrema can also be adapted to use interval arithmetic. In particular, Newton’s method can be adapted, and in fact becomes better-behaved. See [Tuc11, §5.1.3–5.1.4].

$* * *$

What we saw above was a method for finding the minimum, or maximum, of a function f on an interval I in the sense of bounding the minimum, or maximum, of f within I. If we actually want a full list of all local maxima or minima x_i of a function f, together with small intervals in which the values of $f(x_i)$ is contained, the procedure is a little more complicated.

The simplest thing may then be to use the bisection method in two ways – to find maxima or minima and to find roots. Matters are particularly simple if (a) f is differentiable and (b) f and f' have no double roots in $I = [a, b]$. (It is not necessary to check condition (b) before starting; rather, if it fails, it may cause problems that will become obvious during the procedure we are about to describe.)

First, we use bisection (say) to find small intervals $[a_i, b_i] \subset I$ in which the roots $x_i \in I$ of $f'(x) = 0$ are contained. If these intervals are small enough, we will be able to use bisection to show that $f''(x)$ is bounded away from zero within each interval. (If this is not the case, then f'' may have a double root after all, and working with higher-order derivatives may be necessary.) Then we verify that $\text{sgn}(f'(a_i)) \neq \text{sgn}(f'(b_i))$. This shows that there is exactly one root of $f'(x) = 0$ within each interval $[a_i, b_i]$. We finish by evaluating f at $[a_i, b_i]$ by interval arithmetic; what we obtain is precisely a lower and an upper bound for f within $[a_i, b_i]$ by interval arithmetic. There are no local extrema elsewhere, except perhaps at the endpoints a, b of I. We check them and are done.

Of course, carrying out this procedure involves figuring out f' and f''. This matter will be our next topic.

4.1.2 Automatic and symbolic differentiation

Whether we are finding extrema or integrating a function numerically, we will often need to know the derivatives $f^{(k)}$, $k \geq 1$, of the function f we are working with. If f is simple enough, we can, of course, figure them out by hand, but that is cumbersome and error-prone in general. It is much better to use *symbolic* or *automatic differentiation*.

Symbolic differentiation is what we have come to expect from a computer algebra system, such as Macsyma and its many descendants: the system is given an expression – written out much as for human readers – and returns a expression for its derivative. The computer works matters out much like a human being who has taken first-year calculus, namely, using some knowledge on particular functions (such as $(x^n)' = nx^{n-1}$ and $\sin' = \cos$) and the chain rule.

A system for *automatic differentiation*, is given a procedure for computing a function and returns a procedure, or else calculates derivatives at the same time that a procedure is carried out. In principle, it need only know that $(f + g)' = f' + g'$, $(f - g)' = f' - g'$, the product rule $(fg)' = f'g + fg'$, the quotient rule and the chain rule. It relies on the fact that all a computer can do is to compose the four basic operations in some way. (Come to think of it, interval arithmetic is based on the same fact.) See, for instance, [Tuc11, Ch. 4].

Whether we use symbolic or automatic differentiation is a matter of convenience. Both VNODE-LP and ARB can do automatic differentiation.

4.1.3 Numerical integration

Let us be given a function $\phi : [a, b] \to \mathbb{C}$ to integrate from a to b. We can estimate the integral rigorously as follows:

1. partition the interval $[a, b]$ into N intervals $[x_j, x_{j+1}]$;
2. compute ϕ on $[x_j, x_{j+1}]$ using interval arithmetic, i.e., ask an interval-arithmetic

routine to find intervals $[y_j, y_{j+1}]$ containing $\phi([x_j, x_{j+1}])$;

3. sum the intervals $[y_j, y_{j+1}]$ using interval arithmetic. The result is an interval $[A, B]$ such that
$$A \leq \int_a^b \phi(x)dx \leq B.$$

This procedure (the rigorous form of "brute-force integration") may be acceptable if you find yourself in the mountains with little other than a compiler, a computer and an integrand ϕ. The error term $B - A$ in this procedure is inversely proportional to N.

We can do much better. For instance, we can use the Euler-Maclaurin formula (3.7) to estimate the integral $\int_a^b \phi(x)dx = (1/N)\int_{Na}^{Nb} \phi(x/N)dx$ in terms of the sum $\sum_{Na<n\leq Nb} \phi(n/N)$, or instead use, say, the midpoint rule for numerical integration:

$$\int_a^b \phi(x)dx = \frac{b-a}{N}\sum_{n=0}^{N-1} \phi\left(a + \frac{n+1/2}{N}(b-a)\right) + O^*\left(\frac{(b-a)^2}{24N^2}\int_a^b |\phi''(x)|dx\right) \tag{4.1}$$

for ϕ piecewise continuously differentiable and ϕ' of bounded variation, with ϕ'' understood in the sense of distributions, as usual.

Exercise 4.1. Prove (4.1) using either Euler-Maclaurin or your bare hands.

We can bound the integral on the right side of (4.1) by the naïve procedure we gave at first. In other words, (4.1) gives us that

$$\begin{aligned}\int_a^b \phi(x) = &\frac{b-a}{N}\sum_{n=0}^{N-1} \phi\left(a + \frac{n+1/2}{N}(b-a)\right)\\ &+ O^*\left(\frac{(b-a)^3}{24N^3}\sum_{n=0}^{N-1} \max_{t\in[x_n,x_{n+1}]} |\phi''(t)|\right),\end{aligned} \tag{4.2}$$

where $x_n = a + (b-a)n/N$, and we can compute the maximum of $\phi''(t)$ for $t \in [x_n, x_{n+1}]$ just applying interval arithmetic to compute $\phi''([x_n, x_{n+1}])$ for each n. We obtain an error term proportional to $(b-a)^3/24N^2$.

One can generally do even better by going to higher derivatives. There is a wide variety of procedures for numerical integration ("quadrature"); see [AS64, Ch. 25] or [OLBC10, §3.5], or also [Tuc11, §5.3].

It might seem that, in order to have access to ϕ'' (or higher derivatives), we would either need to figure out ϕ'' (by symbolic integration, say) and pass it to the integration routine, or to code everything so that the routine can figure out ϕ'' by automatic differentiation. There is actually a third option, implemented by ARB: if ϕ extends to a complex-analytic function, and this function is passed to the integration routine, then the routine can figure out all derivatives of ϕ on its own, simply by contour integration.

There is a challenge to this last approach when we are to integrate an expression of the form $|\phi(x)|$, ϕ the restriction to $[a, b]$ of a function holomorphic in a complex neighborhood of $[a, b]$. If ϕ is real-analytic, then we first find the zeros of $\phi(x)$ on $[a, b]$, thus reducing the problem to the case where $\phi(x)$ only vanishes at the endpoints, if anywhere; then the complex-analytic extension of $|\phi(x)|$ is just $\phi(z)$, or $-\phi(z)$, and we can proceed. See [Joh18]. However, if ϕ takes complex values on $[a, b]$, matters are more complicated: while $f(x) = \Re\phi(x)$ and $g(x) = \Im\phi(x)$ are real-analytic, and thus admit holomorphic extensions to a neighborhood R of $[a, b]$, the function $f(z) + ig(z)$ extending ϕ could vanish arbitrarily close to $[a, b]$, and so $\sqrt{f(z)^2 + g(z)^2}$ (which extends $\phi(x)$) could have a branch point arbitrarily close to $[a, b]$. Thus, it sometimes seems necessary to pass bounds on $\phi^{(k)}$ onto the integration routine, or, if all else fails, use brute-force integration.

* * *

We have been discussing integrals over compact intervals $[a, b]$ of functions without singularities in $[a, b]$. We will have to take some integrals over $(-\infty, \infty)$. We will start by restricting to a large interval $[-L, L]$. In order to bound the integral over $(-\infty, -L) \cup (L, \infty)$, we could change variables $x \mapsto 1/x$ and then deal with the possible singularity arising at 0 by means of Taylor or Puiseux series. Instead, we will simply use some crude bound on our integrand for $x \in (-\infty, -L) \cup (L, \infty)$.

4.2 SPECIAL FUNCTIONS

Special functions are something with a partly social and partly mathematical definition. A special function is

- non-elementary: it should not be expressible in terms of a finite number of applications of $+$, $-$, $\cdot$, $/$, nth roots, exponentiation, logarithms, $\arcsin$, and a few other basic functions, such as $\Re z$ or $|z|$;
- an object of special attention: that is, it is one of a few chosen functions that are studied in depth – one can then try to reduce other functions to them.

Special functions typically arise as integrals of elementary functions.

We are about to go over a few special functions that will appear several times in the text. We will, in particular, give summary explanations of how to compute some of them in certain ranges. We need methods of computation that give upper and lower bounds, and can be implemented in interval arithmetic. Since we aim at interval-arithmetic results that are merely correct and useful, rather than of maximal precision for a given data type, this will be straightforward.

The main subject of [AS64] and [OLBC10] is special functions. Even though both texts are mainly collections of formulae for immediate usage, no computer package seems to come close to using most of the knowledge contained there. The situation will no doubt improve over time.

Some interval-arithmetic libraries – notably MPFI [RR05] and the relatively recent package ARB [Joh13] – do implement most of the special functions we need. Our aim here is in part to discuss some details that will enable readers to carry out their own implementations, if they so wish.

4.2.1 The error function erf

The *error function* erf is defined by

$$\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-u^2} du. \tag{4.3}$$

Most interval-arithmetic packages do implement $\operatorname{erf}(z)$, at least for z real. The *imaginary error function* is defined by

$$\operatorname{erfi}(x) = -i \operatorname{erf}(ix) = \frac{2}{\sqrt{\pi}} \int_0^x e^{t^2} dt. \tag{4.4}$$

We also define the *Dawson function*

$$D_+(x) = e^{-x^2} \int_0^x e^{t^2} dt = \frac{\sqrt{\pi}}{2} e^{-x^2} \operatorname{erfi}(x). \tag{4.5}$$

The *complementary error function* erfc is simply

$$\operatorname{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-u^2} du = 1 - \operatorname{erf}(x), \tag{4.6}$$

since, of course, $\int_0^\infty e^{-x^2} dx = \sqrt{\pi}/2$.

For $x \geq 0$ real, we will often make do with the crude bounds

$$\operatorname{erf}(x) \leq \frac{2}{\sqrt{\pi}} \int_0^\infty e^{-(x+u)^2} du \leq \frac{2}{\sqrt{\pi}} \int_0^\infty e^{-x^2-u^2} du = e^{-x^2},$$

$$\operatorname{erf}(x) \leq \frac{2}{\sqrt{\pi}} \frac{1}{x} \int_x^\infty u e^{-u^2} du = \frac{1}{\sqrt{\pi}} \frac{e^{-x^2}}{x} \quad (\text{for } x > 0).$$

ARB gives a ball-arithmetic implementation of $\operatorname{erf}(z)$ for z complex, and thus, in particular, it can compute erfi. The implementation has been carefully optimized; see [Joh16]. See also [CR08], [Che12] on computing $\operatorname{erf}(x)$ to arbitrary precision for x real.

The following is a minimal discussion on how to implement $\operatorname{erfi}(x)$ on one's own for x bounded. The Taylor expansion of erf around $z = 0$ is

$$\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}} \sum_{n=0}^\infty \frac{(-1)^n z^{2n+1}}{n!(2n+1)} \tag{4.7}$$

(see, e.g., [AS64, 7.1.5], or expand e^{-x^2} into a series and take the integral). It gives us that, for any $N \geq 1$ such that $N + 1 > x^2$,

$$\begin{aligned} \operatorname{erfi}(x) = -i \operatorname{erf}(ix) &= \frac{2}{\sqrt{\pi}} \sum_{n=0}^\infty \frac{x^{2n+1}}{n!(2n+1)} \\ &= \frac{2}{\sqrt{\pi}} \left(\sum_{n=0}^{N-1} \frac{x^{2n+1}}{n!(2n+1)} + O^*\left(\frac{x^{2N+1}/(N!(2N+1))}{1 - \frac{x^2}{N+1}} \right) \right). \end{aligned} \tag{4.8}$$

In general, using a continued fraction expansion of $\mathrm{erfi}(x)$ or $D_+(x)$ is a better idea than using (4.8); see [McC74] and references therein. Incidentally, D_+ has the advantage of being bounded: by [Inc19, A133842],

$$|D_+(x)| \leq 0.541045. \tag{4.9}$$

Since we will be computing $\mathrm{erfi}(x)$ only for x bounded, it is enough to use the Taylor expansion (4.8). Even in that case, a Chebyshev approximation as in [CKT70] can be a good alternative.

4.2.2 Incomplete gamma functions

We defined the Gamma function $\Gamma(s)$ in (3.34). The *lower incomplete Gamma function* $\gamma(s, x)$ is defined by

$$\gamma(s, x) = \int_0^x e^{-t} t^{s-1} dt. \tag{4.10}$$

The *upper incomplete Gamma function* $\Gamma(s, x)$ is given by

$$\Gamma(s, x) = \Gamma(s) - \gamma(s, x) = \int_x^\infty e^{-t} t^{s-1} dt. \tag{4.11}$$

Incomplete gamma functions are implemented through ball arithmetic in ARB and (for x real) through interval arithmetic in MPFR. Much on them can be found in [AS64, §6.5] and [OLBC10, Ch. 8]. The following series expansion is easy and useful, particularly for s large and x fixed. By integration by parts, for $\Re s > 0$,

$$\gamma(s, x) = \int_0^x e^{-t} t^{s-1} dt = e^{-x} \frac{x^s}{s} + \frac{1}{s} \int_0^x e^{-t} t^s dt.$$

Iterating, we obtain

$$\gamma(s, x) = e^{-x} \sum_{k=0}^{\infty} \frac{x^{s+k}}{s(s+1)\dots(s+k)}. \tag{4.12}$$

By analytic continuation, (4.12) is actually true for any $s \neq 0, -1, -2, \dots$.

4.2.3 Parabolic cylinder functions

A parabolic cylinder function $f : \mathbb{C} \to \mathbb{C}$ is a solution to an equation

$$f'' + (a_2 z^2 + a_1 z + a_0) f = 0.$$

We can of course complete the square and reach the standard form

$$f'' - \left(\frac{1}{4} z^2 + a\right) f = 0. \tag{4.13}$$

It is clear that, even locally, the space of analytic solutions to (4.13) is at most two-dimensional. In fact, for every fixed $a \in \mathbb{C}$, the space of entire functions satisfying equation (4.13) is two-dimensional. The space of solutions going to 0 for real $z \mapsto +\infty$ is one-dimensional. The parabolic cylinder function $U(a, z)$ (to be defined in §15.3.1) spans that space. We follow the notation in [AS64], rather than that of Whittaker [Whi03], who denotes $U(a, x)$ by $D_{-a-1/2}(x)$.

We will actually be much less interested in computing $U(a, z)$ than in obtaining useful bounds on $U(a, z)$, particularly for z imaginary. Obtaining those bounds will be our main task in Chapter 15.

While the task of computing $U(a, z)$ to high precision will not be our concern, it turns out to be related to what we will do, in that the same basic approach (the *saddle-point method*) can be followed. The point is that the approach gives an expression for $U(a, z)$ where cancellation happens only in clearly lower-order terms, if at all. There are two consequences, the first one relevant to us, the second one important for computations: (a) the main term comes out cleanly; (b) the loss of numerical precision that comes with cancellation becomes much less harmful. See [GST06].

4.3 EXPLICIT METHODS FOR L FUNCTIONS

Let us now speak of explicit results on L-functions in the direction of the Generalized Riemann Hypothesis. There are different kinds of such results; besides zero-free regions ([McC84b], [Kad05]), already mentioned in §1.3.1, there are, for instance, zero-density estimates ([Kad13], [KN12]), which bound the proportion of zeros of an L-function in a given strip to the right of the critical line. We will use neither zero-free regions for L-functions nor zero-density estimates, however.

As we already said in §1.3.1, we will use a computation due to D. Platt [Pla16] showing that GRH holds for all characters of conductor up to a constant and all zeros up to a given height.

Proposition 4.2 (Platt [Pla16])**.** *For all primitive Dirichlet characters χ of modulus $q \leq 400000$, every non-trivial zero of the L-function $L(s, \chi)$ with $|\Im s| \leq H_q$ satisfies $\Re s = 1/2$, where*

$$H_q = \begin{cases} \max(10^8/q, 200 + 3.75 \cdot 10^7/q) & \text{for } q \text{ odd,} \\ \max(10^8/q, 200 + 7.5 \cdot 10^7/q) & \text{for } q \text{ even.} \end{cases} \tag{4.14}$$

What Platt did is a true verification, i.e., a proof, and not just a numerical experiment. How does one carry out such a verification?

The first computation of zeros of the Riemann zeta function was done by Riemann himself; one of the main tools still used for such computations, the Riemann-Siegel formula, was already known to him. A general method started to take shape in the hands of Gram [Gra03], Backlund [Bac14], Hutchinson [Hut25] and Titchmarsh [Tit35], among others. Turing carried out the first fully mechanical verification on an early electronic computer [Tur53]. Working on a computer forced him to improve the procedure, making it into a genuine algorithm that required no human intervention. D. H. Lehmer [Leh56] amended and extended Turing's work. What was by far the largest computation for L-functions before Platt was due to Rumely [Rum93]. In the case of $|\Im s|$ large, Platt's work uses ideas from Booker [Boo06a] related to work by Odlyzko and Schönhage [OS88]. In summary – there is a procedure, or family of procedures, that has evolved over time; what follows is simply a brief overview, based in part on [Pla16] and [Boo06b].

4.3.1 Finding zeros

4.3.1.1 Sign changes

Let $\Lambda(s,\chi)$ be a completed L function as in (3.82); let $\varepsilon(\chi)$ be as in (3.83). By (3.81), for $t \in \mathbb{R}$,

$$\Lambda\left(\frac{1}{2}+it,\chi\right) = \varepsilon(\chi)\Lambda\left(\frac{1}{2}-it,\overline{\chi}\right) = \varepsilon(\chi)\overline{\Lambda\left(\frac{1}{2}+it,\chi\right)}.$$

It is then evident that $|\varepsilon(\chi)| = 1$. We choose a ε'_χ such that $(\varepsilon'_\chi)^2 = 1/\varepsilon(\chi)$, and obtain that

$$\varepsilon'_\chi\Lambda\left(\frac{1}{2}+it,\chi\right) = \overline{\varepsilon'_\chi\Lambda\left(\frac{1}{2}+it,\chi\right)},$$

i.e., the function $Z_\chi(t) = \varepsilon'_\chi\Lambda(1/2+it,\chi)$ is real-valued.

Since $Z_\chi(t)$ is continuous and real-valued, every sign change indicates a zero: if $t_- < t_+$, and either $Z_\chi(t_-) < 0 < Z_\chi(t_+)$ or $Z_{t_+} < 0 < Z_\chi(t_-)$, then, evidently, there is a $t \in [t_-,t_+]$ such that $Z_t = 0$. Let $N(T,\chi)$ be the number of non-trivial zeros of $L(s,\chi)$ with $0 \le \Im s \le T$. Then, if we manage to find $N(T,\chi)$ sign changes of $Z_\chi(t)$ between 0 and T, we will be able to conclude that all non-trivial zeros of $L(s,\chi)$ with $0 \le \Im s \le T$ do fall on the critical line $\Re s = 1/2$, as we wished to show. In order to show the same for all non-trivial zeros with $-T \le \Im s \le 0$, we just need to recall that the zeros of $L(s,\chi)$ are the same as the zeros of $L(1-s,\overline{\chi})$, and so it is enough to show that all non-trivial zeros of $L(s,\overline{\chi})$ with $0 \le \Im s \le T$ fall on the critical line.

There are two tasks ahead of us: (a) finding a way to compute $N(T,\chi)$ exactly, (b) computing $L(s,\chi)$ (and thus $Z_\chi(t)$) rapidly, for many $s = 1/2+it$ and many χ. One could add the additional concern that there might be zeros of multiplicity > 1 (i.e., points s_0 around which $L(s,\chi)$ behaves like $(s-s_0)^k$, $k > 1$) on the critical line; those would be a problem, since there would be no sign change associated to them. We would get fewer than $N(T,\chi)$ sign changes, and would look for the missing sign changes forever. However, this situation never occurs in practice, as is completely unsurprising, given that L-functions $L(s,\chi)$ are conjectured not to have zeros of multiplicity > 1.

4.3.1.2 Counting zeros

Task (a) might seem daunting at first, but, in fact, the classical results in §3.5.3.3 and §3.7.3.3 take us very close to it. We have an expression for $N(T,\chi)$ with an error term of $O(\log qT)$. To compute $N(T,\chi)$ exactly, it would be enough to have an expression with an error term less than $1/2$, since we could eliminate it using the fact that $N(T, \chi)$ is an integer.

Turing proved an explicit version [Tur53, Thm. 4] of an estimate of Littlewood's [Lit24, Thm. 40], showing that $S(T)$ oscillates around 0, in a strong sense:

$$\left|\int_{T_1}^{T_2} S(t)dt\right| \leq 2.3 + 0.128 \log \frac{T_2}{2\pi} \tag{4.15}$$

for $T_2 > T_1 > 0$. Rumely [Rum93] proved a more general inequality with $S(T, \chi)$ instead of $S(T)$ for an arbitrary Dirichlet character χ. The numerical constants have been improved by Lehman [Leh70] (who also fixed some errors in Turing's paper) and by Trudgian [Tru11].

Most importantly, Turing pointed out how to use (4.15): if we miss zeros when considering sign changes, there will be an undercount that will show not only for a given $T = T_1$, but for all larger T; it will be enough to go up to $T_2 = T_1 + C \log T_1$, C a constant, for this error to accumulate to the point of giving a contradiction with (4.15). If we are in that situation, we go back and look more closely to find the missing sign changes; otherwise, we know that we have indeed verified that all non-trivial zeros with imaginary part up to T_1 lie on the critical line.

4.3.2 Computing $L(s, \chi)$

Let us look at task (b) in the above, viz., computing $L(s, \chi)$ rapidly. The computation should be done differently for $\Im s$ small and for $\Im s$ large. Part of the reason is that, for our application and many others, when $\Im s$ small, we need to know $L(s, \chi)$ for many characters χ, whereas, for $\Im s$ large, we need to know $L(s, \chi)$ only for characters χ of rather small modulus q. This means, in particular, that, when $\Im s$ is small, we may try to organize matters so that some of the computations are useful for many different characters χ. Proceeding in this manner amortizes the cost of these computations, so to speak.

Case 1: computing $L(s, \chi)$ for $|\Im s|$ moderate. We write $L(s, \chi)$ in terms of the Hurwitz zeta function $\zeta(s, \alpha)$, as in (3.86). For any given t, we follow the procedure in §3.7.2 to compute $\zeta(1/2 + it + r, d/D)$ for $0 \leq d < D$, $0 \leq r \leq r_0$, where D and r_0 are fairly large. (Here [Pla16] uses $r_0 = 15$, $D = 2048$.) For $r \geq 0$,

$$\frac{d^r}{d\alpha^r}\zeta(s, \alpha) = (-1)^r s(s + 1) \dots (s + r - 1)\zeta(s + r, \alpha),$$

and so we obtain $\zeta^{(r)}(1/2 + it, d/D)$. Given these values, we are able to compute $\zeta(1/2 + it, a/q)$ very quickly and to high accuracy for a, q arbitrary, simply by a Taylor expansion. (This procedure is standard: see [Coh07, §9.6.1].)

Once we know the values $\zeta(1/2 + it, a/q)$ for $1 \leq a \leq q$, computing $L(s, \chi)$ by means of (3.86) does not take long. The reason is that we can consider the expression $\sum_a \chi(a)\zeta(s, a/q)$ in (3.86) as the Fourier transform of the function $a \mapsto \zeta(s, a/q)$ on the finite abelian group $(\mathbb{Z}/q\mathbb{Z})^*$, evaluated at the point χ of the dual of $(\mathbb{Z}/q\mathbb{Z})^*$. We can use what is known as a fast Fourier transform (FFT) – that is, a fast algorithm for computing a discrete Fourier transform, proceeding by recursion – to tabulate all values of the Fourier transform of a function on a finite abelian group with m elements in time $O(m\log m)$. Here $m = \phi(q) < q$. The computation of $\zeta(1/2 + it + k, d/D)$ for given (t, k, d) is valid for all q and all χ, and thus it is as if its cost per character χ were divided by total number of characters χ of all moduli we consider. This reduction in the cost per character is what we meant by "amortizing".

Case 2: computing $L(s,\chi)$ for $|\Im s|$ large. For $t = \Im s$ large, we need a different algorithm: a computation time equal to a constant times $|t|$ is too large. There are different ways to proceed, all of them in some sense variants of the *approximate functional equation*. We start with a decomposition

$$L(s,\chi) = \sum_{n\leq N} \chi(n)n^{-s} + \sum_{n>N} \eta\left(\frac{n}{N}\right)\chi(n)n^{-s}. \tag{4.16}$$

It turns out that we can express the second sum in (4.16) as an integral involving $L(s,\chi)$, and then use the functional equation of $L(s,\chi)$ to obtain a functional equation for that sum, in the sense of expressing it in terms of $\sum_{m\leq M}\chi(m)m^{1-s}$ for M such that $NM = qt/2\pi$. In this way, we obtain that, for χ primitive and $N, M \geq \sqrt{q}$ with $NM = qt/2\pi$,

$$L(s,\chi) = \sum_{n\leq N} \chi(n)n^{-s} + A(s,\chi)\sum_{m\leq M} \overline{\chi(m)}m^{-(1-s)} + R, \tag{4.17}$$

where $A(s,\chi)$ is the factor appearing in the functional equation (3.81) when written in the form $L(s,\chi) = A(s,\chi)L(1-s,\overline{\chi})$ (that is, $A(s,\chi) = \varepsilon(\chi)q^{1/2-s}\frac{\gamma(1-s)}{\gamma(s)}$, where $\gamma(s) = \pi^{-(s+\kappa)/2}\Gamma((s+\kappa)/2)$ and $\varepsilon(\chi)$, κ are as in (3.83)) and R is a remainder term. It is clear that it is best to set $N, M \sim \sqrt{qt/2\pi}$, and thus the time required to compute a value of $L(s,\chi)$ is $O(\sqrt{qt})$.

The *Riemann-Siegel formula* [Sie32] gives an explicit expression for the remainder term R. (See also [Hia16].) Yet another possibility is to proceed as in [IK04, Thm. 5.3], replacing the sharp truncations in (4.17) by two transforms defined in terms of a holomorphic function $G(u)$, with the effect of removing the remainder term altogether for $q > 1$.

Since the density of zeros increases as $|t|$ becomes larger, it is good to have a method that amortizes the cost of computations for many neighboring t, much as before we could amortize computations for many different q and χ. Such a method was described in [OS88]; the average time spent per value of t is then $O(|t|^\epsilon)$, as opposed to $O(\sqrt{|t|})$. A similar algorithm in [Boo06a, §5] served as the basis for the fully explicit work in [Pla16]. The main tool in this amortization is, again, a fast Fourier transform: as explained at the end of [Boo06a, §5.1], one can express a certain crucial quantity as a convolution over the integers, and, as we know, a convolution $f * g$ can be expressed as an inverse Fourier transform of a product $\widehat{f}\cdot\widehat{g}$ of Fourier transforms $\widehat{f}, \widehat{g}$.

4.4 CALCULATING EULER PRODUCTS

An *Euler product* is an expression of the form $\prod_p f(p^{-s_1},\dots,p^{-s_n})$, where f is a rational function on $n \geq 1$ variables and the product $\prod_p$ is taken over all primes. We would like to compute the value of such a product for given values of $s_1,\dots,s_n$ for which the product converges.

Simplest approach. Let us study an example we will need later. We would like to compute

$$F(\sigma) = \prod_p f(p^{-1},p^{-\sigma}) \tag{4.18}$$

for different values of $\sigma > 0$, where

$$f(x,y) = \left(1 + \frac{xy}{(1+x)(1-y)}\right).$$

The case $\sigma = 1$ is easy: $F(1) = \zeta(2)$. Let $0 < \sigma < 1$.

If σ is not too small (say, if it is not much smaller than 1), then a naïve computational approach works: we can simply write

$$F(\sigma) = \prod_{p\leq N} f(p^{-1},p^{-\sigma}) \cdot \prod_{p>N} f(p^{-1},p^{-\sigma})$$

for some large N, and bound the tail $\prod_{p>N} f(p^{-1},p^{-\sigma})$ as follows. Since, for $x > 2$ and $\sigma > 0$, by the convexity of $t \to 1/t^{\sigma+1}$,

$$\sum_{p\geq x} \frac{1}{p^{\sigma+1}} < \sum_{\substack{n\geq x\\ n \text{ odd}}} \frac{1}{n^{\sigma+1}} < \frac{1}{2}\int_{x-1}^{\infty} \frac{1}{t^{\sigma+1}}dt = \frac{(x-1)^{-\sigma}}{2\sigma}, \tag{4.19}$$

and since $x \mapsto 1 - x^{-\sigma}$ is increasing on x, we see that

$$\begin{aligned}\log \prod_{p\geq x} f(p^{-1},p^{-\sigma}) &\leq \sum_{p\geq x} \frac{p^{-(\sigma+1)}}{(1+p^{-1})(1-p^{-\sigma})}\\ &\leq \frac{1}{1-x^{-\sigma}} \sum_{p\geq x} \frac{1}{p^{\sigma+1}} \leq \frac{(x-1)^{-\sigma}/2\sigma}{1-x^{-\sigma}}.\end{aligned} \tag{4.20}$$

for $x > 1$. Thus, for $\sigma \in (0,1)$ and $N > 1$ an integer,

$$\prod_{p\leq N} f(p^{-1},p^{-\sigma}) \leq F(\sigma) \leq \prod_{p\leq N} f(p^{-1},p^{-\sigma}) \cdot \exp\left(\frac{N^{-\sigma}/2\sigma}{1-N^{-\sigma}}\right).$$

As we said, these inequalities are enough to estimate $F(\sigma)$ efficiently for σ close to 1. The issue with smaller σ is that the product $\prod_{p\leq N} f(p^{-1},p^{-\sigma})$ then converges very slowly, and so a very large N would be needed.

4.4.1 Accelerating convergence by products of $\zeta(s)$

The solution will be to do express $F(\sigma)$ in terms of an Euler product converging much more rapidly than the one defining $F(\sigma)$. We will be able to do so by first dividing (or multiplying) $F(\sigma)$ by values of $\zeta(s)$ at some well-chosen points s. This is a well-known technique, probably rediscovered several times; a systematic treatment of the one-variable case can be found in [Mor00], but the idea is much older, going back at least to Littlewood (see [Wes22]). If the aim is to obtain analytic continuations, rather than to compute, then there is a method using the same idea, and it is also old: it goes back at least to [Est28] and [Dah52]. There is also a close relation with methods for computing sums of analytic functions over prime reciprocals $1/p$, as in [Coh].

For our particular choice of $f(x,y)$, the observation to make is that

$$f(x,y)\prod_{j=1}^{k}(1-xy^j) = 1+\frac{P_k(x,y)}{(1+x)(1-y)}, \tag{4.21}$$

where

$$P_k(x,y) = (1+x-y)\prod_{j=1}^{k}(1-xy^j) - (1+x)(1-y) \tag{4.22}$$

can be easily proved to be of the form $xy^{k+1}+x^2yQ_k(x,y)$, Q_k a polynomial. (Use induction on k.) Hence

$$F(\sigma) = \prod_{j=1}^{k}\zeta(1+j\sigma)\prod_p\left(1+\frac{P_k\left(p^{-1},p^{-\sigma}\right)}{\left(1+p^{-1}\right)\left(1-p^{-\sigma}\right)}\right). \tag{4.23}$$

The advantage is that the product over p here converges rapidly for k large enough – much more rapidly than the product in (4.18).

It only remains to bound the tail

$$\prod_{p>N}\left(1+\frac{P_k\left(p^{-1},p^{-\sigma}\right)}{\left(1+1/p\right)\left(1-p^{-\sigma}\right)}\right).$$

To do so, it is enough to bound $P_k(p^{-1},p^{-\sigma})/p^{\rho}$ for $p>N$, where $\rho=\min(2+\sigma,1+(k+1)\sigma)$; we then bound

$$\sum_{p>N}\frac{p^{-\rho}}{(1+p^{-1})(1-p^{-\sigma})}$$

as in (4.20).

Giving a useful bound on $P_k(t,t^{\sigma})/t^{\rho}$ is simple. Let P_k^+ consists of the terms in P_k with positive coefficients. Since $P_k(x,y)\equiv xy^{k+1} \bmod x^2y$, every term in P_k^+ is either divisible by x^2y or divisible by xy^{k+1}. For $x=t_0$, $y=t_0^{\sigma}$, we obtain that every term in $P_k^+(t,t^{\sigma})/t^{\rho}$ equals a power t^{α}, $\alpha\geq 0$, times a positive coefficient. Hence

$P_k(t, t^\sigma)/t^\rho \leq P_k^+(t_0, t_0^\sigma)/t_0^\rho$. In turn, $P_k^+(t_0, t_0^\sigma)/t_0^\rho$ is bounded by the sum C_k^+ of the coefficients of all monomials in P_k^+. We conclude that

$$\sum_{p>N} \frac{P_k\left(p^{-1}, p^{-\sigma}\right)}{(1+1/p)\left(1-p^{-\sigma}\right)} \leq C_k^+ \sum_{p>N} \frac{p^{-\rho}}{(1+1/p)\left(1-p^{-\sigma}\right)} \leq C_k^+ \frac{N^{-\rho}/2\rho}{1-N^{-\rho}}$$

for $N > 1$, by (4.19), and so

$$\prod_{p>N} \left(1 + \frac{P_k\left(p^{-1}, p^{-\sigma}\right)}{(1+1/p)\left(1-p^{-\sigma}\right)}\right) \leq \exp\left(C_k^+ \frac{N^{-\rho}/2\rho}{1-N^{-\rho}}\right).$$

In the same way,

$$\sum_{p>N} \frac{P_k\left(p^{-1}, p^{-\sigma}\right)}{(1+1/p)\left(1-p^{-\sigma}\right)} \geq -C_k^- \frac{N^{-\rho}/2\rho}{1-N^{-\rho}}$$

for C_k^- the sum of the absolute values of all negative coefficients of monomials in P_k. Since, as can be easily proved, $\prod_{1\leq i\leq n}(1-\epsilon_i) \geq 1 - \sum_i \epsilon_i$ for $\epsilon_i \in [0, 1]$, it follows that

$$\prod_{p>N} \left(1 + \frac{P_k\left(p^{-1}, p^{-\sigma}\right)}{(1+1/p)\left(1-p^{-\sigma}\right)}\right) \geq 1 - C_k^- \frac{N^{-\rho}/2\rho}{1-N^{-\rho}}.$$

Thus we have an upper and a lower bound on the tail.

It is clear that the procedure above can be made to work in general, for f a rational function on n variables and $s_1, \ldots, s_n$ such that $\prod_p f(p^{-s_1}, \ldots, p^{-s_n})$ converges. We will later see further examples in two and three variables. As for the values of zeta by which we multiply our product, we can choose them iteratively: if after the jth step in the iteration the term with the smallest negative exponent is $a_r p^{-e_r}$, we factor out $\zeta(e_r)^{a_r}$ on the $(j+1)$th step. (In (4.23), we chose all factors $\zeta(1+j\sigma)$ at the same time, and, for the sake of simplicity, our choice was not necessarily optimal.) It should also be clear that the procedure – including both the choices of e_r and a_r and the tail bounds – can be automated, thanks to symbolic algebra.[4]

See [Mor00] and [MRS] for an alternative treatment. In particular, [Mor00] chooses the factors $\zeta(e_r)^{a_r}$ at the beginning, on the basis of an expression for $F(\sigma)$ as an infinite product of values of ζ. See also [Mor05] for a connection with classical algebraic work, viz., the necklace identity and the Witt transform.

4.5 THE SIEVE OF ERATOSTHENES AND ITS VARIANTS

While we will generally use the word "sieve" in the meaning or meanings it has in analytic number theory – small sieve, large sieve – we should also discuss the kind of

[4] A routine along these lines is included in PARI. Unfortunately, as of 2019, it does not seem to do tail bounds, and thus its results are not guaranteed to be correct. Doing one's implementation instead is not hard.

sieves with which most people are familiar from their basic schooling: algorithms that go through a long list of consecutive integers to find the primes among them.

We will restrict our discussion to the best-known sieve, namely, the sieve of Eratosthenes. It has the advantage, crucial to us, of being applicable not just to finding primes, but to factoring integers, or to computing arithmetical functions ($\mu(n)$, $\phi(n)$, etc.) that depend on factorization.

The reader is presumably familiar with a simple sieve of Eratosthenes as in Algorithm 2. (Indeed, memories of the algorithm should be enough to make the conventions of the pseudocode[5] self-explanatory.) The code, as it stands, already contains two well-known and rather common-sensical improvements over the simplest versions of the sieve. First, to find the primes $\leq N$, it crosses out multiples $\leq N$ of integers d, where d goes up to only $\sqrt{N}$, not N. The reason why it is enough for d to go up to $\sqrt{N}$ is that, if a number $\leq N$ is composite, then it has a divisor in $(1, \sqrt{N}]$. The other improvement present in Algorithm 2 is that, once the even numbers > 2 are sieved out, we sieve out only odd multiples of the primes up to $\sqrt{N}$, rather than all multiples of all $1 < d \leq \sqrt{N}$. One can go further, eliminating multiples of 3, 5 and 7 (say) at the beginning as well, and then sieving out only numbers of the form mp, where $7 < p \leq \sqrt{N}$, $p \leq m \leq \sqrt{N}$, where $m \bmod 210 \in (\mathbb{Z}/210\mathbb{Z})^*$. There are many other tricks for cutting running time, though the version in Algorithm 2 already runs in essentially linear time.

The main limitation is often not time but memory. Or, to be more precise: (a) an algorithm that takes up more memory than is available will not run at all, (b) a program that uses much memory will often be slowed down significantly in practice. A computer has different kinds of memory, working at different speeds; the fastest and smallest one – much smaller than total memory – is called the cache. To simplify matters: a typical computer processor may be able to store and access 10^6 integers in cache, 10^9 in RAM, and 10^{12} in secondary storage, such as a hard drive. A program will generally run faster, possibly much faster, if it is able to work mainly with cache. We will not get into the details of computer architecture; at our level of abstraction, our task is to attempt to reduce both the time and space our algorithms use. It is simply that we should be aware that large space savings at a small cost in time in theory often lead to savings in time in practice.

4.5.1 A segmented sieve of Eratosthenes

It has been known since at least the 1960s [Sin69] that one can sieve the integers up to N using no more than $O(\sqrt{N})$ units of memory, while keeping time essentially linear on N. In fact, for any $\Delta \gg \sqrt{N}$, we can find the primes in an interval of the form $[n, n + \Delta] \subset N$ in time and space essentially linear[6] on Δ. The procedure is called

[5] *Pseudocode* denotes an explanation of an algorithm that is somewhere in between a verbal explanation and an actual computer program. It is meant to be relatively readable, while also making it relatively simple for someone who can program to turn it into actual computer code.

[6] When we say "essentially", we omit factors of log or log log. We give more precise estimates on space and time consumption at the end of our pseudocode (Algorithms 2–4). There and elsewhere, we follow the

a *segmented sieve* (Algorithm 3). If – as is the case typically for us – we do not need to store all primes $p \leq N$, but rather we just want to obtain them one after the other so as to compute a sum $\sum_{p\leq N} f(p)$, then we just have to split the interval $[1, N]$ into intervals of length about $\sqrt{N}$, and apply the segmented sieve. The memory used to sieve one interval is of course reused for the next one.

There are many carefully coded implementations of the segmented sieve of Eratosthenes, amounting to a small literature. See the references in [Wal].

Algorithm 2 A simple sieve of Eratosthenes

```
function SimpleSiev($N$)
Ensure: for $1 \leq n \leq N$, $P_n = 1$ if $n$ is prime, $P_n = 0$ otherwise
    $P_1 \leftarrow 0$, $P_2 \leftarrow 1$, $P_n \leftarrow 0$ for $n \geq 2$ even, $P_n \leftarrow 1$ for $n \geq 3$ odd
    $d \leftarrow 3$
    while $d \leq \sqrt{N}$ do
        if $P_d = 1$ then
            $n \leftarrow d \cdot d$
            while $n \leq N$ do
                $P_n \leftarrow 0$, $n \leftarrow n + 2d$          ▷ sieves out odd multiples $\geq d^2$ of $d$
        $d \leftarrow d + 2$
    return $P$
  Time: $O(N \log\log N)$.  Space: $O(N)$.
```

Algorithm 3 A segmented sieve of Eratosthenes

```
function SimpleSegSiev($n_0$,$\Delta$)                    ▷ finds primes in $[n_0, n_0 + \Delta]$
Ensure: for $0 \leq j \leq \Delta$, $S_j = 1$ if $n_0 + j$ prime, $S_j = 0$ otherwise
    $S_j \leftarrow 1$ for all $0 \leq j \leq \Delta$
    $S_j \leftarrow 0$ for $0 \leq j \leq 1 - n_0$                     ▷ 0 and 1 are not prime
    $P \leftarrow$ SimpleSiev($\lfloor\sqrt{n_0 + \Delta}\rfloor$)
    for $p \leq \sqrt{n_0 + \Delta}$ do
        if $P_p = 1$ then                                   ▷ if $p$ is a prime...
            $n \leftarrow \max(p \cdot \lceil n_0/p \rceil, 2p)$      ▷ smallest multiple $\geq \max(n_0, 2p)$ of $p$
            while $n \leq n_0 + \Delta$ do       ▷ $n$ goes over multiples of $p$ in $n_0 + [0, \Delta]$
                $S_{n-n_0} \leftarrow 0$, $n \leftarrow n + p$
    return $S$
  Time: $O((\sqrt{n_0} + \Delta) \log\log(n_0 + \Delta))$.  Space: $O(\sqrt{n_0} + \Delta)$.
```

convention that integers take constant space to store (as is often the case in practice). If we truly need to work with integers of unbounded size, then the space taken to store an integer n is proportional to $\log n$, and so space estimates need to be multiplied by a factor of log.

Algorithm 4 A segmented sieve of Eratosthenes for computing $\mu(n)$

```
function SimpleSegSievMu($n_0$,$\Delta$)        ▷ computes $\mu(n)$ for $n$ in $[n_0, n_0 + \Delta]$
Ensure: for $0 \leq j \leq \Delta$, $m_j = \mu(n_0 + j)$
    $m_j \leftarrow 1, \Pi_j \leftarrow 1$ for all $0 \leq j \leq \Delta$
    $P \leftarrow$ SimpleSiev($\lfloor\sqrt{n_0 + \Delta}\rfloor$)
    for $p \leq \sqrt{n_0 + \Delta}$ do
        if $P_p = 1$ then                            ▷ if $p$ is a prime...
            $n \leftarrow p \cdot \lceil n_0/p \rceil$                  ▷ smallest multiple $\geq n_0$ of $p$
            while $n \leq n_0 + \Delta$ do      ▷ $n$ goes over multiples of $p$ in $n_0 + [0, \Delta]$
                $m_{n-n_0} \leftarrow -m_{n-n_0}, \Pi_{n-n_0} = p \cdot \Pi_{n-n_0}, n \leftarrow n + p$
            $n \leftarrow p^2 \cdot \lceil n_0/p^2 \rceil$              ▷ smallest multiple $\geq n_0$ of $p^2$
            while $n \leq n_0 + \Delta$ do     ▷ $n$ goes over multiples of $p^2$ in $n_0 + [0, \Delta]$
                $m_{n-n_0} \leftarrow 0, n \leftarrow n + p^2$
    for $0 \leq j \leq \Delta$ do
        if $m_j \neq 0 \wedge \Pi_j \neq n_0 + j$ then
            $m_j \leftarrow -m_j$
    return $m$
```

Time: $O((\sqrt{n_0} + \Delta) \log\log(n_0 + \Delta))$. **Space:** $O(\sqrt{n_0} + \Delta)$.

Computation of $\mu(n)$. It is well-known (see [Dre93, §4], say) that – as we are about to see – the sieve of Eratosthenes (segmented or not) can be used to compute $\mu(n)$ for all n in a range.

We first set m_{n-n_0} and Π_{n-n_0} to 1 for all n in an interval $[n_0, n_0 + \Delta]$. We find the primes $p \leq \sqrt{n_0 + \Delta}$ using a simple sieve (or a segmented sieve). For each such p, we go through all n divisible by p in the interval; we set m_{n-n_0} to $-m_{n-n_0}$ (or rather to 0, if $p^2|n$) and Π_{n-n_0} to $p \cdot \Pi_{n-n_0}$. At the end, we go through all n in the interval one last time, and, if $\Pi_{n-n_0} \neq n$, we set m_{n-n_0} to $-m_{n-n_0}$, as, if $m_{n-n_0} \neq 0$, the fact that $\Pi_{n-n_0} \neq n$ means there is a prime divisor $> \sqrt{n_0 + \Delta}$ of n we have not considered. Now m_{n-n_0} equals $\mu(n)$ for all n in the interval. See Algorithm 4.

The procedure takes space $O(\Delta)$ and time $O((\sqrt{n} + \Delta) \log\log(n_0 + \Delta))$ per segment $[n_0, n_0 + \Delta]$. We can easily apply it to compute $\mu(n)$ for all $n \leq N$ in space $O(\sqrt{N})$ and time $O(\sqrt{N} \log\log N)$.

Optimization. There are several simple improvements to the basic algorithm. In the interest of keeping the pseudocode readable, let us discuss them briefly, rather than including them in the pseudocode.

Since we sieve by the primes $p \leq \sqrt{n_0 + \Delta}$ in increasing order, we can produce the list of those primes by a segmented sieve as well, and thus decrease total memory usage from $O(n^{1/2} + \Delta)$ to $O(n^{1/4} + \Delta)$.

We can initialize m_{n-n_0} by repeating a pattern mod $2 \cdot 3 \cdot 5 \cdots 11$ (say), so that it takes the effect of primes $p \leq 11$ into account from the beginning and we do not need to go through such primes and their multiples. It is also possible to replace several multiplications by additions, and to save on space in the process: instead of initializing

Π_{n-n_0} to 1, and multiplying Π_{n-n_0} by p whenever n divisible by p, we can initialize an array of variables s_j to 0, and add $\lceil \log_4 p \rceil$ (an operation that takes very time to compute on a binary computer) to s_{n-n_0} whenever n is divisible by p. We can already find this trick in [Hur18] and [Kuz]. It is an improvement because addition generally takes less time than multiplication.

Since $\lceil \log_4 p \rceil < 2 \log_4 p$ for $p > 2$, the loss of precision incurred in rounding up $\log_4 p$ is not enough to ruin the test of whether $\prod_{p \leq \sqrt{n_0+\Delta}: p|n} p$ equals n: if the value of ν_{n-n_0} at the end (namely, $\nu_{n-n_0} = \sum_{p \leq \sqrt{n_0+\Delta}: p|n} \lceil \log_4 p \rceil$) is less than $\lceil \log_4 n \rceil$, then, evidently, n has a prime factor that was unaccounted for, and so μ_j must be flipped, so as to give the correct value of $\mu(n)$. Conversely, if $\sum_{p \leq \sqrt{x_0}: p|n} \lceil \log_4 p \rceil \geq \lceil \log_4 n \rceil$, then

$$\sum_{\substack{p \leq \sqrt{x_0} \\ p|n}} \log_4 p \geq \frac{1}{2} \sum_{\substack{p \leq \sqrt{x_0} \\ p|n}} \lceil \log_4 p \rceil \geq \frac{1}{2} \lceil \log_4 n \rceil,$$

and so n cannot have a prime factor larger than $\sqrt{n}$. Then μ_j must not be flipped.

We can, of course, combine this idea and the one above, initializing all m_{n-n_0} so as to take the effect of the primes $p < 17$ into account, and then taking logarithms to base 256 rather than base 4. We can take the effect of 2^2, 3^2 into account as well, in the sense of initializing m_{n-n_0} to repeat a pattern of length $2^2 \cdot 3^2 \cdot 5 \cdots 13 = 180180$, so that divisibility by $p < 17$ and by 2^2, 3^2 need not be tested each time.

Factoring n. Essentially the same procedure as for computing $\mu(n)$ works for factoring all n in a range. The main differences are that a little more memory needs to be used, and that the last of the three improvements discussed above is pointless, as we seem to have to store the product Π_{n-n_0} in any event. See Algorithm 5.

4.5.2 Further improvements: shorter segments

Whether we use a segmented sieve to find primes, to compute $\mu(n)$ or to factor integers, we can parallelize the sieving process in a straightforward way. We simply have to let different processors take care of different segments. This kind of parallel programming is called "trivial parallelism"; different processes barely need to communicate. Such parallelism is well within what an occasional programmer ought to be able to implement. The computation of $\mu(n)$ (and $\mu(n)(\log n)/n$) for $n \leq 10^{14}$, necessary for Lemma 5.10, was carried out in such a way.

Oliveira e Silva (ca. 2003; see [OeS]) has shown how to implement the segmented sieve of Eratosthenes using the cache in a particularly efficient way. The length Δ of each segment $[n_0, n_0 + \Delta]$ still has to be $\gg \sqrt{n_0}$, but only part of the interval need to be kept in cache at any given time, together with some auxiliary information. In principle, Oliveira e Silva's method could lead the processor to have only $C n_0^{1/4}$ integers in cache at any one time, and yet have to load data from main memory into cache only a small positive proportion of the time. (A more precise analysis may require studying how particular processors use cache.)

We should also, however, ask ourselves whether it is possible, working on a more abstract level, to write a more efficient sieving algorithm that works in time and space

Algorithm 5 A segmented sieve of Eratosthenes for factoring

```
function SimpleSegSievFac($n_0$,$\Delta$)                    ▷ factors all $n$ in $[n_0, n_0 + \Delta]$
Ensure: for $0 \leq j \leq \Delta$, $F_j = \{(p, v_p(n_0 + j))\}_{p|n_0+j}$
    $F_j \leftarrow \emptyset$, $\Pi_j \leftarrow 1$ for all $0 \leq j \leq \Delta$
    $P \leftarrow$ SimpleSiev($\lfloor\sqrt{n_0 + \Delta}\rfloor$)
    for $p \leq \sqrt{n_0 + \Delta}$ do
        if $P_p = 1$ then                                      ▷ if $p$ is a prime...
            $k \leftarrow 1$, $d \leftarrow p$                    ▷ $d$ will go over the powers $p^k$ of $p$
            while $d \leq n_0 + \Delta$ do
                $n \leftarrow d \cdot \lceil n_0/d \rceil$                 ▷ smallest multiple $\geq n_0$ of $d$
                while $n \leq n_0 + \Delta$ do   ▷ $n$ goes over multiples of $d$ in $n_0 + [0, \Delta]$
                    if $k = 1$ then
                        append $(p, 1)$ to $F_{n-n_0}$
                    else
                        replace $(p, k - 1)$ by $(p, k)$ in $F_{n-n_0}$
                    $\Pi_{n-n_0} \leftarrow p \cdot \Pi_{n-n_0}$, $n \leftarrow n + d$
                $k \leftarrow k + 1$, $d \leftarrow p \cdot d$
    for $0 \leq j \leq \Delta$ do
        if $\Pi_j \neq n_0 + j$ then
            $p_0 \leftarrow (n_0 + j)/\Pi_j$, append $(p_0, 1)$ to $F_j$
    return $F$
```

Time: $O((\sqrt{n_0} + \Delta) \log\log(n_0 + \Delta))$. **Space:** $O(\sqrt{n_0} + \Delta \log(n + \Delta))$.

essentially linear on the length Δ of the interval $[n, n + \Delta]$ even when Δ is smaller than $\sqrt{n}$.

It turns out that one can. Already Galway [Gal00] showed how to modify the Atkin-Bernstein sieve – which is specifically for finding primes – so as to run in essentially linear time and space on segments $[n, n + \Delta]$ of length $\Delta = O(n^{1/3})$. The Atkin-Bernstein sieve, unlike the sieve of Eratosthenes, does not seem useful for factorizing numbers or computing $\mu(n)$.

However, it is also possible to modify the segmented sieve of Eratosthenes so as to work on intervals of length $n^{1/3}(\log n)^{2/3}$, all while keeping time and space essentially linear on the number of integers being sieved. The method is the subject of [Hel20]. The main issue is how to predict when an integer $\Delta < d \leq \sqrt{n + \Delta}$ has a divisor in an interval of length Δ, so as to avoid going through all possible such d. The issue can be rephrased as the problem of finding integer points close to a curve – in this case, a hyperbola $x \mapsto n_0/x$.

Both [Gal00] and [Hel20] are ultimately based on ideas stemming from Voronoi's classic work [Vor03] on Dirichlet's divisor problem, i.e., the problem of estimating the number of points in $\mathbb{Z}^2$ under a hyperbola, extended by Sierpinski [Sie06] to give estimates for the number of points in $\mathbb{Z}^2$ inside a circle (*Gauss's circle problem*). The way to [Hel20] actually went through the computation of $\sum_{n \leq n_0} \tau(n)$ in [TEH12], the method wherein was itself inspired by Vinogradov's simplified version of Voronoi's proof (see, e.g. [Vin54, §3, exercises]). Note here that $\sum_{n\leq n_0}\tau(n)=\sum_{d\leq n_0}\lfloor n_0/d\rfloor$, and that the $\sum_{d\leq n_0}\lfloor n_0/d\rfloor$ equals the number of integer points (x,y) under a hyperbola $x\mapsto n_0/x$.

It would seem that the technique introduced in [OeSHP14] for efficient cache usage and the algorithm in [Hel20] can be combined, with the result that little cache space would be needed. Conversely, as the experimental data in [Gal00, §4] demonstrates, less than optimal cache usage can greatly diminish or nullify gains coming from a superior algorithm.

In the end, we will need neither [OeSHP14] nor [Hel20] in this book, as we will compute $\mu(n)$ only for $n\leq 10^{14}$ (and could make do with a much smaller computation). (We will occasionally use the *primesieve* program [Wal], which does implement [OeSHP14]; it finds all primes in a given interval.) The reader who wants to carry out computations much further is advised to pay attention to details of implementation, and in particular to cache usage. We will not spend more time on such matters, as we wish to discuss algorithms at a more abstract level.

4.6 VERIFYING GOLDBACH'S CONJECTURE FOR SMALL INTEGERS

As we saw in §1.1, analytic tools that prove statements in number theory for all large n often break down for n smaller than a constant C. This is the case, in particular, for the ternary Goldbach conjecture. While this means that we have to check the conjecture for $n<C$ by "brute force", i.e., computations, it does not mean that such force has to be applied without intelligence.

The obvious approach to verifying the ternary Goldbach conjecture for $n<C$ is to take each odd number $n<C$ and to try different decompositions $n=m_1+m_2+m_3$ until we find one for which m_1, m_2 and m_3 are all prime. This would take time at least proportional to n.

Here is a far better strategy, already used by Saouter [Sao98]. Assume that we already know that every even number $4\leq m\leq M$ can be written as the sum of two primes. Then all we have to do is construct a list ("ladder") of *some* primes from 3 up to C such that the difference between any two consecutive primes in the list is at least 4 and at most $M-2$, and such that C minus the last prime in the list is also at least 4 and at most M.

Once we know that such a list exists, we are done: given any odd number $7\leq n\leq C$, let p_1 be the largest prime in the list such that $p_1<n$, unless n minus that prime equals 2, in which case we let p_1 be the second largest prime in the list such that $p_1<n$. In either case, $n-p_1$ is an even number larger than 2 and at most M, and thus, by assumption, it equals the sum p_2+p_3 of some prime numbers p_2 and p_3. Hence, $n=p_1+p_2+p_3$.

Saouter followed this strategy to check the ternary Goldbach conjecture for all $n\leq 10^{20}$. Now, the computations of Oliveira e Silva, Herzog and Pardi [OeSHP14] show that every even number $4\leq m\leq 4\cdot 10^{18}$ can be written as the sum of two primes.

These computations took about 770 core-years [OeSHP14, p. 2], i.e., it would have taken 770 years on a single one-core processor[7]. This number of core-years amounts to more than 6.7 million core-hours; it is at a much larger scale than any other computation that we will use. Note also that $4 \cdot 10^{18} + 2$ is the sum of the prime numbers 211 and $4 \cdot 10^{18} - 209$. Hence, we can take $M = 4 \cdot 10^{18} + 2$. That is, we need to build a list of primes up to C such that the difference between two consecutive primes in the list is always at least 4 and at most $4 \cdot 10^{18}$.

The basic issue in building such a list is how to test for primality. We need our test to be deterministic (at least in the sense of actually giving us a proof that a number is prime, rather than simply making it likely) and reasonably fast. Polynomial-time, deterministic algorithms for testing the primality of any given integer have been known since [AKS04], but they are not yet fast in practice, at the time of writing.

The work of building a list of primes up to a given C was carried out in a joint paper by D. Platt and the author [HP13], with C set to $8.875 \cdot 10^{30}$. Like Saouter, we decided to build our list of primes out of integers of a special kind, whose primality can be checked rapidly. Following a suggestion of A. Booker, we used *Proth primes*, which are very closely related to Saouter's choice. A Proth prime is a prime number of the form $k \cdot 2^n + 1$, k odd, $1 \leq k < 2^n$.

Lemma 4.3 ([Pro78]). *Let $m = k \cdot 2^n + 1$, n an integer, k odd, $1 \leq k < 2^n$. Let a be an integer such that*

$$a^{\frac{m-1}{2}} \equiv -1 \bmod m. \tag{4.24}$$

Then m is prime.

Proof. Let m be a composite number satisfying (4.24) for some integer a. Let $p|m$ be a prime factor of m with $p \leq \sqrt{m}$. We obtain immediately from (4.24) that $a^{(m-1)/2} \equiv -1 \bmod p$, and so the order of a in $(\mathbb{Z}/p\mathbb{Z})^*$ is divisible by 2^n. However, the order of a has to divide $p - 1$, and $p - 1 < \sqrt{m} \leq 2^n$. Contradiction. □

When we build our list of primes, it is enough to test, for each candidate m, whether (4.24) is fulfilled for some a among the first ten primes, say; if it is fulfilled by none of them, we pick another candidate in the desired range.

Building the list up to $8.875 \cdot 10^{30}$ took about 40000 core-hours. This established that every odd integer $5 < n \leq 8.875 \cdot 10^{30}$ is the sum of three primes.

Building a list up to 10^{27} – which is all that is required by the proof of the ternary Goldbach in the version given in this book – takes either a day or a weekend on a single

[7] The two processor types mentioned in [OeSHP14] are Intel 2.83GHz A9550 Core 2 Quad (introduced ca. 2008) and AMD 2.20GHz Athlon64 (a few years older), if you must know. Nowadays, large-scale computations in number theory, as in most other fields, are typically carried out using large numbers of mass-market processors working in parallel. A modern processor can have several cores, that is, independent processing units; for instance, Athlon64 has two, and Core 2 Quad has four (in two chips). We speak of "core-hours" or "core-years" much as others speak of man-hours or man-years, with some of the same provisos: a task taking 100 core-hours can be done by one core working non-stop for 100 hours, or, ideally, by 100 cores working non-stop and at the same time for 1 hour, but organizing matters so that the latter estimate holds true can be far from trivial. If making the latter estimate work *is* trivial, we say that the algorithm we are using is "trivially parallelizable".

core in a modest home computer, depending on the implementation used. There have been independent checks up to 10^{27} [RT16] and 10^{28} ([PTM14], by L. Théry and B. Grégoire). The latter check included a formal proof (via the Coq Proof Assistant) of the existence of a list of primes up to 10^{28} with gaps at most 10^{18}.

There is an alternative approach: using information on the zeros of the Riemann zeta function, one can prove a short-interval estimate, i.e., one can show that there is always a prime in the interval between $(1 - \Delta^{-1})x$ and x for all $x \geq x_0$, where Δ and x_0 are appropriate constants. The best result of this kind to date is that of H. Kadiri and A. Lumley [KL14], which, taken together with [OeSHP14], gives that every odd integer $5 < n \leq 7.864 \cdot 10^{27}$ can be written as the sum of three primes [KL14, Cor. 1.2]. This result improves on [Helc, Prop. C.1]. Kadiri and Lumley's short-interval estimate is stronger than that in the estimates in [Sch76, Thm. 12] and [RS03, Thm. 2]. It relies on D. Platt's rigorous verification[8] of the Riemann Hypothesis up to height $3.061 \cdot 10^{10}$ [Pla11], together with H. Kadiri's explicit zero-free region [Kad05, Thm. 1.1] and other explicit results on the Riemann zeta function [Ros41, Thm. 19], [Kad13].

4.7 AUTOMATED THEOREM PROVING AND COMPLETENESS

There will never be a machine that proves every single true statement. This is an immediate corollary of Gödel's first incompleteness theorem, which, informally speaking, states that there are true statements without proofs: there cannot be a finite set of axioms (or, more generally, a set of axioms produced by a (finite) machine) such that every true statement about the natural numbers can be proven starting from those axioms.

However, the incompleteness theorem does not rule out automated theorem proving: a machine could prove *some* true statements. Moreover, there are, so to speak, small subareas within mathematics that are complete, i.e., every true statement in them does have a proof. In that case, a program can in principle prove every true statement simply by going through all possible proofs, though of course one might hope for a much better algorithm.

At one point in the original version of the proof of the main theorem, an elementary

[8] Using, as always, interval arithmetic for the sake of rigor. The ZetaGrid project, led by [Wed03], had gone to greater heights than Platt did in computations that did not use interval arithmetic. The statement that the ZetaGrid project verified the first $9 \cdot 10^{11}$ zeros (corresponding to $H = 2.419 \cdot 10^{11}$) is often quoted (e.g., [Bom10, p. 29]); this is the point to which the project had got by the time of Gourdon and Demichel's claim [GD04] that they had checked that the first 10^{13} zeros of the Riemann zeta function (meaning all zeros of imaginary part up to $H = 2.44599 \cdot 10^{12}$) lie on the critical line $\Re z = 1/2$. It is unclear whether Gourdon and Demichel's computation was or could be made rigorous. As pointed out in [SD10, p. 2398], it has not been replicated yet. Wedeniwski asserts in private communication that his project verified the first 10^{12} zeros, and that the computation was double-checked (by the same method). Unfortunately, his project's results have never been formally published. We will, at any rate, use only Platt's results, and so did Kadiri and Lumley.

inequality about real numbers needed to be proved:

$$1 + \frac{y_1 y_2}{(1-y_1+x)(1-y_2+x)} \leq \frac{(1-x^3)^2(1-x^4)}{(1-y_1y_2)(1-y_1y_2^2)(1-y_1^2y_2)} \tag{4.25}$$

for $0 \leq x \leq y_1, y_2 < 1$ with $y_1^2 \leq x$, $y_2^2 \leq x$. This sort of inequality could doubtlessly have been proven by tedious casework or by being more clever than the author. Instead, the proof of the (4.25) was partly automated. Due to changes in tactics, we no longer need (4.25), but it would be a pity to waste an opportunity to touch on an interesting topic.

The theory of ordered real closed fields – that is, fields obeying a certain set of axioms (see, e.g., [CK90, p. 41]) obeyed, in particular, by $\mathbb{R}$ – is complete, as was proved by Tarski [Tar51]. Tarski also gave a procedure for *quantifier elimination*, which proves (or disproves) any given statement in the theory. That procedure had non-elementary complexity, i.e., its running time was larger than any tower of exponentials $\left(\left(\left((2^2)^2\right)^{\cdots}\right)^2\right)^n$ of fixed length, where n is the size of the statement. However, since [Col75], we have an algorithm (by *cylindrical decomposition*) whose running time is doubly exponential on the number of variables.

Doubly exponential running time is of course still enormous. It is optimal in the sense that some quantifier-elimination problems require that much time. Nevertheless, it is known that existential and universal statements – that is, statements of the form “there are $x_1, \ldots, x_k$ such that…” or “for all $x_1, \ldots, x_k$”, followed by a formula – can be proven in exponential time [BPCR06, Ch. 13].

As it happens, there is at least one easily available implementation of cylindrical decomposition [HB11], whereas, at the time of writing, there does not yet seem to be a full implementation of an exponential-time algorithm available. The doubly exponential algorithm turned out to run well after the number of variables in (4.25) was lowered from 3 to 2 by some algebraic manipulations done by hand.

* * *

Another interesting matter is that of *proof assistants* and automated proof checkers. When we speak of proofs in the context of results in logic – such as incompleteness theorems – we think of a proof as a sequence of symbols (called a *formal proof*) whose validity can be checked without any sort of thought, simply by letting a monkey turn a crank on some sort of barrel organ (now called “computer”). This is so even though mathematicians never actually write their proofs in this way; they are texts written for human readers. Nowadays, however, producing a formal proof can be an option in practice: there are programs, called proof assistants, that can be used to write out a proof as a formal proof – the dullest parts of the task are automated. So far, this is possible only for some proofs in some areas, simply because there are large areas of mathematics that have not yet been encoded in proof assistants. Work in analytic number theory has already started: there are now two computer-aided formal proofs of the prime number theorem – one elementary, one analytic ([ADGR07], [Har09]) and even a formalization of a large part of an introductory graduate textbook [Ebe19].

As we mentioned in §4.6, a particularly elementary and repetitive part of the present proof – namely, part of the verification of ternary Goldbach for small integers – has been turned into a formal proof with the help of a proof assistant [PTM14]. We are probably still quite some way from being able in practice to turn something like the entirety of the proof in this book into a formal proof. The main practical use of formal proofs at the moment – other than to show that formal-proof systems can already be used – may be to verify proofs whose thorough verification would be extraordinarily arid for any human reader. It is genuinely hoped that the reader will enjoy the present proof enough not to envy a monkey.

Chapter Five

Basic sums of arithmetical functions

When we have to estimate a quantity, we often proceed by reducing it to a quantity of standard type, meaning one for which estimates are by now known. In analytic number theory, the quantities to be estimated are often sums. Of course, for explicit work, we need explicit estimates on some common sums.

The usual procedure for estimating a sum $\sum_{n\leq x} f(n)$ (say) consists in

- obtaining an explicit result for x large, using the same techniques that are used to obtain asymptotic results, but working all constants out, optimizing them whenever possible,
- using computations to obtain bounds, or precise results, for x small.

There are two provisos to be made here. One is that some techniques do not lend themselves well to proving explicit results. This includes not just ineffective results, but also results that are effective in principle but, given the state of our knowledge, bound to give poor constants. Estimates involving contour integration – bread-and-butter for analytic number theorists – sometimes, though not always, involve bounds in which the best known explicit constants are much too large. We will often have to content ourselves with bounds that may not be best asymptotically but are good in practice.

The second proviso is that computations and explicit analytic work should not be thought of as separate. Analytic methods can take computational inputs. It will also become apparent that it can be useful to combine results for large and small x – obtained by different means – into a single bound valid for all x. Such a bound can then be a convenient input for further analytical work.

We estimate simple sums over square-free numbers in §5.1. It is very straightforward to get reasonably good estimates on square-free numbers, simply by inclusion-exclusion. One can obtain better constants by iteration. Here we simply quote a result from the literature [CDE07], complement it with computations for small x and adapt it in various ways.

Estimating sums of $\Lambda(n)$, such as $\psi(x) = \sum_{n\leq x} \Lambda(n)$, is a far deeper issue. The short and the long of it is that, while zero-free regions are the basis of the prime number theorem as it is usually presented, they are neither sufficient nor always indispensable in the explicit regime. We can compute $\psi(x)$ for x bounded, and often can make do with bounds of type $(1-\epsilon)x < \psi(x) < (1+\epsilon)x$ for larger x, provided ϵ is very small. We simply quote such results from the literature, explaining where they come from. We also show how to use them to prove estimates on more general sums of $\Lambda(n)$, in that we will set up a straightforward partial-summation machine.

Estimating sums of $\mu(n)$ is even more delicate, in that a direct analytic approach is not as of yet feasible. There are some simple classical bounds. For better results, the working approach since [Sch69] has been to prove bounds on sums of $\mu(n)$ by the clever usage of bounds on $\psi(x)$. Again, we will be quoting results from the literature, many of them by O. Ramaré.

We will supplement such results by our own computational work (§5.3.2). The efficient computation of $\mu(n)$ for all $n \leq x$ is a non-trivial matter, particularly when it comes to space usage. We are in a setting where space optimization can lead to faster running times in practice.

In §5.3.3, we use convexity to combine results for small x and large x. This way of combining estimates results in bounds whose form is convenient in several senses: it will later allow us to reduce sums to power series, and it also makes it possible to give improved bounds (§5.3.4) on sums of μ with coprimality conditions.

5.1 SUMS OVER SQUARE-FREE NUMBERS

Let $Q(x)$ be the number of square-free numbers $\leq x$. Estimating $Q(x)$ is an example of something that can be done well by means of a very naïve sieve, i.e., inclusion-exclusion. Indeed, by Möbius inversion (2.1),

$$\begin{aligned} Q(x) &= \sum_{n\leq x} \sum_{d:d^2|n} \mu(d) = \sum_{n\leq x} \sum_{\substack{d\leq\sqrt{x} \\ d^2|n}} \mu(d) \\ &= \sum_{d\leq\sqrt{x}} \mu(d) \sum_{\substack{n\leq x \\ d^2|n}} 1 = \sum_{d\leq\sqrt{x}} \mu(d) \left\lceil \frac{x}{d^2} \right\rceil . \end{aligned}$$

Hence

$$\begin{aligned} Q(x) &= \sum_{d\leq\sqrt{x}} \mu(d) \frac{x}{d^2} + O^*\left(\sum_{d\leq\sqrt{x}} \mu^2(d) \right) \\ &= \frac{x}{\zeta(2)} + O^*\left(Q(\sqrt{x})\right) - x \sum_{d>\sqrt{x}} \frac{\mu(d)}{d^2}. \end{aligned}$$

From this equation , we can already obtain bounds of the form $Q(x) = x + cO^*(\sqrt{x})$. It is clear that we can get $c = 2$: use the bound $Q(x) \leq x - 1$ for $x \geq 4$, and a trivial bound on $\sum_{d>x} \mu(d)/d^2$. We can obtain somewhat better values of c by iteration, especially if we express $\sum_{d>x} \mu(d)/d^2$ in terms of $Q(x)$. Doing much better requires a finer analysis – finding cancellation in μ and using a different recurrence formula, as in [CDE07]. (We will discuss how to find cancellation in sums involving μ in §5.3.) A non-explicit approach along these lines can already be found in [Lan09b, p. 606–609]

The following shall be enough for our purposes.

Lemma 5.1. *Let $Q(x)$ be the number of square-free integers $\leq x$. Then*

$$Q(x) = \frac{6}{\pi^2}x + \begin{cases} O^*(0.036438\sqrt{x}) & \text{for } x \geq 82005, \\ O^*(\sqrt{x}/6) & \text{for any } x > 1005, \\ O^*(\sqrt{3}\left(1-\frac{6}{\pi^2}\right)\sqrt{x}) & \text{for any } x > 0. \end{cases} \tag{5.1}$$

Moreover, for $S(x) = \sum_{n>x} \mu^2(n)/n^2$,

$$S(x) = \frac{6}{\pi^2 x} + \begin{cases} O^*\left(\frac{0.085023}{x^{3/2}}\right) & \text{if } x \geq 2909, \\ O^*\left(\frac{0.3}{x^{3/2}}\right) & \text{if } x \geq 10, \end{cases} \tag{5.2}$$

and

$$S(x) \leq \begin{cases} 0.80946/x & \text{for all } x \geq 2, \\ 15/\pi^2 x & \text{for all } x > 0. \end{cases} \tag{5.3}$$

Proof. The cases $x \geq 82005$ and $x \geq 1005$ of equation (5.1) were proved in [CD88]. The case of $x > 0$ arbitrary follows from the case $x \geq 1005$ and a computation up to 1005. (Since $c_0 = \sqrt{3}(1 - 6/\pi^2)$ satisfies $c_0 \leq 6/\pi^2$, it is enough to check the inequality $Q(x) \geq (6/\pi^2)x - c_0\sqrt{x}$ for $x \to n^-$, $1 \leq n \leq 1005$ square-free (note that 1005 is itself square-free), together with the inequality $Q(x) \leq (6/\pi^2)x + c_0\sqrt{x}$, for $x = n$, $1 \leq n \leq 1005$ square-free.)

Let us prove (5.2). Let $c = 0.036438$. By Abel summation,

$$\begin{aligned} S(x) &= 2\int_x^\infty \frac{Q(t)-Q(x)}{t^3}dt = 2\int_x^\infty \frac{Q(t)}{t^3}dt - \frac{Q(x)}{x^2} \\ &= \left(2x\int_x^\infty \frac{dt}{t^2} - 1\right)\frac{6}{\pi^2}\frac{1}{x} + O^*\left(2c\int_x^\infty \frac{dt}{t^{5/2}} + \frac{c}{x^{3/2}}\right) \\ &= \frac{6}{\pi^2}\frac{1}{x} + O^*\left(\frac{7c/3}{x^{3/2}}\right) \end{aligned}$$

for $x \geq 82005$. Thus, (5.2) holds for $x \geq 82005$. We verify (5.2) for $x < 82005$ by brute force, noting that $S(x) = \zeta(2)/\zeta(4) - \sum_{n\leq x}\mu^2(n)/n^2$. (Note also that $S(x)x^{3/2} - \sqrt{x}/\zeta(2)$ has positive derivative on $(n, n+1)$ for every $n \geq 1$, and so it is enough to check (5.2) for $x \to n^-$ and $x = n$, $1 \leq n \leq 1005$ square-free.)

For $x \geq 10$, we obtain (5.3) (or, rather, a stronger bound) from (5.2). For $0 < x \leq 10$, we check (5.3) computationally; the first bound is (almost) tight when $x \to 3^-$, and the second bound is tight when $x \to 1^-$. □

We will sometimes need to restrict to the odd numbers.

Lemma 5.2. *Let $Q_2(x)$ be the number of odd square-free integers $\leq x$. Then*

$$Q_2(x) = \frac{4}{\pi^2}x + \begin{cases} O^*\left(\frac{9}{70}\sqrt{x}\right) & \text{for } x \geq 1573, \\ O^*\left(\left(1-\frac{4}{\pi^2}\right)\sqrt{x}\right) & \text{for } x > 0. \end{cases} \tag{5.4}$$

Moreover, $S_2(x) = \sum_{n>x, n \text{ odd}} \mu^2(n)/n^2$,

$$S_2(x) = \frac{4}{\pi^2 x} + O^*\left(\frac{0.3}{x^{3/2}}\right) \tag{5.5}$$

for $x \geq 3$*, and*

$$S_2(x) \leq \frac{12}{\pi^2 x} \qquad \textit{for all } x > 0. \tag{5.6}$$

Proof. By a computation, for all $41002 \leq x \leq 82005$,

$$Q_2(x) = \frac{4}{\pi^2} x + O^*(8.67142). \tag{5.7}$$

Using inclusion-exclusion and inequalities (5.1) and (5.7), and setting $k_0 = \lfloor \log_2(x/82005) \rfloor$, we obtain that, for $x \geq 82005$,

$$\begin{aligned} Q_2(x) &= \sum_{k=0}^{k_0} (-1)^k Q(x/2^k) + (-1)^{k_0+1} Q_2(x/2^{k_0+1}) \\ &= \sum_{k=0}^{k_0} (-1)^k \frac{6}{\pi^2} \frac{x}{2^k} + (-1)^{k_0+1} \frac{4}{\pi^2} \\ &+ O^*\left(0.036438 \sum_{k=0}^{\infty} \sqrt{\frac{x}{2^k}}\right) + O^*(8.67142) \\ &= \frac{4}{\pi^2} x + O^*\left(0.12441\sqrt{x} + 8.67142\right). \end{aligned} \tag{5.8}$$

Thus, $Q_2(x) \leq (4/\pi^2)x + O^*((9/70)\sqrt{x})$ for $x \geq 4342058$. We check by brute force that this same inequality is also true for $818 \leq x \leq 4342058$. We derive our estimate (5.5) from our estimate on $Q_2(x)$ much as in the proof of Lemma 5.1: for $x \geq 818$,

$$S_2(x) = \frac{4}{\pi^2} \frac{1}{x} + O^*\left(\frac{7}{3} \frac{9/70}{x^{3/2}}\right) = \frac{4/\pi^2}{x} + O^*\left(\frac{3/10}{x^{3/2}}\right).$$

We check by computation that this inequality also holds for $3 \leq x \leq 818$.

We derive (5.6) from (5.5) for $x \geq 7$, and check it by computation for $0 < x < 7$. The first bound in (5.6) is almost tight for $x \to 5^-$, and the second bound is tight for $x \to 1^-$. □

We can use Lemmas 5.1 and 5.2 to prove a poor man's bound on alternating sums with coprimality conditions. It is far from an asymptotic, but it will be useful as a simple, general bound. One can prove stronger bounds for large x by using the results we will quote in §5.3; see [Ramb, Thm. 8.1].

Lemma 5.3. *Let* $m \in \mathbb{Z}^+$*,* $x > 0$*. Then*

$$\left| \sum_{\substack{n>x \\ (n,m)=1}} \frac{\mu(n)}{n^2} \right| \leq \frac{4}{\pi^2 x} + \frac{1}{x^{3/2}} \cdot \begin{cases} \sqrt{8} - \frac{8^{3/2}}{\pi^2} & \textit{if } m \textit{ is odd,} \\ 1 - \frac{4}{\pi^2} & \textit{if } m \textit{ is even.} \end{cases} \tag{5.9}$$

Here $1 - 4/\pi^2 = 0.59471\ldots$ and $\sqrt{8} - 8^{3/2}/\pi^2 = 0.53579\ldots$.

Proof. Write $R_m(x)$ for the sum on the left side of (5.9). For $0 < x < 1$,

$$R_m(x) = R_m(0) = \prod_{p \nmid m} \left(1 - \frac{1}{p^2}\right)$$

is positive and ≤ 1; if m is odd, it is $\leq 3/4$. Taking derivatives, we see that, for any $c_0 > c_1 > 0$, the function $x \mapsto c_0 x^{3/2} - c_1 x^{1/2}$ on $[0,1]$ first decreases and then increases; comparing its values at $x = 0$ and at $x = 1$, we see that it has its maximum at $x = 1$. Thus, for $0 < x \leq 1$, $(1 - (4/\pi^2)/x) \cdot x^{3/2} \leq 1 - 4/\pi^2$ and $(3/4 - (4/\pi^2)/x) \cdot x^{3/2} \leq 3/4 - 4/\pi^2$. Thus (5.9) holds for $0 < x < 1$.

For m even and $1 \leq x < 3$, $R_m(x) = R_m(1) = R_m(0) - 1 \in [(4/3)/\zeta(2) - 1, 0]$. For $c_0, c_1 > 0$ with $c_1 < 3c_0$, the function $x \mapsto c_0 x^{3/2} - c_1 x^{1/2}$ is decreasing on $[1, \infty)$. Hence, the map $x \to (1 - (4/3)/\zeta(2) - (4/\pi^2)/x) \cdot x^{3/2}$ has its maximum on $[1,3]$ at $x = 3$, where it equals $0.28233\ldots$. From (5.5), we know that, for m even and $x \geq 3$,

$$|R_m(x)| \leq \frac{4}{\pi^2 x} + \frac{0.3}{x^{3/2}}. \tag{5.10}$$

Since $0.28233\ldots < 0.3$, the inequality (5.10) holds in fact for m even and $x \geq 1$. We conclude that (5.9) holds for m even and $x > 0$ arbitrary.

For m odd and $1 \leq x < 2$, $R_m(x) = R_m(1) \in [1/\zeta(2) - 1, -1/4]$. By the same reasoning as above, the maximum of $x \mapsto (1 - 1/\zeta(2) - (4/\pi^2)/x) \cdot x^{3/2}$ on $[1,2]$ is attained at $x = 2$. It equals $\sqrt{8} - 8^{3/2}/\pi^2$.

For m odd and general x,

$$R_m(x) = R_{2m}(x) - \frac{1}{4} R_{2m}\left(\frac{x}{2}\right) = \frac{3}{4} R_{2m}(x) - \frac{1}{4} \sum_{\substack{\frac{x}{2} < n \leq x \\ (n,2m)=1}} \frac{\mu(n)}{n^2}.$$

Now, by (5.10), $(3/4)|R_{2m}(x)| \leq 3/\pi^2 x + 0.225/x^{3/2}$ for $x \geq 1$, whereas

$$\left| \frac{1}{4} \sum_{\substack{\frac{x}{2} < n \leq x \\ (n,2m)=1}} \frac{\mu(n)}{n^2} \right| \leq \frac{1}{4} \sum_{\substack{\frac{x}{2} < n \leq x \\ n \text{ odd}}} \frac{\mu^2(n)}{n^2}.$$

By Lemma 5.2,

$$\frac{1}{4} \sum_{\substack{\frac{x}{2} < n \leq x \\ n \text{ odd}}} \frac{\mu^2(n)}{n^2} \leq \frac{1}{\pi^2 \frac{x}{2}} - \frac{1}{\pi^2 x} + O^*\left(\frac{0.3/4}{(x/2)^{3/2}} + \frac{0.3/4}{x^{3/2}}\right) = \frac{1}{\pi^2 x} + \frac{\frac{0.3}{4}(2^{3/2}+1)}{x^{3/2}}$$

for $x \geq 6$, whereas, for $2 \leq x < 6$, a direct computation shows that the weaker bound $1/\pi^2 x + 0.308/x^{3/2}$ holds. Hence, for $x \geq 2$ and m odd,

$$|R_m(x)| \leq \frac{4}{\pi^2 x} + \frac{0.533}{x^{3/2}} = \frac{4}{\pi^2 x} + \frac{0.533}{x^{3/2}}.$$

Thus, (5.9) holds for all odd m. □

5.2 SUMS OVER PRIMES

We will need some bounds on sums involving the von Mangoldt function Λ. They should be fully explicit and robust, in the sense of being valid over wide ranges. For the most part, we will just be quoting the literature and deriving the results we need from it in a fairly straightforward fashion.

5.2.1 Bounds on $\psi(x)$ and $\vartheta(x)$

Define

$$\psi(x) = \sum_{n \leq x} \Lambda(n), \qquad \vartheta(x) = \sum_{p \leq x} \log p.$$

We quote from [RS62, Thm. 9]: for $x > 0$,

$$\psi(x) < 1.03883x. \tag{5.11}$$

By Table 1 in [RR96, p. 419],

$$\psi(x) - x = O^*(0.000213x) \tag{5.12}$$

for all $x \geq 10^{10}$. Bounds such as (5.11) can be proven by finite verifications of the Riemann hypothesis up to a finite height, in the style of §4.3. The use of zero-free regions for $\zeta(x)$ is helpful but not indispensable. In contrast, bounds of the form $\psi(x) = (1 + \mathrm{err}(x))x$, where $\mathrm{err}(x) = o(x)$, do require the use of zero-free regions, in conjunction with finite verifications of RH. Such is the case, for instance, for the estimate

$$\psi(x) - x = O^*\left(0.006409 \frac{x}{\log x}\right), \tag{5.13}$$

valid for $x \geq 1514981$. This bound was proved for $x \geq e^{22}$ in [Dus98, Thm. 1.3], improving on [Sch76, Thm. 7*]. Much as was noted in [Ram13b, Lem. 4.2], the bound also holds for $1514981 \leq x < e^{22}$, as can be verified by a simple computation.

Of course, the fact that there are elementary proofs of the prime number theorem means that one *can* prove that $\psi(x) = (1 + o(1))x$ without using *any* information on $\zeta(s)$ for $\Re s < 1$. Such proofs, however, generally give much weaker bounds than those above, whose proofs do use such information.

In this section, we will have enough with (5.12) and analogous bounds, supplemented by computational checks for small x. By the bottom table in [RR96, p. 423],

$$\psi(x) = x + O^*(\sqrt{x}) \tag{5.14}$$

for $8 \leq x \leq 10^{10}$, and so, by a check by hand for $x \leq 8$,

$$\psi(x) = x + O^*(\sqrt{2x}) \tag{5.15}$$

for all $0 \leq x \leq 10^{10}$. By (5.12), (5.15) and a check of $\psi(n)/n$ for all integers $10^5 \leq n < 2/0.001522^2 \leq 863378$,

$$\psi(x) \leq 1.001522x \tag{5.16}$$

for all $x \geq 10^5$. Here and in what follows, computations are based on the sieve of Eratosthenes (§4.5).

By [Sch76, §9, Thm. 6*] and a check for $10^5 \leq x < 1155901$,

$$\vartheta(x) > 0.995268x \tag{5.17}$$

for all $x \geq 10^5$. (It is enough to check $\vartheta(n-1)/n$ for all integers $10^5 \leq n < 1155901$.) Again by a computation,

$$\begin{aligned} \vartheta(x) &\geq x - 1.83264\sqrt{x} \qquad \text{for } 10^5 \leq x \leq 10^{10}, \\ \vartheta(x) &\leq x - 0.260433\sqrt{x} \qquad \text{for } 1 \leq x \leq 10^{10}. \end{aligned} \tag{5.18}$$

By Table 1 in [RR96, p. 419],

$$\vartheta(x) \geq (1 - 0.000213)x \tag{5.19}$$

for all $x \geq 10^{10}$. Finally, by [PT16, Cor. 2],

$$\vartheta(x) < (1 + 7.5 \cdot 10^{-7})x \tag{5.20}$$

for all $x > 0$.

The bounds on large x we have quoted are not in general tight. It is often the case that a paper gives bounds of the type $|\psi(x) - x| \leq \epsilon x$ with different ϵ for x in different ranges – say, for $x \geq x_0$ and for $x \geq x_1$, where $x_1 > x_0$. We can then obtain an improved bound on $x \geq x_0$ by combining the bound for $x \geq x_1$ with a computational verification for $x_0 \leq x \leq x_1$. For instance, by Table 1 in [RR96, p. 419],

$$\psi(x) - x = O^*(0.000015x) \tag{5.21}$$

for all $x \geq 10^{13}$. A computation[1] gives us that

$$\psi(x) - x = O^*(6.76846 \cdot 10^{-6}x) \tag{5.22}$$

for $10^{10} \leq x \leq e^{30} = 1.06864\ldots \cdot 10^{13}$. Hence

$$\psi(x) - x = O^*(0.000015x) \tag{5.23}$$

for all $x \geq 10^{10}$.

One can also use recent, stronger bounds, combined with computational verifications. For instance, Faber and Kadiri ([FK15], [FK18]) use an explicit zero-density estimate on $\zeta(s)$ [Kad13] together with Platt's verification of the Riemann Hypothesis up to height $3.061 \cdot 10^{10}$ [Pla11] and an improvement on the constant of the classical zero-free region for $\zeta(s)$ [Kad05] to prove that

$$\psi(x) - x = O^*(2.811 \cdot 10^{-6}x) \tag{5.24}$$

[1] Since the range of the computation here is not small, it is best to use a highly optimized implementation of the Eratosthenes sieve for primes, such as the program *primesieve* [Wal]. All computations in the present section up to this point were fairly small; a homebrewed implementation of a sieve would have been enough for them.

for all $x \geq e^{30}$. Combining this result with inequality (5.22), we obtain

$$\psi(x) - x = O^*(6.76846 \cdot 10^{-6} x) \tag{5.25}$$

for all $x \geq 10^{10}$. Notice that the bound in (5.25) is essentially optimal within its range, as the bound is actually reached (up to the last significant digit) for some $x \in [10^{10}, e^{30}]$.

Actually, we will not use (5.25). The results in this section will be used as auxiliary bounds; having the best possible bounds is not necessary and will often not even be useful. In such a situation, it seems reasonable to use older bounds such as those in [RR96, p. 419], simply so as to depend on fewer recent advances.

Remark 1. Given results of the form $\psi(x) - x = O^*(\epsilon x)$ in different ranges, one can put them together with bounds of the form

$$|\psi(x) - x| \leq c_1 \sqrt{\log x} \cdot e^{-c_2 \sqrt{\log x}} \cdot x, \tag{5.26}$$

coming from the full use of classical zero-free regions, to obtain bounds of the form $\psi(x) - x = O^*(\epsilon' x / \log x)$. The best values of c_1 and c_2 in (5.26) to date are those in [Tru16]. For the sake of giving a rough idea of how different bounds compare: the bound from [Tru16] becomes better than those in [FK18, Table I] only at some point between $x = e^{3500}$ and $x = e^{4000}$.

Remark 2. A few words are in order on how one might go about proving stronger results of type $\psi(x) - x = O^*(\epsilon x)$ than those currently available. Most work on explicit bounds of this type has followed [Ros41] in using iterated integration, which amounts to approximating a sharp truncation by a piecewise polynomial smoothing. Faber and Kadiri use a polynomial smoothing from [RS03], optimal in its given framework; however, it is a framework that arguably tilts the matter towards polynomial or piecewise polynomial smoothings, in that it does not take advantage of a possible faster-than-polynomial decay of the Mellin transform.

It would seem clear that one also ought to consider other frameworks and other kinds of smoothing. For instance, one may consider linear combinations of exponential smoothings $e^{-\alpha t/x}$, $\alpha \in \mathbb{C}$. There is an entire literature on optimizing approximations to (for instance) characteristic functions of intervals by means of functions whose Fourier transforms have compact support (starting with [Beu38], [Sel72], [Vaa85]). This possible direction is only one of several.

5.2.2 Bounds on general sums over primes

We would now like to use the bounds on $\psi(x)$ and $\vartheta(x)$ quoted above to derive bounds on sums of the form $\sum_n \Lambda(n) f(n)$ or $\sum_p f(p)$, where f is a weight function, continuous or not. This is the easy, "Abelian" direction, as opposed to the converse direction ("Tauberian"), which is harder and would require additional conditions.

(When the weight f is continuous and precision is critical, it is better to take advantage of the continuity of f to estimate $\sum_n \Lambda(n) f(n)$ directly, rather than using an estimate on $\psi(x)$. We will see how in Ch. **??**. However, when f is not continuous or we just need a quick bound for some auxiliary purpose, it can make sense to use an estimate on $\psi(x)$. Of course, as we have already mentioned in passing, the best existing estimates on $\psi(x)$ involve approximating the characteristic function of the interval $[1, x]$ by a continuous weight.)

Let us first prove a very general partial-summation result of a form fitting our needs.

Proposition 5.4. *Let $g : \mathbb{Z}^+ \to \mathbb{Z}$ satisfy*

$$\sum_{n\leq x} g(n) \geq x - \begin{cases} c_-\sqrt{x} & \text{for } x \leq x_0, \\ \epsilon_- x & \text{for } x \geq x_0, \end{cases} \qquad \sum_{n\leq x} g(n) \leq x + \begin{cases} c_+\sqrt{x} & \text{for } x \leq x_0, \\ \epsilon_+ x & \text{for } x \geq x_0, \end{cases} \tag{5.27}$$

where $c_+, c_-, \epsilon_+, \epsilon_- \in \mathbb{R}$ and $x_0 \geq 0$. Let $f : \mathbb{R} \to \mathbb{R}$ be a function of bounded variation such that $f(t) = o(1/t)$ as $t \to \infty$ and $f(t)$, $tf'(t)$ are in L^1.

Then

$$\begin{aligned} \sum_n g(n)f(n) \geq& \int_0^\infty f(t)dt - c_+\int_0^{x_0} \sqrt{t}\delta_+(t)dt - c_-\int_0^{x_0} \sqrt{t}\delta_-(t)dt, \\ &- \epsilon_-\int_{x_0}^\infty t\delta_-(t)dt - \epsilon_+\int_{x_0}^\infty t\delta_+(t)dt, \\ \sum_n g(n)f(n) \leq& \int_0^\infty f(t)dt + c_-\int_0^{x_0} \sqrt{t}\delta_+(t)dt + c_+\int_0^{x_0} \sqrt{t}\delta_-(t)dt, \\ &+ \epsilon_-\int_{x_0}^\infty t\delta_+(t)dt + \epsilon_+\int_{x_0}^\infty t\delta_-(t)dt, \end{aligned} \tag{5.28}$$

where $\delta_+(t) = f'(t)$ when $f'(t) > 0$ and $\delta_+(t) = 0$ otherwise, and $\delta_-(t) = -f'(t)$ when $f'(t) < 0$ and $\delta_-(t) = 0$ otherwise. In particular, for $c = \max(c_-, c_+)$, $\epsilon = \max(\epsilon_-, \epsilon_+)$,

$$\sum_n g(n)f(n) = \int_0^\infty f(t)dt + O^*\left(c\int_0^{x_0} \sqrt{t}|f'(t)|dt + \epsilon\int_{x_0}^\infty t|f'(t)|dt\right). \tag{5.29}$$

Here, as is usually the case for us, f need not be everywhere differentiable, or even continuous. In such cases, $|f'|$, δ_- and δ_+ are to be understood as a distributions, or alternatively, $\delta_+(t)dt$, $\delta_-(t)dt$ are to be seen as shorthand for the measures μ_+, μ_- into which the measure $df = \mu_+ - \mu_-$ is decomposed. See §2.3.3.

Proof. Let $G(t) = \sum_{n\leq t} g(n)$. By Abel's summation formula (3.12),

$$\begin{aligned} \sum_n g(n)f(n) &= \lim_{x\to\infty}\left(G(x)f(x) - \int_0^x G(t)f'(t)dt\right) \\ &= -\int_0^\infty G(t)f'(t)dt, \end{aligned} \tag{5.30}$$

where we use the assumptions $G(t) \ll t$, $f(t) = o(1/t)$ and $tf'(t) \in L^1$.

By (5.27),

$$-\int_0^{x_0} G(t)f'(t)dt \geq -\int_0^{x_0} tf'(t)dt - \int_0^{x_0} c_+\sqrt{t}\delta_+(t)dt - \int_0^{x_0} c_-\sqrt{t}\delta_-(t)dt$$
$$-\int_0^{x_0} G(t)f'(t)dt \leq -\int_0^{x_0} tf'(t)dt + \int_0^{x_0} c_-\sqrt{t}\delta_+(t)dt + \int_0^{x_0} c_+\sqrt{t}\delta_-(t)dt,$$

and, similarly,

$$-\int_{x_0}^{\infty} G(t)f'(t)dt \geq -\int_{x_0}^{\infty} tf'(t)dt - \int_{x_0}^{\infty} \epsilon_+ t\delta_+(t)dt - \int_{x_0}^{\infty} \epsilon_- t\delta_-(t)dt$$
$$-\int_{x_0}^{\infty} G(t)f'(t)dt \leq -\int_{x_0}^{\infty} tf'(t)dt + \int_{x_0}^{\infty} \epsilon_- t\delta_+(t)dt + \int_{x_0}^{\infty} \epsilon_+ t\delta_-(t)dt.$$

We conclude that (5.28) holds. By partial summation and $f(t) = o(1/t)$,

$$-\int_0^{\infty} tf'(t)dt = \int_0^{\infty} f(t)dt,$$

and thus we obtain (5.29). □

Here is a simple application of Proposition 5.4.

Corollary 5.5.

$$\sum_{n\leq x} \Lambda(n)n \leq 1.24016 \cdot \frac{x^2}{2} \qquad \textit{for } x > 0, \tag{5.31}$$

$$\sum_{n\leq x} \Lambda(n)n \leq 1.00302 \cdot \frac{x^2}{2} \qquad \textit{for } x \geq 10^5. \tag{5.32}$$

Proof. Apply Proposition 5.4 with $g(n) = \Lambda(n)$, $c_- = c_+ = \sqrt{2}$, $\epsilon_- = \epsilon_+ = 0.000213$, $x_0 = 10^{10}$ and

$$f(t) = \begin{cases} t & \text{if } t \leq x, \\ 0 & \text{otherwise.} \end{cases}$$

Condition (5.27) holds by (5.12) and (5.15). We obtain, for $x \leq 10^{10}$,

$$\begin{aligned}\sum_{n\leq x} \Lambda(n)n &= \int_0^x tdt + O^*\left(\sqrt{2}\left(\int_0^x \sqrt{t}dt + \sqrt{x}\cdot x\right)\right) \\ &= \frac{x^2}{2} + O^*\left(\frac{5\sqrt{2}}{3}x^{3/2}\right),\end{aligned} \tag{5.33}$$

where the term $\sqrt{x}\cdot x$ is present because – as should be clear – in the statement of Prop. 5.4, f' is to be understood as a distribution, and thus $\delta_-(t)$ consists of a point mass at $t = x$ and is 0 elsewhere. Both $\delta_-(t)$ and $\delta_+(t)$ vanish for $t > x$; since $x \leq 10^{10}$, they vanish for all $t > x_0 = 10^{10}$. For $x > 10^{10}$,

$$\begin{aligned}\sum_{n\leq x} \Lambda(n)n &\leq \int_0^x t dt + \sqrt{2}\int_0^{10^{10}} \sqrt{t}dt + \epsilon\left(\int_{10^{10}}^x t dt + x^2\right)\\ &= \frac{x^2}{2} + \sqrt{2}\cdot\frac{2}{3}10^{15} + \epsilon\frac{x^2}{2} - \epsilon\frac{10^{20}}{2} + \epsilon x^2\\ &= (1+3\epsilon)\frac{x^2}{2} + \left(\sqrt{2}\cdot\frac{2}{3}10^{-5} - \epsilon\right)\cdot\frac{10^{20}}{2} \leq (1+3\epsilon)\frac{x^2}{2},\end{aligned} \tag{5.34}$$

where $\epsilon = \epsilon_- = \epsilon_+ = 0.000213$.

For $x > 10^{10}$, it is clear that (5.34) is stronger than both (5.31) and (5.32). Inequality (5.33) implies (5.31) for $400 \leq x \leq 10^{10}$ and (5.32) for $2.5\cdot 10^6 \leq x \leq 10^{10}$. For smaller values of x, we verify (5.31) and (5.32) by brute force. (They are essentially tight for $x = 5$ and $x = 102679$, respectively.) □

We will sometimes have to work with functions that may have a sharp cutoff at some $t_0 > 0$, in the sense that $\eta(t) = 0$ for $t < t_0$ and $\lim_{t\to t_0^+}\eta(t)$ is not necessarily zero.

Corollary 5.6 (to Prop. 5.4)**.** *Let $t_0 \geq 0$. Let $\eta : \mathbb{R}\to\mathbb{R}$ be a function of bounded variation satisfying $\eta(t) = 0$ for $t \leq t_0$. Assume as well that $\eta(t) = O(1/t)$ as $t\to\infty$, that $\eta(t)$ and $t\eta'(t)$ are in L^1 and that $\int_0^\infty \eta(t)dt \geq 0$.*

Let $x_0 > 0$. Let $g : \mathbb{Z}^+ \to \mathbb{Z}$ satisfy

$$(1-\epsilon_-)x \leq \sum_{n\leq x} g(n) \leq (1+\epsilon_+)x$$

for all $x \geq x_0$, where $\epsilon_-, \epsilon_+ \geq 0$.

Then, for any $x > 0$ such that $t_0 x \geq x_0$,

$$\begin{aligned}\sum_n g(n)\eta\left(\frac{n}{x}\right) &\geq (1-\epsilon_-)x\int_0^\infty \eta(t)dt - (\epsilon_+ + \epsilon_-)x\int_0^\infty t\eta_+(t)dt\\ \sum_n g(n)\eta\left(\frac{n}{x}\right) &\leq (1+\epsilon_+)x\int_0^\infty \eta(t)dt + (\epsilon_+ + \epsilon_-)x\int_0^\infty t\eta_+(t)dt,\end{aligned} \tag{5.35}$$

where $\eta_+(t) = \max(\eta'(t), 0)$.

In particular, for any $x > 0$ such that $t_0 x \geq 10^5$,

$$\begin{aligned}\sum_p (\log p)\eta\left(\frac{n}{x}\right) &\geq \left(0.995268\int_0^\infty \eta(t)dt - 0.004733\int_0^\infty t\eta_+(t)dt\right)x,\\ \sum_p (\log p)\eta\left(\frac{n}{x}\right) &\leq \left((1+7.5\cdot 10^{-7})\int_0^\infty \eta(t)dt + 0.004733\int_0^\infty t\eta_+(t)dt\right)x.\end{aligned} \tag{5.36}$$

Here, as always, we have to remember to count the contribution of discontinuities to norms such as $|t\eta'(t)|_1$. For instance, if $\lim_{t\to t_0^+}\eta(t) = r \neq 0$, then the discontinuity at t_0 contributes $t_0|r|$. If $\lim_{t\to t_0^+}\eta(t) = r$ and $\eta(t) \geq 0$ for all t, then (5.36) simplifies to

$$\begin{aligned}\sum_p (\log p)\eta\left(\frac{n}{x}\right) &\geq (0.995268|\eta|_1 - 0.004733 t_0 r)\cdot x\\ \sum_p (\log p)\eta\left(\frac{n}{x}\right) &\leq ((1+7.5\cdot 10^{-7})|\eta|_1 + 0.004733 t_0 r)\cdot x.\end{aligned}\tag{5.37}$$

Proof. Apply Proposition 5.4 with $f(t) = \eta(t/x)$. Since $f(t) = 0$ for every $t \leq x_0$, whatever value we choose for c_-, c_+ is irrelevant: $\delta_+(t) = \delta_-(t) = 0$ for $t \leq x_0$. We obtain from (5.28) that

$$\begin{aligned}\sum_n g(n)\eta\left(\frac{n}{x}\right) &\geq |\eta|_1 x - \epsilon_- \int_{x_0}^{\infty} t\delta_-(t)dt - \epsilon_+ \int_{x_0}^{\infty} t\delta_+(t)dt\\ \sum_n g(n)\eta\left(\frac{n}{x}\right) &\leq |\eta|_1 x + \epsilon_- \int_{x_0}^{\infty} t\delta_+(t)dt + \epsilon_+ \int_{x_0}^{\infty} t\delta_-(t)dt.\end{aligned}\tag{5.38}$$

We can, of course, extend the integrals here to range from 0 to ∞, since $\delta_-(t) = \delta_+(t) = 0$ for $t < x_0$. Now,

$$\begin{aligned}\int_0^{\infty} t\delta_+(t)dt - \int_0^{\infty} t\delta_-(t)dt &= \int_0^{\infty} t\eta'(t)dt\\ &= t\eta(t)|_0^{\infty} - \int_0^{\infty}\eta(t)dt = -\int_0^{\infty}\eta(t)dt,\end{aligned}$$

and so

$$\begin{aligned}\epsilon_- \int_0^{\infty} t\delta_-(t)dt + \epsilon_+ \int_0^{\infty} t\delta_+(t)dt &= (\epsilon_+ + \epsilon_-)\int_0^{\infty} t\delta_+(t)dt + \epsilon_- \int_0^{\infty}\eta(t)dt,\\ \epsilon_- \int_0^{\infty} t\delta_+(t)dt + \epsilon_+ \int_0^{\infty} t\delta_-(t)dt &= (\epsilon_+ + \epsilon_-)\int_0^{\infty} t\delta_+(t)dt + \epsilon_+ \int_0^{\infty}\eta(t)dt.\end{aligned}$$

We conclude that (5.35) holds. We apply the bounds (5.17) and (5.20), and obtain (5.36). □

We will of course apply Prop. 5.4 and Cor. 5.6 to specific smoothing functions – besides the one in Cor. 5.5 – but we defer these applications to Appendix B.1.

Lemma 5.7. *For $x \geq \sqrt{2}$,*

$$\sum_{n\leq x}\frac{\Lambda(n)}{n} \leq \log\frac{x}{\sqrt{2}}.\tag{5.39}$$

For $x \geq 3$,

$$\sum_{n\leq x}\frac{\Lambda(n)}{n} \leq \log\frac{x}{3^{2/3}/\sqrt{2}}.\tag{5.40}$$

For $x \geq 1$*,*

$$\sum_{n\leq x} \frac{\Lambda(n)}{n} \geq \log x - \log \frac{3}{\sqrt{2}}. \tag{5.41}$$

For $x_1 \geq x_0 \geq 1$*,*

$$\sum_{x_0<n\leq x_1} \frac{\Lambda(n)}{n} \leq \log \frac{x_1}{x_0} + \frac{\log 3}{3} \tag{5.42}$$

and

$$\sum_{x_0<n\leq x_1} \frac{\Lambda(n)}{n} \log \frac{n}{x_0} \leq \left(\frac{1}{2}\log \frac{x_1}{x_0} + \frac{\log 3}{3}\right) \log \frac{x_1}{x_0}. \tag{5.43}$$

Proof. By [Ram13a, Cor. to Thm 1.1],

$$\sum_{n\leq x} \frac{\Lambda(n)}{n} = \log x - \gamma + O^*\left(\frac{0.0067}{\log x}\right)$$

for $x \geq 23$, where γ is Euler's constant $\gamma = 0.577215\ldots$. Since $\gamma - \log(\sqrt{2}) > \gamma - \log(3^{2/3}/\sqrt{2}) = 0.19138\ldots$ and $\log(3/\sqrt{2}) - \gamma = 0.17482\ldots$, we obtain (5.39), (5.40) and (5.41) immediately for $x \geq 23$. It is easy to check (5.39) and (5.40) numerically for $\sqrt{2} \leq x < 23$ and $3 \leq x < 23$, respectively; it is clearly enough to check integer values of x. It is also easy to check (5.41) for $1 \leq x < 23$; the worst case here is that of x tending to an integer n from the left.

Inequality (5.42) follows immediately from (5.40) and (5.41) for $x_1 \geq 3$. If $x_1 < 3$, the worst case is that of $x_0 \to 2^-$, $x_1 \to 2^+$; in that case, the difference $(\sum_{x_0<n\leq x_1} \Lambda(n)/n) - \log x_1/x_0$ tends to $(\log 2)/2$, which is less than $(\log 3)/3$.

Finally, applying (5.42), we see that

$$\begin{aligned}\sum_{x_0<n\leq x_1} \frac{\Lambda(n)}{n} \log \frac{n}{x_0} &= \sum_{x_0<n\leq x_1} \frac{\Lambda(n)}{n} \int_{x_0}^{n^-} \frac{dt}{t} = \int_{x_0}^{x_1} \sum_{t<n\leq x_1} \frac{\Lambda(n)}{n} \frac{dt}{t} \\ &\leq \int_{x_0}^{x_1} \log\left(\frac{x_1}{t} + \frac{\log 3}{3}\right) \frac{dt}{t} \\ &= \frac{1}{2}\log^2 \frac{x_1}{x_0} + \frac{\log 3}{3} \log \frac{x_1}{x_0}.\end{aligned}$$

□

5.3 SUMS OF μ

Now let us see some estimates on sums involving μ. We care particularly about the following sums:

$$m(x) = \sum_{n\leq x} \frac{\mu(n)}{n}, \quad \check{m}(x) = \sum_{n\leq x} \frac{\mu(n)}{n} \log \frac{x}{n}, \quad \check{\check{m}}(x) = \sum_{n\leq x} \frac{\mu(n)}{n} \left(\log \frac{x}{n}\right)^2, \tag{5.44}$$

and, more generally,

$$\begin{aligned} m_q(x) &= \sum_{\substack{n\leq x\\(n,q)=1}} \frac{\mu(n)}{n}, \qquad \check{m}_q(x) = \sum_{\substack{n\leq x\\(n,q)=1}} \frac{\mu(n)}{n}\log\frac{x}{n}, \\ \check{\check{m}}_q(x) &= \sum_{\substack{n\leq x\\(n,q)=1}} \frac{\mu(n)}{n}\left(\log\frac{x}{n}\right)^2 \end{aligned} \tag{5.45}$$

for $q \in \mathbb{Z}^+$.

We might say that one of the reasons why the smoothed sums $\check{m}(x)$, $\check{\check{m}}(x)$ are natural is that, since $\sum_n \mu(n)n^{-s} = 1/\zeta(s)$,

$$\sum_n \mu(n)(\log n)n^{-s} = -\left(\frac{1}{\zeta(s)}\right)', \qquad \sum_n \mu(n)(\log n)^2 n^{-s} = \left(\frac{1}{\zeta(s)}\right)''.$$

At the same time, we will not be using these identities directly, and even if we see them as underlying reasons, we could overstate their importance. The central fact is that $x \mapsto \log^+(x/n)$ is continuous, and $x \mapsto (\log^+(x/n))^2$ is differentiable, even at $x = n$. Other smoothing functions would also lead to bounds sharper than those for $m(x)$, generally speaking.

5.3.1 Bounds on sums of μ for large x

We expect the sums $m(x)$, $\check{m}(x)$ and $\check{\check{m}}(x)$ to have square-root cancellation, or nearly so. The Riemann Hypothesis implies that the closely related sum

$$M(x) = \sum_{n\leq x} \mu(n) \qquad \text{(Mertens's function)}$$

satisfies $M(x) = O_\epsilon\left(x^{1/2+\epsilon}\right)$. (See [Sou09], [BdR] for sharper estimates.) The two sums $m(x)$ and $M(x)$ are linked by partial summation and also by relations stemming from Möbius inversion; see [Axe10] and [Bal12a]. As a result, one obtains $m(x) = O_\epsilon\left(x^{-1/2+\epsilon}\right)$. Analogously, one can show that $\check{m}(x) = 1+O_\epsilon\left(x^{-1/2+\epsilon}\right)$ and $\check{\check{m}}(x) = 2\log x - 2\gamma + O_\epsilon\left(x^{-1/2+\epsilon}\right)$, again conditionally on the Riemann Hypothesis. See [Axe10, Lemmas 3.2 and 5.1].

(Interestingly, as we shall later discuss, $m(x)$ is in all likelihood *not* $O\left(x^{-1/2}\right)$, but the error terms for $\check{m}(x)$ and $\check{\check{m}}(x)$ probably *are* $O\left(x^{-1/2}\right)$. This is not extremely surprising: $\log^+ x/n$ acts as a smoothing, thanks to the fact that it is continuous at $x = n$.)

Of course, we need unconditional results – indeed, effective and explicit results. The situation here is even less satisfactory than for sums involving Λ. The main reason is that the standard complex-analytic approach to estimating sums of μ would involve $1/\zeta(s)$ rather than $\zeta'(s)/\zeta(s)$, and thus we would need explicit bounds either on the residues of $1/\zeta(s)$ or on the size of $1/\zeta(s)$ within a zero-free region. We do have the latter kind of bound, but the constants are large; see (3.65). In consequence, explicit estimates on sums involving μ are harder to obtain than estimates on sums involving Λ. Such is the situation even though analytic number theorists are generally used (from the habit of non-explicit work) to see the estimation of one kind of sum or the other as essentially the same task.

The overall situation is surveyed in [Rama]. Fortunately, all we will need is a saving of a constant times a power of $\log x$ (often simply $\log x$) over the trivial bound. Of course, we also care about the constant.

For $m(x)$, in occasions when we need to save only a factor of $\log x$, and do not care too much about the constant, we can use the following inequality by Granville and Ramaré [GR96, Lemma 10.2]:

$$|m_q(x)| \leq 1 \tag{5.46}$$

for all $x > 0$, $q \in \mathbb{Z}^+$. This elegant bound has the merit of being completely independent of q.

To show (5.46), we first establish the classical identity in [Mei54, p. 303]:

$$\begin{aligned} \sum_{n \leq x} \mu(n) \left\lfloor \frac{x}{n} \right\rfloor &= \sum_{n \leq x} \mu(n) \sum_{\substack{m \leq x \\ n|m}} 1 \\ &= \sum_{m \leq x} \sum_{n|m} \mu(n) = 1. \end{aligned} \tag{5.47}$$

This identity was already used by Gram [Gra84, p. 197–198] to prove (5.46) in the special case $q = 1$: assuming, as we may, that x is an integer,

$$\begin{aligned} \sum_{n \leq x} \frac{\mu(n)}{n} &= \frac{1}{x} \left(\sum_{n \leq x} \mu(n) \left\lfloor \frac{x}{n} \right\rfloor + \sum_{n \leq x} \mu(n) \left\{ \frac{x}{n} \right\} \right) \\ &= \frac{1}{x} \left(1 + \sum_{1 < n \leq x} O^*(1) \right) = O^*(1). \end{aligned}$$

We can easily prove a generalization of (5.47), namely,

$$\sum_{\substack{n \leq x \\ (n,q)=1}} \mu(n) \left\lfloor \frac{x}{n} \right\rfloor = \sum_{m \leq x} \sum_{\substack{n|m \\ (n,q)=1}} \mu(n) = \sum_{\substack{m \leq x \\ m|q^\infty}} 1.$$

We now obtain (5.46) for q arbitrary:

$$\begin{aligned}\sum_{\substack{n\leq x\\(n,q)=1}} \frac{\mu(n)}{n} &= \frac{1}{x}\left(\sum_{\substack{n\leq x\\(n,q)=1}} \mu(n)\left\lfloor\frac{x}{n}\right\rfloor + \sum_{\substack{n\leq x\\(n,q)=1}} \mu(n)\left\{\frac{x}{n}\right\}\right)\\ &= \frac{1}{x}\left(\sum_{\substack{m\leq x\\m|q^\infty}} 1 + \sum_{\substack{1<m\leq x\\(m,q)=1}} O^*(1)\right) = O^*(1),\end{aligned}$$

where we assume again without loss of generality that x is an integer.

In general, we want bounds stronger than (5.46). Let us look for the moment at results with $q = 1$. Here the best results in the literature to date are those in [Ram15]; some older results along the same lines can be found in [DE93], [EM95], [EM96] and [Ram13b].

By [Ram15, Thm. 1.2],

$$|m(x)| \leq \frac{0.0144}{\log x} \tag{5.48}$$

for $x \geq 96955$. By [Ram15, Thm. 1.5], [2]

$$|\check{m}(x) - 1| \leq \frac{1}{389\log x} \tag{5.49}$$

for $x \geq 3155$. For $x \geq 9$, by [Ram15, Thm. 1.8],

$$\left|\check{\check{m}}(x) - 2\log x + 2\gamma\right| \leq \frac{1}{103\log x}. \tag{5.50}$$

The point of departure here is the bound on $|\psi(x) - x|$ in (5.13). This bound can be carefully manipulated by means of combinatorial identities so as to yield a good estimate for $M(x)$, via an estimate for $\sum_{n\leq x}\mu(n)\log n$ and a recursion. The basic procedure is already to be found in [Sch69]. The realization that estimates on $\psi(x)$ imply estimates on $M(x)$ in an elementary way is of course classical; see, e.g., [Lan06], based in part to even older work.

Careful refinements of this procedure give the following bounds on $M(x)$:

$$|M(x)| \leq \frac{x}{4345} \tag{5.51}$$

[2] In a corrigendum [Ram19], Ramaré gives, among other results, the bound $|\check{m}(x) - 1| \leq 1/396\log x$, valid for $x \geq 3162$. This bound clearly implies (5.49) for $x \geq 3162$; for $3155 \leq x < 3162$, we verify (5.49) by direct computation (see 5.3.2). This is as good a place as any to remark that bounds of the form we are considering (in (5.48)– (5.52), say) are reasonably resistant to human error (most likely deliberately so): they can be verified computationally for small x, and are not tight for large x; thus, while they are tight in the sense they cannot be sensibly improved without increasing the least x for which they hold, they would most likely be unaffected if (*absit omen*) a slip in their proof were to be found.

for $x \geq 2160535$ ([CDE07]; available as a preprint since 1996) and

$$|M(x)| \leq \frac{0.0130 \log x - 0.118}{(\log x)^2} x \tag{5.52}$$

for $x \geq 1078853$ [Ram13b, Thm 1.1].

One could of course go from such estimates on $M(x)$ to estimates on $m(x)$ just by Abel summation: by (3.12),

$$m(x) = \sum_{n \leq x} \mu(n) \cdot \frac{1}{n} = \frac{M(x)}{x} + \int_1^x \frac{M(u)}{u^2} du.$$

However, that would involve a loss of a factor of log. The fact that there is a way to reduce this loss is not at all obvious. In a qualitative form, it can be found in Axer [Axe10]. The proof of the bound in (5.48) is based on an identity in [Bal12b], which yields the bound

$$\left| m(x) - \frac{M(x)}{x} \right| \leq \frac{1}{x^2} \int_1^x |M(u)| du + \frac{8}{3x},$$

valid for all $x \geq 1$ [Bal12b, Prop. 7]. This bound is proved by an ingenious variant of an idea underlying the proof of (5.46) above: instead of the identity (5.47), Balazard uses what may be called a higher-degree generalization thereof [Mac94].

Bounds on $M(x)$ involving higher powers of $\log x$ than (5.51)–(5.52) are known [EM95], and can be used to derive analogous bounds on $m(x)$. However, the constants involved are such as to make those bounds not yet practical.

5.3.2 Bounds on sums of μ for x bounded

For relatively small x, we can simply compute $m(x)$, $\check{m}(x)$ and $\check{\check{m}}(x)$.

Lemma 5.8. *Let $m(x)$, $\check{m}(x)$ and $\check{\check{m}}(x)$ be as in (5.44). We can compute $m(x)$, $\check{m}(x)$ and $\check{\check{m}}(x)$ for $x \leq x_0$ in time $O(x_0 \log \log x_0)$ and space $O(\sqrt{x_0})$.*

Proof. We compute $\mu(n)$ in time $O(x_0 \log \log x_0)$ and space $O(\sqrt{x_0})$ by a segmented sieve of Eratosthenes, applying Algorithm 3 in §4.5 to consecutive subintervals ("segments") of $[1, x_0]$ of length $\asymp \sqrt{x_0}$. We keep track of $m(N)$ for $N \leq x_0$ while the computation of $\mu(n)$ for successive $\mu(n)$ is ongoing. We compare $\sqrt{N+1} \cdot m(N)$ to the bound on $\sqrt{x} \cdot m(x)$ for $x < N$; if it is greater, it becomes the new bound.

We proceed similarly for $\check{m}(x)$. There is a common pitfall to avoid, though. We ought *not* to express $\check{m}(x)$ as

$$\check{m}(x) = (\log x) m(x) - \sum_{n \leq N} \frac{\mu(n)}{n} \log n$$

if we mean to compute $\check{m}(x)$: expressing something as a difference between two quantities that tend to be considerably larger is a recipe for losing precision quickly. Rather, we should use the fact that

$$\check{m}(x) = \check{m}(N) + m(N) \cdot \log \frac{x}{N} \tag{5.53}$$

for $N \leq x \leq N+1$. We use that same identity to compute $\check{m}(N+1)$ given $\check{m}(N)$ and $m(N)$.

It almost goes without saying that what we must actually store and update at each step is $\check{m}(N) - 1$, and *not* $\check{m}(N)$. Otherwise, we lose precision very severely, for the same reason as before. It is helpful, also for the sake of precision, to use an implementation of the function $t \mapsto \log(1+t)$ – based on a Taylor series around 1 – rather than use $\log$ directly.

Similarly, for $N \leq x \leq N+1$,

$$\check{\check{m}}(x) = \check{\check{m}}(N) + 2\check{m}(N) \cdot \log \frac{x}{N} + m(N) \left(\log \frac{x}{N}\right)^2,$$

and so

$$\begin{aligned}\check{\check{m}}(x) - 2\log x + 2\gamma = \check{\check{m}}(N) &- 2\log N + 2\gamma \\ &+ 2\left(\check{m}(N) - 1\right) \cdot \log \frac{x}{N} + m(N) \left(\log \frac{x}{N}\right)^2.\end{aligned}$$

Again, we store and update $\check{\check{m}}(N) - 2\log N + 2\gamma$, rather than $\check{\check{m}}(N)$: we can see from (5.50) that we should regard $2\log N - 2\gamma$ as the main term of $\check{\check{m}}(N)$. □

Now we simply have to compute $m(x)$, $\check{m}(x)$ and $\check{\check{m}}(x)$ for all x up to a reasonable x_0, and verify that they satisfy good bounds. As always, we will keep matters strictly rigorous by means of interval arithmetic.

The computation that proves the following lemma is trivial, in the sense of being doable in a weekend on an ordinary desktop computer. The bottlenecks are space and accuracy, more than the theoretical running time. (Space usage affects running time in practice; see §4.5.) Our computation, in double-precision interval arithmetic, gives us that

$$-8.72884 \cdot 10^{-9} \leq \check{m}\left(10^{12} - 1\right) \leq -8.61050 \cdot 10^{-9}.$$

We can clearly see a decrease in precision at this point. Going much further would involve using higher than double precision, making matters much less efficient on current hardware.[3]

Lemma 5.9. *Let $\check{m}(x)$ and $\check{\check{m}}(x)$ be as in (5.44). Then*

$$|\check{m}(x) - 1| \leq \begin{cases} \frac{0.0234188}{\sqrt{x}} & \textit{for all } 11 \leq x \leq 10^{12} \\ \frac{1}{\sqrt{x}} & \textit{for all } 0 < x \leq 10^{12}. \end{cases} \tag{5.54}$$

[3] There is an IEEE standard for quadruple precision, but it has not been implemented in hardware yet (as of 2019).

and

$$|\check{m}(x) - 1| \leq \frac{1}{x} + \frac{0.0199824}{\sqrt{x}} \quad \textit{for all } 0 < x \leq 10^{12}. \tag{5.55}$$

Moreover,

$$|\check{\check{m}}(x) - 2\log x + 2\gamma| \leq \begin{cases} \frac{2\gamma}{\sqrt{x}} & \textit{for } 1 \leq x \leq 10^{12}, \\ \frac{4e^{\frac{\gamma}{2}-1}}{\sqrt{x}} & \textit{for } 0 < x \leq 10^{12}, \end{cases} \tag{5.56}$$

$$|\check{\check{m}}(x) - 2\log x + 2\gamma| \leq \frac{2\gamma}{x} + \frac{0.00232347}{\sqrt{x}} \quad \textit{for all } 1 \leq x \leq 10^{12}. \tag{5.57}$$

where γ is the Euler-Mascheroni constant.

The constants here may not be tight, due to accumulated errors. Incidentally, bounds like (5.54) were proven in [Ram15] for x in a smaller range. In particular, [Ram15, Lem. 10.2] states that $|\check{m}(x) - 1| \leq 0.0218/\sqrt{x}$ for $11 \leq x \leq 1.2 \cdot 10^7$.

For us, the terms proportional to $1/x$ in (5.55) and (5.57) are just convenient ways to account for the effects of x small; an actual lower-order term to be got from analysis (assuming various conjectures) might look different.

Proof. Compute $\check{m}(x)$ and $\check{\check{m}}(x)$ as in the proof of Lemma 5.8. For each N, we check whether

$$\sqrt{N+1} \cdot \left| \check{m}(N) - 1 + m(N) \cdot \left[0, \log\left(1 + \frac{1}{N}\right)\right]\right| \tag{5.58}$$

contains elements greater than the current bound by means of interval arithmetic. We do the same for $\sqrt{N+1}$ times the absolute value of

$$\check{\check{m}}(N) - 2\log N + 2\gamma + 2\left(\check{m}(N) - 1\right) \cdot \left[0, \log\left(1 + \frac{1}{N}\right)\right] \\ + m(N) \cdot \left[0, \log^2\left(1 + \frac{1}{N}\right)\right].$$

We have to check the lowest ranges separately, as follows.[4] First, $|\check{m}(x) - 1|$ equals 1 for $0 \leq x \leq 1$ and $1 - \log x$ for $1 \leq x \leq 2$. For all $1 \leq x \leq 2$, we know that $1 - \log x \leq 1/x$, since equality holds for $x = 1$ and the derivative of the left side is bounded from above by the derivative of the right side for $x \geq 1$. We conclude that (5.55) and the second inequality in (5.54) hold for all $0 \leq x \leq 2$.

To check (5.56) and (5.57) for $1 \leq x \leq 2$, note first that $(\log x)^2 - 2\log x + 2\gamma$ is positive (since it is positive for $x = 2$ and its derivative $2(-1 + \log x)/x$ is negative for $1 \leq x \leq 2$). We can see that $((\log x)^2 - 2\log x + 2\gamma)x$ is $\leq 2\gamma$ for $1 \leq x \leq 2$, since it equals 2γ for $x = 1$ and its derivative $2\gamma + (\log x)^2 - 2$ is negative for $1 \leq x \leq 2$.

Lastly, $\check{\check{m}}(x) = 0$ for $x \leq 1$, and a derivative test shows that $\sqrt{x} \cdot (-2\log x + 2\gamma)$ reaches its maximum $4e^{\gamma/2-1}$ at $x = e^{\gamma-2}$. □

[4] These verifications are of the kind that a computer-algebra package ought to do automatically, with next to no human intervention. Add this desideratum to the list in §4.1.

It has been conjectured by Hejhal [Hej89] and then by Gonek [Gon89] that

$$\sum_{0<\Im\rho\leq T} \frac{1}{|\zeta'(\rho)|} \ll T(\log T)^{1/4}, \tag{5.59}$$

where ρ ranges over zeros of $\zeta(s)$. If this conjecture is true, then[5]

$$|\check{m}(x)| < \frac{c}{\sqrt{x}} \tag{5.60}$$

for all $x > 0$ and some $c > 0$. In fact, it would be enough to have a much weaker bound $\ll T^{2-\epsilon}$ in place of the right side of (5.59) to reach the same conclusion (5.60).

This situation contrasts with the situation for $m(x)$. It is known to be false that $|M(x)| \leq \sqrt{x}$ for all x (Mertens's conjecture, disproved in [OtR85]), it is generally believed that $\limsup_{x\to\infty} |M(x)|/\sqrt{x} = \infty$ (see [OtR85] and references therein), and, by partial summation ($M(N) = Nm(N) - \sum_{n\leq N-1} m(n)$), that implies that

$$\limsup_{x\to\infty} \sqrt{x}|m(x)| = \infty.$$

In practice, though, $\sqrt{x}|m(x)|$ stays bounded for x within a reasonable range, though, as we will see, the bound is noticeably larger than for $\sqrt{x}|\check{m}(x)|$ or $\sqrt{x}|\check{\check{m}}(x)|$. The reason why we are about to treat the computation $\sqrt{x}|m(x)|$ separately is the following. Computing $m(x)$ involves fewer floating-point operations than computing $\check{m}(x)$ or $\check{\check{m}}(x)$. Hence, when we compute $m(x)$, accuracy decreases more slowly – and it is also the case that sieving, and not floating-point operations, tends to take most of the running time. This means that it is worthwhile to take a higher x, and to optimize the sieving process more than we did above.

For the sake of verification, we record that the computation below gives us that

$$-8.02622 \cdot 10^{-9} \leq m\left(10^{14}\right) \leq -8.01187 \cdot 10^{-9}. \tag{5.61}$$

Lemma 5.10. *Let $m(x)$ be as in (5.44). Then*

$$|m(x)| \leq \begin{cases} \frac{0.569449}{\sqrt{x}} & \text{for } 3 \leq x \leq 10^{14}, \\ \sqrt{2/x} & \text{for } 0 < x \leq 10^{14}. \end{cases} \tag{5.62}$$

Ramaré [Ram14] showed that $|m(x)| \leq 1/2\sqrt{x}$ holds for all $1 \leq x \leq 10^{10}$. It turns out that that stronger bound holds for all $3 \leq x \leq 7727068587$, but not for $x = 7727068588 - \epsilon$. Interestingly, the smallest integer $n > 200$ such that $M(n) > \sqrt{n}/2$ is 7725030629 [CD88].

The computation used to prove Lemma 5.10 is on the higher end of "homebrew": the program takes about two weeks when run on 8- or 12-core servers from the early 2010s. Interval arithmetic is of course used throughout.

[5] Thanks are due both to a pseudonymous MathOverflow contributor and to P. Humphries for this observation.

Proof. We compute $m(x)$ as in the proof of Lemma 5.8, optimizing for memory, and also parallelizing, as described in §4.5. □

It would actually have been enough for our purposes to compute a bound for $x \leq 10^{12}$, as elsewhere. It was simply interesting to see what could be done within the constraints of fairly casual programming.

On a different note: there is an algorithm for computing $M(x)$ for given x in time $O(x^{2/3}(\log\log x)^{1/3})$ [DR96], based on a combinatorial identity due to Lehman. (A variant of Vaughan's or Heath-Brown's identity would also do). It would be interesting to adapt it to compute $m(x)$ or $\check{m}(x)$. There is also an algorithm for computing $\pi(x)$ in time $O(x^{1/2+\epsilon})$ using integrals involving $\zeta'(s)/\zeta(s)$ in the complex plane ([LO87], first implemented in [Pla15]). Adapting it to compute $M(x)$, $m(x)$ or $\check{m}$ would be more challenging, and possibly unfeasible in a practical sense, due to the same issues involving residues of $1/\zeta(s)$ that make a direct complex-analytic approach to estimating $M(x)$, $m(x)$ or $\check{m}$ for large x rather daunting, indeed seemingly impracticable for the moment. (Recall our discussion at the beginning of §5.3.1.)

If one could find an algorithm to compute $m(x)$ for individual x in time less than $\sqrt{x}$, one would have an algorithm that verifies an inequality of the form $m(x) \leq C/\sqrt{x}$ for all $x \leq x_0$ in less than linear time, bypassing the need for a sieve entirely. The same goes for computing $M(x)$: an algorithm running in time less than $\sqrt{x}$ might be useful to help find the first x for which Mertens's conjecture fails. (See [KvdL04] on complementary approaches.) No such algorithm is known at the time of writing, either for $M(x)$ or for $m(x)$.

At any rate, Lemmas 5.9 and 5.10 are more than enough for our purposes, so we proceed.

5.3.3 Convexity. Bounds on sums of μ for all x

We have discussed bounds on $m(x)$, $\check{m}(x)$ and $\check{\check{m}}(x)$ for small x and for large x. We would like to put these bounds in a form that is valid for all x. This will show its usefulness shortly, when we derive bounds for $m_q(x)$, $\check{m}_q(x)$, etc.

As will later become clear, it is particularly useful to have bounds on $m(t)$ and $\check{m}(t) - 1$ that are linear combinations of powers $t^{-\beta}$, $\beta > 0$. This would seem at first sight to be impossible: for t large, the bounds we have are proportional to $1/\log t$. Let us see what we will do.

Lemma 5.11. *Let $x \geq x_0 > 0$. Then, for all $x_0 \leq t \leq x$,*

$$\frac{1}{\log t} \leq \frac{x^\beta}{\log x}\frac{1}{t^\beta}, \tag{5.63}$$

where $\beta = 1/\log x_0$.

Old hands at the estimation of sums in analytic number theory will see that, while the trick involved here isn't really the same as what is often called "Rankin's trick", it resembles it, in that it is a simple tool for reducing the task of bounding a sum to the task of bounding a series. It is nothing more mysterious than bounding a convex function by a linear function, after a change of variables.

Proof. Let $x_0 \leq t \leq x$. Obviously, $1/\log t \leq (1/\log x) \cdot (\log x/\log t)$. Now, the function log is concave, and so, since $x_0 < t \leq x$,

$$\frac{\log\log x - \log\log t}{\log x - \log t} \leq (\log s)'|_{s=\log x_0} = \frac{1}{\log x_0}.$$

Hence $\log x/\log t \leq (x/t)^{1/\log x_0}$. We conclude that $1/\log t \leq (1/\log x)(x/t)^{1/\log x_0}$. □

We can now derive a bound on $m(t)$ valid for all t.

Lemma 5.12. *For any $y > 1$, $0 < t \leq y$,*

$$|m(t)| \leq \left(\frac{2}{t}\right)^{1/2} + 0.0144 \frac{y^{\frac{1}{\log 10^{14}}}}{\log y} \frac{1}{t^{\frac{1}{\log 10^{14}}}}. \tag{5.64}$$

The constant 2 could be improved very slightly.

Proof. By (5.48) and Lemma 5.11 (with $x_0 = C$),

$$|m(t)| \leq 0.0144 \frac{y^{1/\log C}}{\log y} \frac{1}{t^{1/\log C}} \tag{5.65}$$

for all $C \leq t \leq y$, where $C \geq 96955$. We set $C = 10^{14}$, use (5.62) and are done. □

We can sometimes push the first exponent in expressions such as (5.64) all the way from $1/2$ to 1 by being slightly clever about it. We will not bother to do it for $m(t)$, but we will do it for $\check{m}(t)$ (and later for $\check{m}_2(t)$).

Lemma 5.13. *Let $3155 \leq C \leq 10^{12}$. For any $y \geq C$, $0 < t \leq y$ and any $\sigma \in [0, 1]$,*

$$|\check{m}(t) - 1| \leq \frac{1}{t^\sigma} + \frac{1}{389} \frac{y^{\frac{1}{\log C}}}{\log y} \frac{1}{t^{\frac{1}{\log C}}}. \tag{5.66}$$

For any $y > 1$, $0 < t \leq y$,

$$|\check{m}(t) - 1| \leq \frac{1}{\sqrt{t}} + \frac{1}{389} \frac{y^{\frac{1}{\log 10^{12}}}}{\log y} \frac{1}{t^{\frac{1}{\log 10^{12}}}}. \tag{5.67}$$

Here $1/t^\sigma$ can actually be improved to $0.999748/t^\sigma$.

Proof. By (5.49) and Lemma 5.11 (with $x_0 = C$),

$$|\check{m}(t) - 1| \leq \frac{1}{389} \frac{y^{1/\log C}}{\log y} \frac{1}{t^{1/\log C}} \tag{5.68}$$

for all $C \leq t \leq y$. Let $0 < t \leq C \leq y$. Taking derivatives on y, we see that, for $y \geq C$, the right side of (5.68) is at least $t^{-1/\log C} \cdot e/(389 \log C)$. Taking the derivative of this on C, we see that this is at least $t^{-1/\log 10^{12}} \cdot e/(389 \log 10^{12})$. Thus, to verify (5.66), all that is left to do is to check that

$$|\check{m}(t) - 1| \leq \frac{1}{t} + \frac{e}{389 \log 10^{12}} \frac{1}{t^{\frac{1}{\log 10^{12}}}} \tag{5.69}$$

for $0 < t \leq 10^{12}$.

For $11 \leq t \leq 10^{12}$, we have the bound $|\check{m}(t) - 1| \leq 0.0234188/\sqrt{t}$ from (5.54). Since $0.0234188/\sqrt{t} < t^{-1/\log 10^{12}} \cdot e/(389 \log 10^{12})$ for $t \geq 17389$, we see that (5.69) holds for all $17389 \leq t \leq y$. We check the correctness of (5.69) for $2 \leq t < 17389$ by a simple computation. Finally, $|\check{m}(t) - 1| = 1 - \log t < 1/t \leq 1/t^{\sigma}$ for $1 \leq t \leq 2$ and any $\sigma \geq 0$, and $|\check{m}(t) - 1| = 1 \leq 1/t^{\sigma}$ for $0 < t \leq 1$ and any $\sigma \geq 0$.

To verify (5.67), we simply put (5.68) and the last bound in (5.54) together. □

Lemma 5.14. *Let* $9 \leq C \leq 10^{12}$. *For any* $y \geq C$, $0 < t \leq y$,

$$\left|\check{\check{m}}(t) - 2(\log x - \gamma)\right| \leq \frac{2e^{\gamma-1}}{t} + \frac{1}{103} \frac{y^{\frac{1}{\log C}}}{\log y} \frac{1}{t^{\frac{1}{\log C}}}. \tag{5.70}$$

Proof. By (5.50) and Lemma 5.11,

$$\left|\check{\check{m}}(t) - 2(\log t - \gamma)\right| \leq \frac{1}{103} \frac{y^{1/\log C}}{\log y} \frac{1}{t^{1/\log C}} \tag{5.71}$$

for all $C \leq t \leq y$. Let $0 < t \leq C \leq y$. By the same argument as in the proof of Lemma 5.13, it is enough to check that

$$\left|\check{\check{m}}(t) - 2(\log t - \gamma)\right| \leq \frac{2e^{\gamma-1}}{t} + \frac{e}{103 \log 10^{12}} \frac{1}{t^{\frac{1}{\log 10^{12}}}} \tag{5.72}$$

for $0 < t \leq 10^{12}$.

For $1 \leq t \leq 10^{12}$, $|\check{\check{m}}(t) - 2(\log t - \gamma)| \leq 2\gamma/t + 0.00232347/\sqrt{t}$ by (5.57). Since $2\gamma < 2e^{\gamma-1}$ and $0.00232347 < (e/103 \log 10^{12})/t^{1/\log 10^{12}}$ for $t \geq 35$, we get that (5.72) holds for all $t \geq 35$. We verify (5.72) for $1 \leq t \leq 35$ by a computation. The constant $2e^{\gamma-1}$ is the smallest constant c such that $|2(\log t - \gamma)| \leq c/t$ for all $0 < t \leq 1$; in particular, (5.72) holds for $0 < t \leq 1$ as well. □

5.3.4 Coprimality conditions

We would like to estimate the sums m_q, $\check{m}_q$, $\check{\check{m}}_q$ defined in (5.45). We are particularly interested in good bounds for the special case $q = 2$.

The first step is simple: for any $n, q \in \mathbb{Z}^+$,

$$\sum_{d|(q^\infty,n)} \mu\left(\frac{n}{d}\right) = \mu\left(\frac{n}{(q^\infty,n)}\right) \sum_{d|(q^\infty,n)} \mu\left(\frac{(q^\infty,n)}{d}\right) = \begin{cases} \mu(n) & \text{if } (q,n) = 1, \\ 0 & \text{if } (q,n) > 1. \end{cases}$$

It follows easily that, for any function $h : \mathbb{Z}^+ \to \mathbb{C}$,

$$\sum_{\substack{n \\ (n,q)=1}} \frac{\mu(n)}{n} h(n) = \sum_{d|q^\infty} \frac{1}{d} \sum_n \frac{\mu(n)}{n} h(dn). \tag{5.73}$$

As Ramaré points out in the proof of [Ram15, Lemma 1.9], identities such as (5.73) make it easy to derive some bounds on sums with coprimality conditions $(n,q) = 1$ from bounds without such conditions. For example, as was already stated in [Bal12a], $0 \leq \check{m}(x) \leq 1.00303$ for all $x > 0$. (This can be shown easily using (5.49).) We apply (5.73) with $h(n) = \log^+(x/n)$, and obtain [Ram15, Cor. 1.10]:

$$\check{m}_q(x) = O^*\left(1.00303 \frac{q}{\phi(q)}\right) \tag{5.74}$$

for all x and all positive integers q.

Generalizing more precise bounds is trickier, particularly when x/q is small. Again using (5.73), Ramaré managed to show that [Ram15, Thm. 1.12]:

$$m_q(x) = O^*\left(\frac{1}{\log x/q} \cdot \frac{4}{5} \frac{q}{\phi(q)}\right) \tag{5.75}$$

for all $x \geq 1$ and all positive integers $q \leq x$. He also gave analogous bounds for $\check{m}_q$ and $\check{\check{m}}_q$.

Incidentally, while these results are based on the bounds (5.48)–(5.50), which rely on known facts on the Riemann zeta function, we are not using other L-functions. This is a good thing: results based on known facts on L-functions are usually valid for q in much smaller ranges, simply because our knowledge of L-functions is in some senses even more deficient than our knowledge of the Riemann zeta function.

As it turns out, we can use our results from §5.3.3 to prove results like (5.75), but with far better constants, at least for most q.

Proposition 5.15. *For any $y > 1$ and any $q \geq 1$,*

$$\begin{aligned} |m_q(y)| &\leq \prod_{p|q} \left(1 - \frac{1}{\sqrt{p}}\right)^{-1} \cdot \sqrt{\frac{2}{y}} + \prod_{p|q} \left(1 - \frac{1}{p^{1-1/\log 10^{14}}}\right)^{-1} \cdot \frac{0.0144}{\log y}. \\ \left|\check{m}_q(y) - \frac{q}{\phi(q)}\right| &\leq \prod_{p|q} \left(1 - \frac{1}{\sqrt{p}}\right)^{-1} \cdot \frac{1}{\sqrt{y}} \\ &\quad + \prod_{p|q} \left(1 - \frac{1}{p^{1-1/\log 10^{12}}}\right)^{-1} \cdot \frac{1}{389 \log y}. \end{aligned} \tag{5.76}$$

Proof. By (5.73),

$$|m_q(y)| \leq \sum_{d|q^\infty} \frac{1}{d} \left| m\left(\frac{y}{d}\right)\right|.$$

We apply Lemma 5.12, and obtain

$$\begin{aligned} |m_q(y)| &\leq \frac{1}{\sqrt{y}} \sum_{d|q^\infty} \frac{\sqrt{2}}{\sqrt{d}} + \frac{0.0144}{\log y} \sum_{d|q^\infty} \frac{1}{d^{1-1/\log 10^{14}}} \\ &= \frac{\sqrt{2}}{\sqrt{y}} \prod_{p|q} \left(1 - \frac{1}{\sqrt{p}}\right)^{-1} + \frac{0.0144}{\log y} \prod_{p|q} \left(1 - \frac{1}{p^{1-1/\log 10^{14}}}\right)^{-1}. \end{aligned} \tag{5.77}$$

In the same way, starting from (5.67), we obtain

$$\left|\check{m}_q(y) - \frac{q}{\phi(q)}\right| \leq \frac{1}{\sqrt{y}} \prod_{p|q} \left(1 - \frac{1}{\sqrt{p}}\right)^{-1} + \frac{1}{389 \log y} \prod_{p|q} \left(1 - \frac{1}{p^{1-1/\log 10^{12}}}\right)^{-1}. \tag{5.78}$$

□

It is interesting to see that the bounds in Proposition 5.15 – and, in particular, the products $\prod_{p|q}$ therein – resemble what one would get from contour integration, if only (a) we had good bounds on the residues of $1/\zeta(s)$ on the critical line, and (b) all zeros of $\zeta(s)$ were on $\Re s = 1/2$ with at most one exception s_0. (Of course, s_0 would lie outside the classical zero-free region.) Needless to say, we have neither (a) and (b), but we nevertheless managed to obtain Proposition 5.15 by an indirect route.

* * *

The case $q = 2$ is of particular interest to us; we should examine it separately. For $y \leq 10^{12}$, we will use the following inequalities, easy to establish by computation, just as in the proof of Lemma 5.9:

$$|m_2(y)| \leq \begin{cases} 0.390056/\sqrt{y} & \text{for } 1423 \leq y \leq 10^{12}, \\ \sqrt{3/y} & \text{for } 0 < y \leq 10^{12}, \end{cases} \tag{5.79}$$

$$|\check{m}_2(y) - 2| \leq \begin{cases} 0.025358/\sqrt{y} & \text{for } 19341 \leq y \leq 10^{12}, \\ 0.068199/\sqrt{y} & \text{for } 2001 \leq y \leq 10^{12}, \\ 2/\sqrt{y} & \text{for } 0 < y \leq 10^{12}, \end{cases} \tag{5.80}$$

$$|\check{m}_2(y) - 2| \leq \frac{2.9}{y} + \frac{0.013464}{\sqrt{y}} \qquad \text{for } 0 < y \leq 10^{12}, \tag{5.81}$$

$$\left|\check{\check{m}}_2(y) - 4 \log \frac{x}{2} + 4\gamma\right| \leq \begin{cases} \frac{0.01998}{\sqrt{y}} & \text{for } 77581 \leq y \leq 10^{12}, \\ \frac{4(\log 2+\gamma)}{\sqrt{y}} & \text{for } 1 \leq y \leq 10^{12}. \\ \frac{8\sqrt{2}e^{\frac{\gamma}{2}-1}}{\sqrt{y}} & \text{for } 0 \leq y \leq 10^{12}. \end{cases} \tag{5.82}$$

and

$$|\check{\check{m}}_2(y) - 4 \log \frac{y}{2} + 4\gamma| \leq \frac{5.77}{y} + \frac{0.0018124}{\sqrt{y}} \qquad \text{for all } 1 \leq x \leq 10^{12}. \tag{5.83}$$

The alternative would have been to derive bounds from our inequalities for $m(y)$ and $\check{m}(y)$ for y small, but that would have resulted in considerably worse constants.

We can derive bounds for y larger than a constant proceeding as in the proof of Prop. 5.15, with some improvements specific to $q = 2$.

Lemma 5.16. *For any* $3155 \leq C \leq 10^{12}$ *and any* $y \geq C$*,*

$$|\check{m}_2(y) - 2| \leq \left(1 - \frac{1}{2^{1-\frac{1}{\log C}}}\right)^{-1} \frac{1}{389 \log y} + \frac{\lfloor \log_2 C \rfloor + 2}{y}, \tag{5.84}$$

For any $9 \leq C \leq 10^{12}$ *and any* $y \geq C$*,*

$$\begin{aligned} \left|\check{\check{m}}_2(y) - 4\left(\log \frac{y}{2} - \gamma\right)\right| \leq \left(1 - \frac{1}{2^{1-\frac{1}{\log C}}}\right)^{-1} \frac{1}{103 \log y} \\ + \frac{2e^{\gamma-1}\lfloor \log_2 C \rfloor + 4 \log 2 + 4\gamma}{y}. \end{aligned} \tag{5.85}$$

Proof. Assume now that $y \geq C$, where $3155 \leq C \leq 10^{12}$. Then (5.66) applies, and, for $t \geq C$, so does (5.68). Thus, by (5.73),

$$\begin{aligned} |\check{m}_2(y) - 2| &\leq \sum_{d=2^k: k \geq 0} \frac{1}{d} \left|\check{m}\left(\frac{y}{d}\right) - 1\right| \\ &\leq \sum_{d=2^k: 0 \leq k \leq \log_2 y} \frac{1}{d} \left|\check{m}\left(\frac{y}{d}\right) - 1\right| + \sum_{k > \log_2 y} \frac{1}{2^k} \\ &\leq \sum_{d=2^k: 0 \leq k \leq \log_2 y} \frac{1}{d} \cdot \frac{1}{389 \log y} \cdot d^{\frac{1}{\log C}} \\ &+ \sum_{d=2^k: 1 \leq y/d < C} \frac{1}{d} \cdot \frac{1}{y/d} + \sum_{k > \log_2 y} \frac{1}{2^k} \\ &\leq \left(1 - \frac{1}{2^{1-\frac{1}{\log C}}}\right)^{-1} \frac{1}{389 \log y} + \frac{\lfloor \log_2 C \rfloor + 2}{y}. \end{aligned} \tag{5.86}$$

Let us now examine $\check{\check{m}}_2(y)$. Let $y \geq C$, where $9 \leq C \leq 10^{12}$. Again by (5.73),

$$\check{\check{m}}_2(y) - 4\left(\log \frac{y}{2} - \gamma\right) = \sum_{d=2^k: k \geq 0} \frac{1}{d} \left(\check{\check{m}}\left(\frac{y}{d}\right) - 2\left(\log \frac{y}{d} - \gamma\right)\right), \tag{5.87}$$

where we use the fact that $\sum_{k\geq 0} k/2^k = 2$. Hence, by Lemma 5.14,

$$\begin{aligned}
&\left|\check{\check{m}}_2(y) - 4\left(\log\frac{y}{2} - \gamma\right)\right| \\
&\leq \sum_{d=2^k:0\leq k\leq \log_2 y} \frac{1}{d}\left|\check{\check{m}}\left(\frac{y}{d}\right) - 2\left(\log\frac{y}{d} - \gamma\right)\right| + 2\sum_{k>\log_2 y} \frac{\log\frac{2^k}{y} + \gamma}{2^k} \\
&\leq \frac{1}{103\log y}\sum_{d=2^k:0\leq k\leq \log_2 y} \frac{d^{\frac{1}{\log C}}}{d} + \sum_{d=2^k:1\leq y/d<C} \frac{1}{d}\cdot\frac{2e^{\gamma-1}}{y/d} + \frac{4\log 2 + 4\gamma}{y} \\
&\leq \left(1 - \frac{1}{2^{1-\frac{1}{\log C}}}\right)^{-1}\frac{1}{103\log y} + \frac{2e^{\gamma-1}\lfloor\log_2 C\rfloor + 4\log 2 + 4\gamma}{y},
\end{aligned} \tag{5.88}$$

where we are using the fact that $t(\gamma + \log 2 - \log t)$ has its maximum on $[1/2, 1]$ at 1. □

We can now get good bounds for $m_2(y)$, $\check{m}_2(y)$ and $\check{\check{m}}_2(y)$, y large, by combining (5.79) and (5.80) with Lemma 5.16.

Lemma 5.17. *For $y \geq 5379$,*

$$|m_2(y)| \leq \frac{0.0296}{\log y}. \tag{5.89}$$

For $y \geq 4957$,

$$|\check{m}_2(y) - 2| \leq \frac{2}{379}\frac{1}{\log y}. \tag{5.90}$$

For $y \geq 2209$,

$$\left|\check{\check{m}}_2(y) - 4\left(\log\frac{y}{2} - \gamma\right)\right| \leq \frac{1}{50\log y}. \tag{5.91}$$

Proof. By (5.76),

$$|m_2(y)| \leq \frac{2}{\sqrt{2}-1}\frac{1}{\sqrt{y}} + \frac{0.02944}{\log y} \leq \frac{0.0296}{\log y} \tag{5.92}$$

for $y \geq 10^{12}$. The bound (5.79) implies that $|m_2(y)| \leq 0.0296/\log y$ for $16484 \leq y \leq 10^{12}$. We obtain the same bound for $5379 \leq y \leq 16484$ by direct computation.

By (5.84) with $C = 10^{12}$,

$$|\check{m}_2(y) - 2| \leq \frac{41}{y} + \frac{2}{379.118\log y} \leq \frac{2}{379}\frac{1}{\log y}$$

for $y \geq 10^{12}$. The bound (5.80) implies (5.90) for $15701 \leq y \leq 10^{12}$. We establish (5.90) for $4957 \leq y < 15701$ by direct computation.

Finally, by (5.85) with $C = 10^{12}$,

$$\left|\check{\check{m}}_2(y) - 4\left(\log\frac{y}{2} - \gamma\right)\right| \leq \frac{56.19}{y} + \frac{0.019924}{\log y} \leq \frac{1}{50\log y} \tag{5.93}$$

for $y \geq 10^{12}$. For $77581 \leq y \leq 10^{12}$, we get the same inequality immediately from (5.82). We verify the inequality for $2209 \leq y \leq 77581$ by direct computation. □