Currently there are no reasons to doubt the Riemann Hypothesis

The zeta function beyond the realm of computation

David W. Farmer

Abstract

We examine published arguments which suggest that the Riemann Hypothesis may not be true. In each case we provide evidence to explain why the claimed argument does not provide a good reason to doubt the Riemann Hypothesis. The evidence we cite involves a mixture of theorems in analytic number theory, theorems in random matrix theory, and illustrative examples involving the characteristic polynomials of random unitary matrices. Similar evidence is provided for four mistaken notions which appear repeatedly in the literature concerning computations of the zeta-function. A fundamental question which underlies some of the arguments is: what does the graph of the Riemann zeta-function look like in a neighborhood of its largest values? We explore that question in detail and provide a survey of results on the relationship between L-functions and the characteristic polynomials of random matrices. We highlight the key role played by the emergent phenomenon of carrier waves, which arise from fluctuations in the density of zeros. The main point of this paper is that it is possible to understand some aspects of the zeta function at large heights, but the computation evidence is misleading. [1]

1 Introduction

Should one believe the Riemann Hypothesis (RH)? Since it is a conjecture with no proposed roadmap to prove it, one point of view is that it should neither be believed nor disbelieved. Yet many mathematicians have an opinion, presumably backed up by logical reasoning.

Here we consider all published arguments for doubting RH. A paper of Ivić [65, 66] lists 4 reasons, and a paper of Blanc [14] provides a 5th reason. Three of those reasons involve speculation about the distribution of zeros and their relationship to the value distribution of the ζ-function. The fundamental question there is: what does the ζ-function look like in a neighborhood of its largest values? The majority of this paper is a survey of prior results, and speculations and heuristics based on those results, which lead to an answer to that question.

Our primary goal is to provide intuition and to be persuasive. The arguments against RH generally take the form "It would be surprising if **X**." So, the burden we bear in refuting that argument is to give good reasons why **X** is not surprising. To crystallize our main points we present 33 *Principles* which we hope are also useful for future reference.

Is the purpose of this paper to persuade that RH is true? Certainly not. But perhaps those who continue to doubt RH will realize that their belief is

[1] MSC2020: 11M26, 11M50

not based on good evidence. As for those who believe RH: perhaps someone will write a companion paper: *Currently there are no good reasons to believe the Riemann Hypothesis.* Note the subtle difference from the opposite of the title to this paper. Also useful would be a paper explaining why every currently known equivalence to RH is unlikely to be helpful for proving RH (as in [41]).

The ζ-function is the simplest example of an *L-function.* All L-functions have properties similar to the ζ-function, and all L-functions have an analogue of the Riemann Hypothesis. As much as possible we try to discuss the ζ-function in isolation, but in a few places it is necessary to expand our perspective. We attempt to keep this paper self-contained, providing definitions and background as needed.

The themes. Many of the surprising properties of the ζ-function arise because there are different facets to the same object. A classic example is the function usually denoted $S(t)$: it is the error term in the counting function of the zeros of the ζ-function, and it also is the imaginary part of the logarithm of $\zeta(\frac{1}{2}+it)$. That equivalence is basic complex analysis. But when pondering a specific question about $S(t)$, sometimes one perspective gives good intuition, and sometimes another. Combining perspectives can lead to surprising relations, such as Principle 8.7: in regions where $|\zeta(\frac{1}{2}+it)|$ is large, $S(t)$ tends to be decreasing.

A second theme is multiple levels of randomness. We will make extensive use of the fact that the ζ-function and characteristic polynomials of random unitary matrices have a similar sort of randomness (and also some differences, which we will describe). In the random matrix world, it is a theorem that everything just follows from the quadratic repulsion between the eigenvalues. Everything. But (thank you to an anonymous referee for helping me see it this way), that perspective is reductive and it is beneficial to have other points of view, particularly on larger scales where the individual zeros are not visible. In the ζ-function world this means that sometimes the primes will enter the discussion, and other times they will be ignored.

The sections. In Section 2 we introduce the main theme by pulling together various ideas in the paper to answer the question: What does the zeta function look like beyond the realm of computation? The remainder of the paper is primarily devoted to providing intuition which will dispel misconceptions about the answer to that question — misconceptions largely due to poor extrapolation from available computations. In Section 3 we provide basic definitions and background. In Section 4 we describe four Mistaken Notions which appear repeatedly in discussions of computations of the ζ-function, some of which play an important role in the claimed reasons to doubt RH. In Section 5 we describe the connection between the distribution of zeros and the size of the ζ-function, introducing *carrier waves* as a way to separate local from long-range behavior. In Section 6 we briefly describe the connection between the ζ-function and unitary polynomials, and in Section 7 we provide an historical account of the connections to Random Matrix Theory. This leads to Section 8, where we use large unitary matrices to illustrate phenomena which occur far outside the range in which we can compute the ζ-function. By the end of Section 8 we have a good understanding of the "typical" large values of the ζ-function, and the relationship between the carrier wave, the density wave, and $S(t)$, but it is not until Section 9 that we address the most extreme values. The primes are only briefly mentioned up to this point, a shortcoming we address in Section 10. The zeros have dominated most of our discussion, but in Section 11 we discuss randomness of the ζ-function or characteristic polynomials without reference to

zeros or eigenvalues. After all that preparation, in Section 12 we use information from the prior sections to refute the three arguments against RH based on the distribution of zeros and values of the ζ-function, and for completeness in Section 13 we cite recent results to refute the other two arguments against RH. Finally, in Section 14 we use the Principles to explain why the Mistaken Notions of Section 4 are, in fact, mistaken.

Acknowledgments. I thank Louis-Pierre Arguin, Juan Arias de Reyna, Emma Bailey, Sir Michael Berry, Philippe Blanc, Richard Brent, Brian Conrey, Jon Keating, Hugh Montgomery, Eero Saksman, Tim Trudgian, and Christian Webb for clarifying several points in this article. I also thank Jonathan Bober, Xavier Gourdon, and Ghaith Hiary for making available extensive data from their computations of the ζ-function. In addition, useful suggestions from a referee led to many improvements. This paper was written in PreTeXt [84].

2 The ζ-function as a random object

We present some of the main points in this paper in a thought experiment to give intuition about the behavior of the ζ-function far beyond the range where it can be computed. It is hoped that the reader will view the claims in this overview with some skepticism, leading to closer scrutiny of the ideas and perspective in the remainder of the paper.

2.1 Snapshots of the ζ-function

As described in Section 3, we will consider the function $Z(t)$, which contains the same information as the Riemann ζ-function, but it is a real-valued function of a real variable, so we can graph it. What might the graph of $Z(t)$ look like for $t \approx 100^{100^{100}}$? Current methods fail well before 10^{40}, so such a computation is not remotely plausible. But suppose in some distant future a computer algebra package allowed one to graph $Z(t)$ for such large t, on an interval of, for example, width $40/(2\pi \log(t/2\pi))$ — an interval which should contain around 40 zeros. By Principle 7.3 and Principle 7.4, combined with Principle 5.1 or Principle 10.2, it is plausible that the graph might look like this:

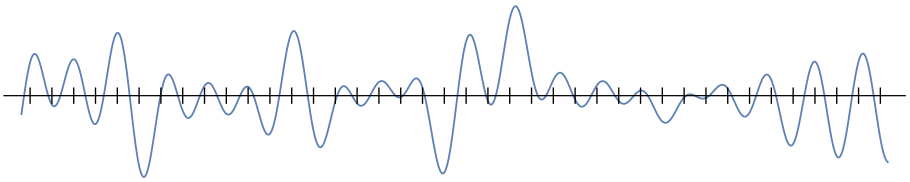

Figure 2.1 What $Z(t)$ looks like, on some interval which is expected to contain around 40 zeros, near $t \approx 100^{100^{100}}$. The tick marks are separated by the average gap between zeros at that height.

Neither axis in Figure 2.1 has a scale, although we are told enough information to determine the width covered by the horizontal axis. The scale on the vertical axis is the mystery.

By Principle 5.1, the Z-function is close to being a polynomial. In particular, locally the behavior of the Z-function is determined by its zeros, up to a local constant factor. So, if the relative spacing of the zeros in Figure 2.1 is plausible for t of that size, then the given graph is accurate: as is typical for computer algebra systems, the vertical scale is automatically adjusted to show the features of the graph.

By Principle 7.3, the distribution of zeros in Figure 2.1 is indeed plausible, and furthermore we expect to find many intervals of that width at that height, on which the graph of the Z-function looks similar to Figure 2.1. The only thing we don't know is: what is the vertical scale of that graph?

By Selberg's Theorem (3.15), $\log Z(t)$ is normally distributed with mean zero and variance $\sqrt{\frac{1}{2}\log\log t}$. So, at height $t \approx 100^{100^{100}}$, half the time $\log Z(t)$ is likely to be of size ≈ 15, and the other half of the time it is likely to be ≈ -15. Exponentiating, we find that the vertical scale in Figure 2.1, or in any other interval at that height where the graph looks like Figure 2.1, is likely to be around 10^6, or it is likely to be around 10^{-6}, and both possibilities are equally probable. In particular, the local distribution of zeros is not the main factor in the size of the ζ-function.

This fundamental fact, that the size of the ζ-function is not strongly dependent on the local distribution of zeros, is hard to believe when your intuition only comes from computations in the modest range currently accessible by computers.

Snapshots do not show the big picture. By **snapshot** we refer to an image showing the graph of a function, having an aspect ratio between 1/10 and 2, over an interval where the function has a moderate number of interesting features, with the vertical scale adjusted so that those features are visible. Figure 2.1 is an example, as are the numerous other graphs in this paper. A snapshot of the ζ-function typically covers an interval containing between 10 and 100 zeros. As mentioned above, a snapshot of the Z-function depends only on the location of the nearby zeros. More precisely, it depends only on the relative sizes of the gaps between zeros. Near the middle of the snapshot all that matters is the zero gaps within the snapshot; near the edges, the gaps immediately outside the snapshot are relevant.

Any specific (finite) sequence of normalized zero gaps will never happen (except once, if it is specifically constructed in that way), but precision is not relevant, because a snapshot is a continuous function of the zero gaps. Given an open neighborhood of any specific (finite) gap sequence, such an approximate gap sequence will occur infinitely many times, and in fact will occur a positive proportion of the time among all normalized gap sequences of that length — the precise proportion depending on the given sequence and the size of the neighborhood. See Principle 7.3 or Principle 7.4. For example, Figure 4.1 will occur a positive proportion of the time (to within the resolution of the human eye) as a snapshot of the Z-function.

To summarize:

Principle 2.2 Snapshots of the ζ-function are well understood. *Any given finite sequence of normalized gaps between zeros will (approximately) occur a positive proportion of the time, as predicted by the Random Matrix Model for zeros of the zeta function. The appearance of the graph of the Z-function over that region, ignoring the vertical scale, is determined by the relative spacing of those zeros.*

In particular, one does not expect to see anything surprising when looking at a new snapshot of the Z-function.

What counts as "surprising" changes over time. In the early computations of zeros of the ζ-function, it was surprising that occasionally two zeros would be very close together (known as **Lehmer pairs**). Today such pairs of close zeros are expected, and those occur at the frequency predicted by Principle 7.3.

3 Background on the ζ- function

The Riemann zeta-function (which we will call the ζ-function) is defined by

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s} \tag{3.1}$$

for $\sigma > 1$, where $s = \sigma + it$ is a complex variable. The ζ-function has a meromorphic continuation to the complex plane, with a simple pole at $s = 1$ with residue 1.

The ζ-function has a symmetry, known as the **functional equation**, which can be expressed in several ways. With

$$X(s) := \pi^{s-\frac{1}{2}} \frac{\Gamma(\frac{1}{2} - \frac{1}{2}s)}{\Gamma(\frac{1}{2}s)} \tag{3.2}$$

where Γ is the Euler Gamma-function, we have

$$\zeta(s) = X(s)\zeta(1-s), \tag{3.3}$$

or equivalently

$$\xi(s) := \tfrac{1}{2}s(s-1)\pi^{-\frac{1}{2}s}\Gamma(\tfrac{1}{2}s)\zeta(s) \tag{3.4}$$

$$= \xi(1-s), \tag{3.5}$$

or equivalently

$$Z(t) := X(\tfrac{1}{2} + it)^{-\frac{1}{2}}\zeta(\tfrac{1}{2} + it) \quad \text{is real if} \quad t \in \mathbb{R}. \tag{3.6}$$

The factor $\frac{1}{2}s(s-1)$ in (3.4) is irrelevant to the invariance under $s \leftrightarrow 1-s$, but it is traditionally included so that $\xi(s)$ is an entire function. In (3.6) the square root is chosen so that $Z(0) = \zeta(\frac{1}{2}) \approx -1.46$ and $Z(t)$ is analytic for $|\Im(t)| < \frac{1}{2}$. It is common to refer to $|t|$, the magnitude of the imaginary part of $s = \sigma + it$, as the **height** when referring to the behavior of $\zeta(s)$ or $Z(t)$ in a particular region.

The **Hardy Z-function** (3.6) is useful because it can be graphed, and it tells us essentially everything we might want to know about the ζ-function because $|Z(t)| = |\zeta(\frac{1}{2} + it)|$ if $t \in \mathbb{R}$. Figure 3.1 shows $Z(t)$ and $\log|Z(t)|$ for $5429.29 < t < 5466.44$, along with the function $S(t)$ which we will introduce shortly.

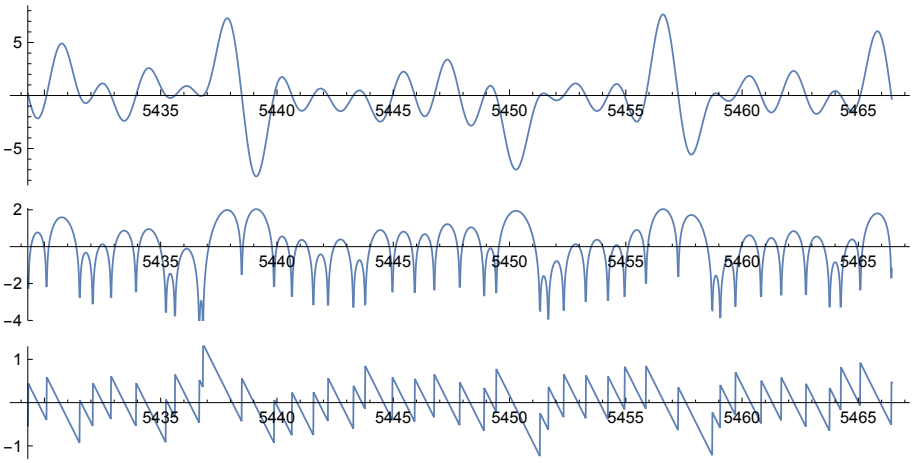

Figure 3.1 $Z(t)$, $\log|Z(t)|$, and $S(t)$ near the 5000th zero $\gamma_{5000} \approx 5447.86$.

The critical facts (pun intended) about $Z(t)$ are that it is smooth, and if t is real then $Z(t)$ is real and $|Z(t)| = |\zeta(\frac{1}{2} + it)|$. Those facts do not uniquely determine $Z(t)$: there is still a global choice of ± 1. One might ask whether the choice matters, and if it does, is the standard choice for the square-root in (3.6) the correct option?

Principle 3.2 *$Z(t)$ is statistically indistinguishable from $-Z(t)$.*

In other words, any measure of distance between probability distributions cannot distinguish between the distributions of $Z(t)$ and $-Z(t)$ for $t \in [T, 2T]$, as $T \to \infty$.

Thus, the global choice of sign for $Z(t)$ does not matter. Indeed, $\xi(\frac{1}{2} + it)$ is also real for $t \in \mathbb{R}$, and has the same critical zeros as $Z(t)$, but it has the opposite sign of $Z(t)$. So, the standard normalizations for those functions disagree with each other. A consequence of Principle 3.2 is that if k is a non-negative integer then

$$\int_T^{2T} Z(t)^{2k+1}\, dt = o(T), \tag{3.7}$$

and as $T \to \infty$ that integral changes sign infinitely many times. Equation (3.7) is a theorem for $k = 0$, use the Riemann-Siegel formula (4.1) and integrate term-by-term, and a conjecture for larger k.

Principle 3.2 is a stronger statement than asserting that to leading order the value distribution of $Z(t)$ is symmetric. In contrast, the average value of $\zeta(\frac{1}{2} + it)$ is 1 in a very strong sense: if k is a non-negative integer then

$$\int_T^{2T} \zeta(\tfrac{1}{2} + it)^k\, dt \sim T, \tag{3.8}$$

which can be shown by moving the integral to the region of absolute convergence of the Dirichlet series and then integrating term-by-term. Such small biases (meaning: effects of size 1, when the main term is growing very slowly) are the underlying cause of many misconceptions arising from numerical computations, see (4.4) and Principle 4.2.

3.1 We care about zeros because we care about primes

Riemann's great insight was that the zeros of the ζ-function encode information about the primes. Based at least partially on numerical computation, he formulated the conjecture which is now known as the **Riemann Hypothesis**:

Conjecture 3.3 The Riemann Hypothesis (RH). *The zeros of $\zeta(s)$ with $0 < \sigma < 1$ lie on the line $\sigma = \frac{1}{2}$; equivalently, all zeros of $\xi(s)$ lie on the line $\sigma = \frac{1}{2}$; equivalently, all zeros of $Z(t)$ with $|\Im(t)| < \frac{1}{2}$ are real.*

Riemann's original formulation involved yet another function: all the zeros of $\Xi(t) := \xi(\frac{1}{2} + it)$ are real. RH has been rigorously verified for $t < 3 \times 10^{12}$, involving more than 12.3×10^{12} zeros [82].

The zeros of $\zeta(s)$ with $0 < \sigma < 1$ are traditionally denoted by $\rho = \beta + i\gamma$. By the functional equation and the fact that $\zeta(s)$ is real if s is real, if $\rho = \beta + i\gamma$ is a zero of $\zeta(s)$ then so is $\rho = 1 - \beta + i\gamma$. So either $\beta = \frac{1}{2}$, as predicted by RH, or there is a pair of zeros of $\zeta(s)$ (or $\xi(s)$) located symmetrically around the **critical line** $\sigma = \frac{1}{2}$. Equivalently, a failure of RH corresponds to a pair of complex conjugate zeros of $Z(t)$. Assuming RH, the zeros of $\zeta(s)$ with positive imaginary part are denoted $\frac{1}{2} + i\gamma_n$ with $0 < \gamma_1 < \gamma_2 < \cdots$. That notation would break down if there were repeated zeros, but no plausible reason has been

given to expect a multiple zero. Various discussions in this paper may implicitly assume all zeros are simple, an assumption which generally is irrelevant to the points being made.

There are 41 zeros of the ζ-function visible in Figure 3.1. RH implies that (for $|t| > 3$) all local maxima of $Z(t)$ (or $\xi(\frac{1}{2} + it)$) are positive and all local minima are negative, a condition which is visible in the first two graphs in Figure 3.1. The converse is not necessarily true — a failure of RH need not cause a negative maximum or a positive minimum, although that is mistakenly asserted in the official statement of the Riemann Hypothesis Millennium Problem [19]. The issue is the distance of a hypothetical non-critical zero from the critical line, see [42]. So, one cannot trivially deduce from the graph of $Z(t)$ or $\log(|Z(t)|)$ in Figure 3.1 that RH holds for $5430 < t < 5448$.

It is possible to verify RH on an interval by a computer calculation, based on two properties of the ζ-function. One property we have already seen: $Z(t)$ is real when t is real, so one can count critical zeros by looking for sign changes. The other is that the zeros of the ζ-function have a nice counting function with a small and computable error term. Let $N(T)$ be the number of zeros of $\zeta(s)$ with $0 < \gamma_n \le T$. We have [101], assuming T is not the imaginary part of a zero of $\zeta(s)$,

$$N(T) = \frac{1}{2\pi} T \log T - \frac{\log 2\pi e}{2\pi} T + \tfrac{7}{8} + S(T) + O(T^{-1}) \tag{3.9}$$

where

$$S(T) = \frac{1}{\pi} \arg \zeta(\tfrac{1}{2} + iT) = \frac{1}{\pi} \Im \log \zeta(\tfrac{1}{2} + iT). \tag{3.10}$$

The argument in (3.10) is determined by continuous variation along the line $2 + it$ for $0 \le t \le T$, and then along the line $\sigma + iT$ for $2 \ge \sigma \ge \frac{1}{2}$. One can take either (3.9) or (3.10) as the definition, with the other as a theorem.

We will see that the function $S(t)$ grows very slowly. The third plot in Figure 3.1 illustrates the basic properties of $S(t)$: it has a jump discontinuity (of height 1) at a simple critical zero, and where it is continuous it is approximately linear with slope $-\frac{1}{2\pi} \log t$.

By (3.9) and (3.10) one can rigorously prove that RH holds on an interval: use sign changes to count real zeros of $Z(t)$, then compute the change in $S(t)$ to determine the change in $N(t)$ and thus find the total number of zeros. Check if those two quantities are equal. A sophisticated version of this idea is known as **Turing's method**; see [21] for a modern treatment. In Figure 3.1 a failure of RH would correspond to a jump discontinuity in $S(t)$ of height 2, at a point where $Z(t)$ does not have a zero. Such a jump does not occur in the graph of $S(t)$, so by also considering either $Z(t)$ or $\log |Z(t)|$, one can "see" a proof in Figure 3.1 that RH is true for $5430 < t < 5448$.

3.2 Unfolding the zeros

By (3.9) the zeros at larger height are on average closer together. Specifically, at height T the average gap between zeros is $2\pi/\log T$, and $\gamma_n \approx 2\pi n/\log n$. When discussing the statistics of the zeros, in particular the gaps between zeros, it is helpful to use the **normalized** or **unfolded** zeros $\tilde{\gamma}_n = \frac{1}{2\pi} \gamma_n \log \gamma_n$. Note that $\tilde{\gamma}_n \sim n$ and $\tilde{\gamma}_{n+1} - \tilde{\gamma}_n$ equals 1 on average.

When it is necessary to be more precise than the asymptotic behavior, we set $\tilde{\gamma}_n = \widetilde{N}(\gamma_n)$, where

$$\widetilde{N}(T) = \frac{1}{2\pi} T \log T - \frac{\log 2\pi e}{2\pi} T + \tfrac{7}{8}. \tag{3.11}$$

We leave it as an exercise to determine the average value of $\tilde{\gamma}_n - n$ as $n \to \infty$. The answer is in Subsection 8.6. That answer also explains the usual definition of $S(t)$ in the case t is the imaginary part of a zero of the ζ-function.

3.3 The size of $Z(t)$ and $S(t)$

The function $S(t)$ is both the error term in the zero counting function $N(t)$ and the imaginary part of $\log \zeta(\frac{1}{2} + it)$. Viewed as an error term for a counting function, it is perhaps surprising that $S(t)$ grows very slowly: much more slowly that the error term in the prime number theorem, or the Dirichlet divisor problem, or other counting problems, for example. A consequence is that the zeros cannot stray too far from their expected location. We will see that this has profound implications for the behavior of the ζ-function.

Assuming RH, Littlewood [101] showed that $S(T) = O(\log T / \log\log T)$. It is conjectured [43] that

$$|S(T)| \le (1 + o(1)) \frac{1}{\pi} \sqrt{\frac{1}{2} \log T \log\log T}, \tag{3.12}$$

and that bound is sharp. The results for $Z(t)$ are analogous: on RH we have [101] $\log |Z(t)| = O(\log t / \log\log t)$, and the conjecture (with sharp constant) is

$$\log |Z(T)| \le (1 + o(1)) \sqrt{\frac{1}{2} \log T \log\log T}. \tag{3.13}$$

In the other direction [18], the current best result is that there exists $C > 0$ such that there exist arbitrarily large T with:

$$\log |Z(T)| > C \sqrt{\frac{\log T \log\log\log T}{\log\log T}}. \tag{3.14}$$

The typical size of $Z(t)$ is much smaller. Selberg [94] proved that if t is chosen uniformly at random from $[T, 2T]$, then

$$\frac{\log \zeta(\frac{1}{2} + it)}{\sqrt{\frac{1}{2} \log\log T}} \to N(0,1) \qquad \text{as} \qquad T \to \infty, \tag{3.15}$$

where $N(0,1)$ is the standard (complex) Gaussian, and (3.15) indicates convergence in distribution. In particular, $\log |Z(t)|$ and $S(t)$ each have a (real) Gaussian distribution, which are different only because the definition of $S(t)$ contains a factor of $1/\pi$, and those distributions are independent.

In numerical computations of the ζ-function the scale factor in Selberg's theorem is practically irrelevant: at $T_{BH} = 10^{33}$, which is approximately the largest height where the ζ-function has been calculated [17], Selberg's theorem (assuming it is valid at that low height) says that the typical size of $\log |Z(t)|$ is $\sqrt{\frac{1}{2} \log\log T_{BH}} \approx 1.5$. The conjectured extreme values are larger, $\sqrt{\frac{1}{2} \log T_{BH} \log\log T_{BH}} \approx 12.8$, so within the realm of current computation one might expect to see $|Z(t)|$ larger than $300,000$. Unfortunately, the extreme values are rare, and possibly (due to lower order terms) the predicted largest values do not occur until significantly greater heights. The largest value of $Z(t)$ found in [17] is approximately 16244 near 3.92×10^{31}, which is 3.7% less than the largest computed value [100]. The largest calculated value of $S(t)$ is 3.345 near $t = 7.7573 \times 10^{27}$.

Figure 3.4 shows $Z(t)$, $\log|Z(t)|$, and $S(t)$ in a neighborhood of the largest value of $Z(t)$ found in [17]. (Note that there is no good linear scale on which to plot $Z(t)$ in that region.) The author thanks Jonathan Bober and Ghaith Hiary for providing open access to their extensive data [59].

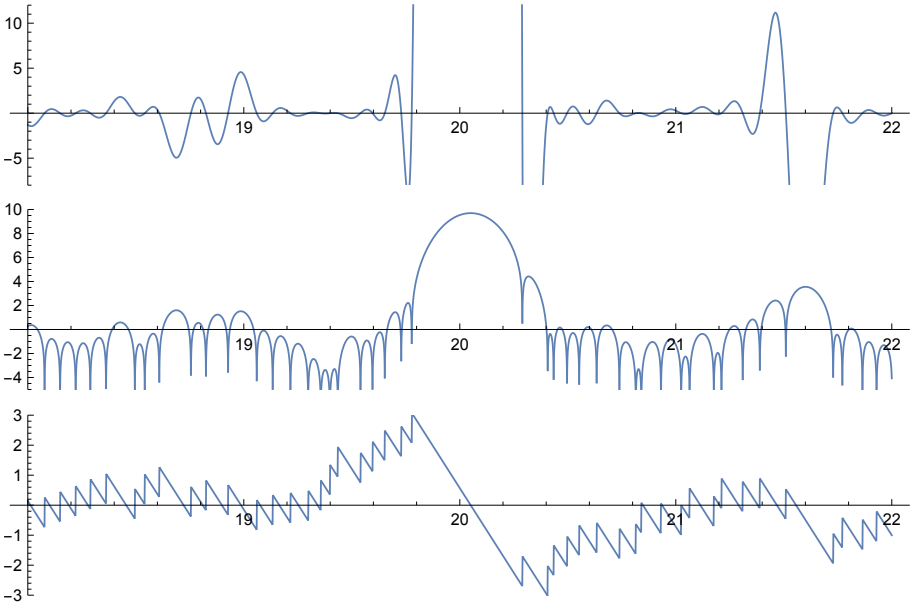

Figure 3.4 Plots of $Z(T_{big}+t)$, $\log|Z(T_{big}+t)|$, and $S(T_{big}+t)$ where $T_{big} = 392467645898943091552511692840 84 \approx 3.9\times 10^{31}$.

In Figure 3.4 we see that this particular large value of $Z(t)$ arises from a large gap between neighboring zeros. The large zero gap also contributes to the large value of $S(t)$: a zero gap of K times the local average spacing must be accompanied by a value $|S(t)| \geq K/2$. The large zero gap in Figure 3.4 is 5.93 times the local average. The relationship between the size of an isolated large zero gap and the local maximum of $Z(t)$ is subtle: see Principle 9.2. The relationship between a large gap and $S(t)$ is almost trivial, but we record it for later use. We write $S^+(t)$ and $S^-(t)$ for the right and left limiting values, respectively.

Principle 3.5 *If $\tilde{\gamma}_{j+1}-\tilde{\gamma}_j = K$, then $S^+(\gamma_{j+1})-S^-(\gamma_j) = K$. So in particular either $|S^+(\gamma_{j+1})| \geq K/2$ or $|S^-(\gamma_j)| \geq K/2$. Thus, an upper bound on $|S(t)|$ implies a comparable upper bound on the size of the normalized zero gaps.*

Note that Principle 3.5 does not say that a large value of $S(t)$ must be accompanied by a large gap between zeros.

As will be explored in detail, Figure 3.4 does not illustrate the typical behavior for the large values of $Z(t)$ and $S(t)$. One way to see this is from Selberg's theorem [94] that $\log|Z(t)|$ and $S(t)$ are independently distributed. One of those being typically large should have no effect on the other. Figure 3.4 illustrates that a large zero gap causes large values of $Z(t)$ and $S(t)$ to occur in close proximity, therefore that cannot be the typical behavior near a large value. (This discussion does not address the question of independence in the tails of the distributions of $\log|Z(t)|$ and $S(t)$.)

In Section 4 we briefly explore the history of finding large values of $Z(t)$ and explain why that work has inadvertently led to a mistaken impression of what the graph of $Z(t)$ looks like in a neighborhood of its largest values. In Section 5 we introduce *carrier waves*, which are the actual cause of the largest values of $Z(t)$.

4 Misleading ideas about large values

Computation has been an important tool for studying the ζ-function ever since Riemann calculated the first few zeros by hand. Computers have enabled large-scale computations: large on the the human scale but minuscule on an absolute scale. In reference to whether existing computations should be seen as evidence for RH, Andrew Odlyzko [79] has sounded a cautionary note: the true nature of the ζ-function is unlikely to be revealed until we reach regions where $S(t)$ is routinely over 100. Since (by Selberg's theorem) $S(t)$ is typically of size $\sqrt{\frac{1}{2}\log\log t}$, such regions will be inaccessible for a long time.

Despite Odlyzko's warning, there are certain aspects of the ζ-function which appear in computations and have influenced the direction of research, but which do not accurately portray the true nature of the ζ-function. These misconceptions are partially based on the way the ζ-function is computed, which we describe next.

4.1 Computations of $Z(t)$

The earliest large-scale computations of the ζ-function calculated $Z(t)$ using the **Riemann-Siegel formula**:

$$Z(t) = 2 \sum_{n<\sqrt{t/2\pi}} n^{-\frac{1}{2}} \cos(\theta(t) - t\log n) \;+\; \text{remainder} \tag{4.1}$$

where

$$\begin{aligned}\theta(t) &= \arg\left(\pi^{-it/2}\Gamma\left(\frac{1}{4}+i\frac{t}{2}\right)\right)\\ &= \frac{t}{2}\log\left(\frac{t}{2\pi}\right) - \frac{t}{2} - \frac{\pi}{8} + O(t^{-1}).\end{aligned} \tag{4.2}$$

The function $\theta(t)$ arises in an alternate expression for $Z(t)$:

$$Z(t) = e^{i\theta(t)}\zeta(\tfrac{1}{2}+it). \tag{4.3}$$

The $n=1$ term in (4.1) is the largest, and if t is small then that term has a strong influence on the overall sum. The points where $\cos(\theta(t)) = \pm 1$ are known as **Gram points**, numbered so that $\theta(g_m) = m\pi$. Almost equivalently, Gram points are the locations where $\zeta(\frac{1}{2}+it)$ is real but nonzero. (The "almost" is because 0 is not considered to be a Gram point, and it has not been proven (but certainly it is true) that the ζ-function does not vanish at any Gram point.)

Gram [54] noted that the first several zeros of $Z(t)$ interlace the Gram points. In other words, when t is small the other terms are insufficient to flip the sign of the first term at a Gram point.

In the course verifying RH for the first 75 million zeros, Brent [23] found that $Z(t)$ was unusually large, more than 79.6, at the 70 354 406th Gram point. Furthermore, at that point *the first* 72 *terms in the Riemann-Siegel formula (4.1) were positive.* Figure 4.1 shows a graph of $Z(t)$ near that Gram point.

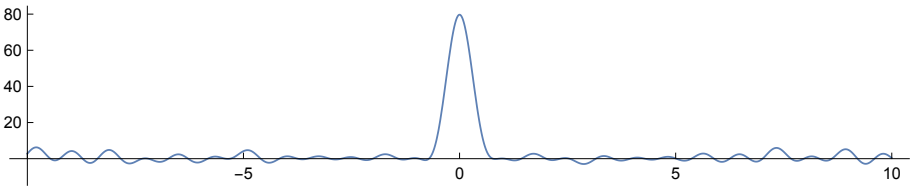

Figure 4.1 Plot of $Z(g_{70354406}+t)$ for $-10<t<10$, where $g_{70354406}\approx 30694257.761$ is a Gram point.

At a Gram point the first term in the Riemann-Siegel formula is 1 or -1, and the other terms have no particular bias, so it is not surprising that [101]

$$\sum_{n\le N,\ n\ \text{even}} Z(g_n)\sim N \qquad \text{and} \qquad \sum_{n\le N,\ n\ \text{odd}} Z(g_n)\sim -N. \tag{4.4}$$

Those averages are similar to the fact that the average value of $\zeta(\frac{1}{2}+it)$ is 1, see (3.8). Such biases, combined with Selberg's central limit theorem, are major contributors to:

Principle 4.2 *For questions about the size of the ζ-function, no numerical computation can give reliable evidence because the true nature of the ζ-function reveals itself on the scale of $\sqrt{\log\log T}$, which within the realm of computation is indistinguishable from a bias of order $O(1)$.*

Principle 4.2, and other sources for misinterpreting the data, are explored throughout this paper.

4.2 Three unfounded inferences

Brent's observations have re-appeared in subsequent numerical computations. The result is that reinforcing those observations has been (at least partially) a goal of recent computational work, particularly as it relates to understanding the largest values of the ζ-function. That is unfortunate because the largest values as they appear in computations are not representative of the largest values at greater height. Thus, the impressions one has from those data are not helping to build intuition for the true nature of the ζ-function. Quite the opposite: those numerical examples tempt one into mistaken notions.

The notions we refer to are:

Mistaken Notion 4.3 *The largest values of the ζ-function occur when there is a particularly large gap between zeros.*

Mistaken Notion 4.4 *The largest values of the ζ-function occur when a large number of initial terms in the Riemann-Siegel formula (4.1) have the same sign.*

Mistaken Notion 4.5 *Counterexamples to RH are more likely to occur near an unusually large gap between zeros.*

These Notions are part of the folklore of the ζ-function and the author has observed conversations on these Notions at many conferences, often accompanied by refutations. Thus, it seemed prudent to address these issues explicitly.

We do not attribute those Notions as conjectures due to any specific person. Indeed, in many cases where those Notions appear in the literature (implicitly or explicitly), the writer acknowledges the Notion without necessarily endorsing it. For example, concerning Mistaken Notion 4.4, Brent [23] says "This suggests that 'interesting' regions might be predicted by finding values of t such that the first few terms in the Riemann-Siegel sum reinforce each other." The "interesting" behavior of an unusually large gap does indeed occur when the initial terms reinforce, at least in the range accessible by computers. That

Notion is not Mistaken in the limited context of computing the ζ-function at moderate height.

The even more "interesting" behavior of a possible counterexample to RH requires a second step to get to Mistaken Notion 4.5. Odlyzko [81] explains it this way: "One reason for the interest in large values of $\zeta(1/2+it)$ is that one could think of a large peak as 'pushing aside' the zeros that would normally lie in that area, and if these zeros were pushed off of the critical line, one would find a counterexample to the RH." Odlyzko is not endorsing Mistaken Notion 4.5, merely noting the reasoning which might have led to its formulation.

Brent's and Odlyzko's use of "scare quotes" is a further indication that they are not endorsing those Notions as fundamental principles. Indeed, it may be that individually the vast majority of experts understand the limitations of what can be learned from existing computations. But the frequency with which those Notions have been repeated, not always with qualifiers such as "it has been said that...", indicates the need for clarification. In Section 14 we revisit these Notions in the context of the Principles discussed in this paper, justifying the claim that the Notions are indeed "Mistaken".

4.3 Gram's law

If (g_n) are the Gram points for the Z-function, then we say **Gram's law holds** for the **Gram interval** (g_m, g_{m+1}) if that interval contains exactly one zero of the Z-function. Gram noted that this law holds for the first 15 zeros, and Hutchinson [63] found that it holds for the first 126 zeros, and he also coined the term "Gram's law". Subsequent work found that it holds for more than 91% of the first 1000 zeros, and most papers on zero calculations report statistics on Gram's law. Such numerical evidence, and repeated attention, leads to the archetypal mistaken notion concerning zeros:

Mistaken Notion 4.6 *Gram points are special, and Gram's law is helpful for developing intuition about the location of zeros of the ζ-function.*

It has been known for a long time that Gram's law fails infinitely often [101] and in fact fails a positive proportion of the time (although there are good reasons to believe it is true more than 66% of the time, see [102, 55]). Gram's law is not mistaken merely because it is not always true: it is mistaken because it paints a picture of the zeros which inhibits one from gaining a proper intuition about the behavior of the ζ-function at large height. Also, as we will explain in Subsection 14.4, if Gram's law is stated properly, it is true 0% of the time.

5 Separating out the local zero spacing

Figure 3.1 and Figure 3.4 show that the local spacing of the zeros has an influence on the size of the ζ-function: when there is a large gap the function is larger, and when there is a small gap the function stays small. Unfortunately, those observations require one to *ignore the vertical scale*. The specific spacing of nearby zeros strongly influences the *relative* sizes of the *nearby* maxima and minima. But the local spacings have very little to do with whether or not the actual function values are particularly large or small, compared to what one would expect for the ζ-function in that region.

Indeed, by the end of this section we will see that the local spacing of zeros is not the leading order contribution to the size of the ζ-function.

In general, the size of the ζ-function is controlled by the ζ-function's **carrier wave**. The terminology is due to Hejhal [58], with detailed discussion by

Bombieri and Hejhal [20]. The idea of carrier waves is based on some unpublished speculations of H.L. Montgomery. The main idea is that, on a logarithmic scale, the ζ-function changes its size slowly. If it is large, then outside a set with small measure, it usually stays large for a while. (The "set with small measure" is small neighborhoods of the zeros.) If it is small, it usually stays small for a while. Here "a while" can be interpreted as "on an interval containing many zeros". In particular, the typical large values occur as clusters of large local maxima and minima, not as a single isolated large maximum. That is contrary to what we see in graphs of $Z(t)$, which is why we have Mistaken Notion 4.3. That is why for many people, this idea is in the "I find that hard to believe" category. Indeed, the phenomenon of carrier waves is not visible in any numeric computation of the ζ-function, because the scale of the carrier waves, just like the scale of $S(t)$, grows so slowly that it appears bounded within the range we can compute. However, we can take a first-principles approach to defining what we mean by "carrier wave", and then build intuition by looking at illustrative examples.

5.1 The wave as a local constant factor

Suppose $f(t)$ is a high degree polynomial, with zeros $\gamma_1 < \ldots < \gamma_M$, and suppose we want to understand the graph of f near t_0. Further suppose that t_0 is roughly near the middle of the γ_j. We can write

$$f(t) = a_0 \prod_{\gamma_j \text{ near } t_0} \left(1 - \frac{t}{\gamma_j}\right) \prod_{\gamma_j \text{ far from } t_0} \left(1 - \frac{t}{\gamma_j}\right) \tag{5.1}$$

$$\approx A_0 \prod_{\gamma_j \text{ near } t_0} \left(1 - \frac{t}{\gamma_j}\right) \qquad \text{for } t \text{ very close to } t_0. \tag{5.2}$$

The approximation in the second line above will be valid for t in a neighborhood of t_0 if the zeros far from t_0 are balanced on either side of t_0, meaning that the product over "γ_j far from t_0" is approximately constant near t_0.

L-functions are much like high degree polynomials, and the global spacing of their zeros is very regular (as can be seen from the small error term in the zero counting function $N(T)$, see (3.9)). Thus we have:

Principle 5.1 *The behavior of $Z(t)$ near t_0 depends on three things: a global factor independent of t_0, the arrangement of the zeros near t_0, and a scale factor which depends on the zeros far from t_0 and which does not change too quickly as a function of t_0.*

Figure 5.3 illustrates some of the ideas in Principle 5.1.

Principle 5.1 does not specify which of the two factors that depend on t_0 are most relevant to the size of $Z(t)$. Any graph of $Z(t)$ in the range accessible by current computers, such as in Figure 3.4, makes it appear that the arrangement of the zeros is more important. We will see that those examples are misleading.

5.2 Carrier waves

The scale factor $A_0 = A_0(t_0)$ has been termed the **carrier wave** by Hejhal [58] and Bombieri and Hejhal [20], making rigorous a speculation of Montgomery. Montgomery's preliminary calculations suggested that the carrier wave for the Riemann zeta-function should not vary too much over a window of width $\exp(\delta_T \log\log T)/\log T$, for some function $\delta_T \to 0$. Bombieri and Hejhal [20]

show that, for most T, the carrier wave does not vary significantly over a window of width $M/\log T$, for any fixed $M > 0$, as $T \to \infty$. That is: across a span of M consecutive zeros, for any fixed M, the size of $\log|\zeta(\frac{1}{2}+it)|$ usually varies very little. (The proof in [20] might actually show that one can take $M = \log\log(T)^\kappa$ for any $\kappa < \frac{1}{4}$, with (6.21) in that paper providing the limiting constraint, but those details would need to be checked.) See Subsection 11.3 for a discussion of how wide is the carrier wave.

Bombieri and Hejhal used carrier waves as the key idea toward their proof of the following surprising theorem: if $L_1(s)$ and $L_2(s)$ are L-functions which individually satisfy RH and have the same functional equation, and $\alpha \in \mathbb{R}$, then $L_1(s) + \alpha L_2(s)$ has 100% of its zeros on the critical line (there are technical conditions which we have omitted). The proof is: the $\log(L_j(\frac{1}{2}+it))$ are independently and normally distributed with a large variance, so most of the time one of the L_j is much larger than the other, and furthermore (because of the carrier wave) it stays larger across an arbitrarily many zeros. Therefore in that region the zeros of the linear combination are very close to the zeros of the larger L-function, and that accounts for most of the zeros. (The large variance is an important ingredient: given two Gaussians with variance σ^2, the probability that they differ by less than $\sqrt{\sigma}$ is $O(\sigma^{-\frac{1}{2}})$. The above argument requires $\sigma \to \infty$.)

Note that Selberg's theorem on the normal distribution of $L(\frac{1}{2}+it)$ is not sufficient: one needs the additional fact that the carrier wave causes the larger L-function to stay large over a significant range.

Thus we have:

Principle 5.2 *The carrier wave is responsible for the bulk of the value distribution of an L-function, with the variation due to the local zero distribution playing a secondary role. In particular, it is the carrier wave which obeys Selberg's central limit theorem.*

Principle 5.2 explains why, in the range accessible by current computers, the observed value distribution of $\log|Z(t)|$ departs significantly from normal. The local zero spacing contributes a lesser (typically, bounded) amount. But a bounded amount is significant in the range where the carrier waves are very small.

5.3 Measuring the wave

We now describe a way to measure and observe the carrier wave. The goal is to isolate the contribution of the nearby zeros, as suggested in (5.2). The idea is to think of the zeros as parameters which can change: we can slide the zeros side-to-side, and this will cause a change in the graph of $Z(t)$. (This is easier to picture if all the zeros are real, which we will assume for the purposes of this thought experiment.) If we slide the zeros apart, making a large zero gap, then the function will acquire a large local maximum. If we slide the zeros so they become more equally spaced, then the maxima of the function will be approximately the same size. In the extreme case of moving the zeros to be equally spaced, the result will be the (scaled and shifted) cosine function.

It is not possible to achieve a *globally* equal spacing for the zeros, because the local average spacing of the zeros is not constant. Instead we focus on "nearby" zeros, where the average spacing is close to constant. Suppose $t_0 \in [\gamma_M, \gamma_{M+1}]$. If K is large enough and we slide the zeros $\gamma_{M-K}, \ldots, \gamma_{M+K}$ to be equally spaced, then near t_0 the resulting function will look like $a\cos(b(x+c))$ for some $a, b, c \in \mathbb{R}$.

We will run the above process in reverse. That is, start with $a\cos(b(x+c))$, pick a region of interest in the critical strip containing zeros $\gamma_{M-K}, \ldots, \gamma_{M+K}$,

and then "move" the cosine zeros to the ζ-zeros. This suggests that

$$Z(t) \approx A_Z\, a \cos(b(t+c)) \prod_{j=M-K}^{M+K} \frac{t-\gamma_j}{t-g_j}, \tag{5.3}$$

where the g_j are zeros of $\cos(b(t+c))$, indexed so that g_j is close to γ_j, the parameters a, b, c depend on t_0 and K, and A_Z is a normalization factor. The approximation should be good near t_0 if K is large enough and t_0 is near the middle of the interval $[\gamma_{M-K}, \gamma_{M+K+1}]$. This approximation works because locally the Z-function is determined by the nearby zeros and a scale factor.

A similar approximation was considered by Hiary and Odlyzko ([60], Section 6), who refer to it as "HP", because essentially it uses a partial Hadamard Product for the approximation. They find that the quality of the approximation improves linearly with the number of zeros in the product. A difference in the approach here is to recognize that the zeros far from t_0 are basically contributing a multiplicative constant near t_0.

We apply (5.3) to $Z(t)$ from Figure 3.1, with

$$\begin{aligned} t_0 &= 5448 \\ M &= 5000 \\ \gamma_M &\approx 5447.8619 \\ K &= 10 \\ A_Z &= 2 \\ a &= 1.0991 \\ b &= 3.3825 \\ c &= 0.17008. \end{aligned} \tag{5.4}$$

The parameter $A_Z = 2$ is explained after Note 5.4, as is the fact that b and c are actually simple functions of t_0. The parameter a is chosen so that the approximation is exact at t_0. So, once t_0 is selected, the only choice is K, the number of zeros to be matched on either side of t_0. The result is shown in the top plot in Figure 5.3, where we superimpose $Z(t)$ and its approximation based at $t_0 = 5447.86$ using 20 matched zeros.. The bottom plot shows the same idea based at $t_0 = \gamma_{7010} \approx 7273.70$, which has scale factor $a = 0.9130$. In both cases one sees that the behavior near t_0 is determined by the nearby zeros and a local scale factor (which depends on t_0).

In the top plot in Figure 5.3 we can interpret the scale factor $a = 1.0991$ as the magnitude of the carrier wave at $t_0 = 5447.86$, and similarly for the scale factor $a = 0.9130$ at $t_0 = 7273.70$. Those scale factors are different, which shows that the magnitude of $Z(t)$ is not solely due to the local zero spacing and a single global factor.

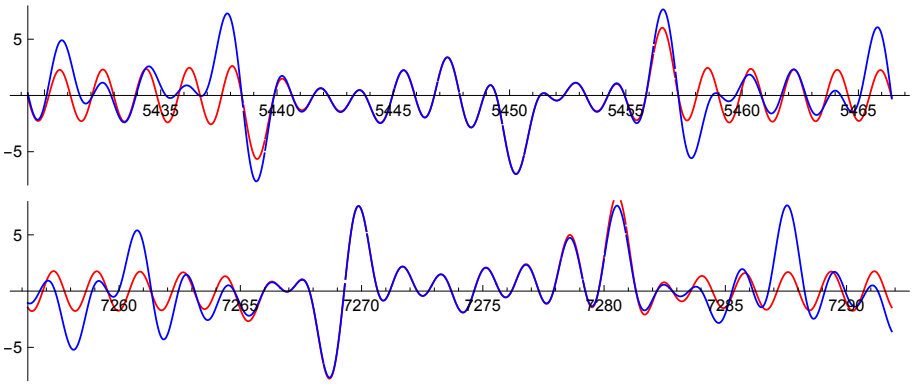

Figure 5.3 The function $Z(t)$ together with a local approximation as in (5.3) with $K = 10$. The top graph has t near 5447.86 and the local scale factor is 1.0991. The bottom graph has t near 7273.70 and the scale factor is 0.913.

Here we are focusing on (5.3) because it explicitly separates the contribution of the local zeros from the contribution of the carrier wave. There are other ways to approximate the ζ-function, see for example Subsection 10.2.

5.4 Some caveats

The scale factors in Figure 5.3 are not the carrier wave in the sense of Bombieri-Hejhal, because Montgomery's heuristic calculation suggests that at the small height of that example the carrier wave should only be approximately constant in a very narrow window. In Figure 5.3 we have fit to the zeros in a much wider window. Nevertheless, those plots illustrate that the nearby zeros and a single local scale factor determine the local behavior of $Z(t)$.

Note 5.4 The scale factor at a point is not well-defined: it depends on K, the number of zeros, or more precisely the width of the window, over which one has done the fit. If the window covers an enormous number of zeros then the local scale factor will just be 1, because of the long-range rigidity of the zeros; see Subsection 8.2. If the function changes scale in the window, i.e., the carrier wave is not approximately constant in the window, then the scale factor is not providing useful information. This is discussed further in Note 8.6.

It remains to justify that the global scale factor (independent of t_0) is $A_Z = 2$. That comes from the overall factor of 2 in the Riemann-Siegel formula (4.1).

The parameters b and c in (5.3) are also extraneous: instead of $A_Z \cos(bt+c)$ we could use the first term in the Riemann-Siegel formula:

$$2\cos(\theta(t)).$$

Indeed $\theta'(5447.86) \approx 3.38255$, which is the value for b in (5.4). Using $2\cos(\theta(t))$ in (5.3) does not change the point illustrated by Figure 5.3.

In order to really "see" the carrier wave, one must go to enormously larger values of t. That is not computationally feasible in the L-function world, but it is easy in the random matrix world. In Section 6 and Section 7 we review the connections between the ζ-function and the characteristic polynomials of random unitary matrices, returning to carrier waves in Section 8.

5.5 The highest tone

Many observations about the zeta-function can be traced to the fact that $S(t)$, the error term in the zero counting function $N(t)$, grows very slowly and is zero

on average. Here we collect some additional consequences.

Principle 5.5 *The Z-function, interpreted as a sound wave made up of separate tones, has a highest frequency component which is the loudest tone and which is separated in frequency from the lower tones. In particular, the Z-function is (approximately) a band-limited function with discrete support.*

Specifically, at height t the highest frequency is $\frac{1}{2}\log(t/2\pi)$, the factor of $1/2$ coming from the fact that sin or cos have two zeros in each period. Each subsequent frequency is lower by $\frac{1}{2}\log(2)$, $\frac{1}{2}\log(3)$, ..., and its contribution is smaller by a factor $1/\sqrt{2}$, $1/\sqrt{3}$,.... These observations come straight from the Riemann-Siegel formula (4.1), and are illustrated in Figure 5.6, which shows the Fourier transform of $Z(t)$ for $t \approx 10^6$. Note that $\log(10^6/2\pi) \approx 11.978$.

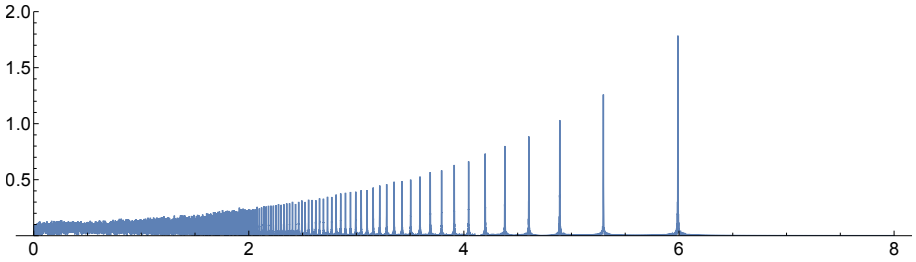

Figure 5.6 The Fourier transform of $Z(t)$, computed numerically, for $10^6 \le t \le 10^6 + 5000$.

The Z-function is not actually band-limited, a fact which is relevant to the phenomenon of superoscillations, see [10, 11]. However, it is close enough to band-limited that band-limited interpolation has become an indespensible tool in large-scale computations of the zeros [79, 17].

Principle 5.7 *The highest tone of the Z-function comes from the* Γ*-factor in the functional equation, and the lower tones come from the Dirichlet coefficients. The* Γ*-factor places the zeros in precise well-spaced locations, and then the Dirichlet coefficients make adjustments.*

To see that interpretation, view the Riemann-Siegel formula (4.1) as arising from the expression $Z(t) = X(\frac{1}{2}+it)^{-1/2}\zeta(\frac{1}{2}+it)$ for the Z-function in terms of the zeta-function, see (3.6), combined with the **approximate functional equation**, (4.12.4) of [101],

$$\zeta(s) = \sum_{1\le n\le x} \frac{1}{n^s} + X(s) \sum_{1\le n\le y} \frac{1}{n^{1-s}} + \text{ correction terms}, \tag{5.5}$$

where $xy = t/2\pi$. Setting $x = y$, combining terms, and using Stirling's formula for the Γ-functions in $X(\frac{1}{2}+it)$ yields the Riemann-Siegel formula. The $n = 1$ terms contribute $2\,\Re X(\frac{1}{2}+it)^{-1/2}$, which is bounded by 2 and has regularly spaced zeros with the same counting function as the zeros of the zeta function (but with error term bounded by 1). Those zeros interlace the Gram points. We see that Gram's law is a consequence of a more general principle:

Principle 5.8 The first term dominates at low height. *When the analytic conductor of an L-function is small, the first term in the Riemann-Siegel formula has a strong influence on the location of the critical zeros.*

We used the term "analytic conductor" so that the principle applies more widely. For the ζ-function at $s = \frac{1}{2}+it$, the analytic conductor is proportional to $\log(2+|t|)$. Figure 5.9 illustrates the principle by graphing the ζ-function and and the first term in its Riemann-Siegel formula, over the interval $[1000, 1050]$. The Gram points are the locations of the local maxima and local minima of the

red curve which oscillates between -2 and 2. Gram's law (see Subsection 14.4 for further discussion) follows from the expectation that at low height, the zeros of the ζ-function will be close to the zeros of the first term.

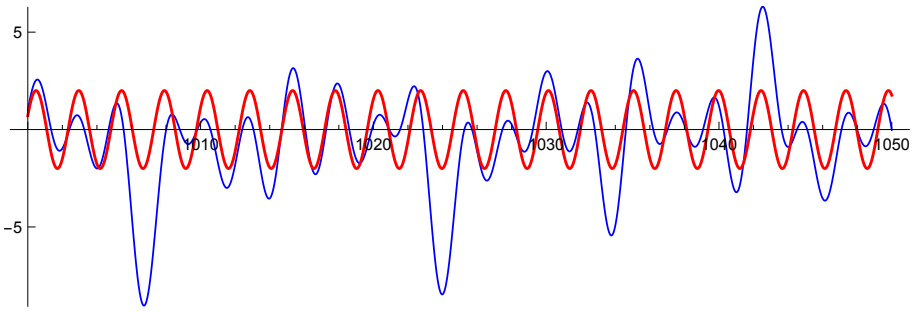

Figure 5.9 The functions $Z(t)$ and $2\cos(\vartheta(t))$ on the interval $1000 \le t \le 1050$.

There is a Riemann-Siegel formula for an L-function of degree d, with the modification that xy is of size t^d, the numerators of the two halves of the approximate functional equation are a_n and $\overline{a}_n$, respectively, and $X(s)$ is replaced by $\varepsilon X(s)$. For large t there is highest frequency term, with frequency asymptotically $\frac{d}{2}\log t$, involving only the data in the functional equation. And just like in the classical case, initially the Dirichlet coefficients make small adjustments to the location of the zeros arising from the Γ-factors. The Γ-factors have a surprising influence on the initial zeros, particularly for nonarithmetic L-functions, whose trivial zeros have nonzero imaginary part. That influence can be captured by working directly with $X(\frac{1}{2}+it)^{-1/2}$ instead of using the asymptotics from Stirling's formula. See [47].

6 A random model for the ζ-function

A graph of $Z(t)$ gives the impression of randomness: the function wiggles, and there is no apparent pattern to those wiggles. Obviously $Z(t)$ is not random, because it is a specific function which has no randomness in its definition. But we can make a random function in the following way. Fix an interval $\Upsilon = [t_0, t_1] \subset \mathbb{R}$. If $T \in \mathbb{R}$ is random, then $Z_T(t) := Z(T+t)$ is a random function on Υ.

If we had another set of random functions $\mathcal{Z}_T$ on Υ, and it was possible to prove theorems about $\mathcal{Z}_T$, and if furthermore we had reason to believe that Z_T and $\mathcal{Z}_T$ had similar properties, then we could turn theorems about $\mathcal{Z}_T$ into conjectures about Z_T. That would be illuminating, particularly if precise conjectures about Z_T were in short supply. That is how random matrix theory made fundamental contributions to the study of L-functions.

6.1 Self-reciprocal polynomials and the functional equation

Before getting into the details, let's consider a certain class of polynomials which have constant term 1:

$$f(z) = 1 + a_1 z + a_2 z^2 + \cdots + a_{N-1} z^{N-1} + a_N z^N. \tag{6.1}$$

The polynomial f is **self-reciprocal** if

$$a_j = a_N \overline{a_{N-j}} \quad \text{for} \quad 0 \le j \le N, \tag{6.2}$$

or equivalently

$$f(z) = \mathcal{X}(z)\overline{f}(z^{-1}) \qquad \text{where} \qquad \mathcal{X}(z) = a_N z^N, \tag{6.3}$$

or equivalently

$$\mathcal{Z}_f(\theta) = \mathcal{X}(e^{i\theta})^{-\frac{1}{2}} f(e^{i\theta}) \quad \text{is real if} \quad \theta \in \mathbb{R}. \tag{6.4}$$

An important consequence of any of these conditions is that

$$f(re^{i\theta}) = 0 \qquad \text{if and only if} \qquad f(r^{-1}e^{i\theta}) = 0. \tag{6.5}$$

In (6.3), $\overline{f}$ is the **Schwarz reflection** of f: $\overline{f}(z) := \overline{f(\overline{z})}$. The parallel with the functional equation for the ζ-function is evident if one compares (6.3) to (3.3), and (6.4) to (3.6). At the risk of belaboring the point, in both cases there is an involution of the plane, the set of zeros is fixed by the involution, the **Riemann Hypothesis** is the assertion that every zero is fixed by the involution, and there is a Z-function which is real on the set of points which are fixed by the involution. The unit circle is the analogue of the critical line for self-reciprocal polynomials. Thus, if a collection of polynomials is claimed to be a random model for the ζ-function, those polynomials must be self-reciprocal.

6.2 Random polynomials and trigonometric polynomials

We seek random self-reciprocal polynomials which can serve as a model for $\zeta(s)$ or equivalently $Z(t)$. How to choose the randomness? There are three reasonable options.

The first option is to choose the coefficients of $f(z)$ so that they satisfy (6.2). The second is to recognize that Z_f is a trigonometric polynomial (i.e. a finite Fourier series), so its coefficients can be chosen randomly with only the requirement that they are real. The third option is to choose the zeros so that they satisfy (6.5). We will argue that only the third of those options is reasonable.

Figure 6.1 shows $Z(t)$ at height 1.5×10^{30}, in a region where nothing special is happening. (Data courtesy of Bober and Hiary [17][59].)

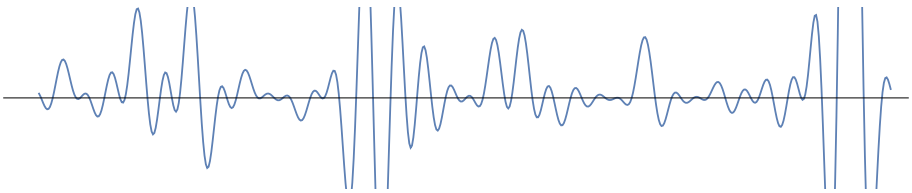

Figure 6.1 $Z(t)$ on an interval of width 2π near 1.5×10^{30}

We wish create random polynomials which capture the type of randomness visible in Figure 6.1. But what does that mean? The following features are worthy of attention.

(A) Do all the zeros appear to be real?

(B) How often does the graph get big? Almost equivalently, how often are there large gaps between zeros?

(C) How often are pairs of zeros close together? How often are there clusters of close together zeros?

(D) Are the maxima and minima generally around the same size, with only occasional much larger or smaller local extrema?

Item D is obviously a leading question. The answer is "yes".

Item A refers to RH. We can't escape the fact that RH is true within the realm where we can experiment, so the random polynomials we seek must satisfy RH. That is, we seek **unitary polynomials**, which is standard terminology for polynomials with all zeros on $|z| = 1$.

The requirement of generating unitary polynomials effectively rules out the first two options for choosing random polynomials. If the coefficients are chosen randomly, then one can choose small coefficients to obtain a perturbation of $z^N - 1$. The zeros will be on the unit circle, but they will be close to regularly spaced and the function will not have randomness similar to Figure 6.1. If the coefficients are random and large, then the polynomial will usually not be unitary (although it may have a large proportion of its zeros on the unit circle). Choosing the coefficients randomly and then conditioning on having all zeros on the unit circle turns out to produce the wrong type of randomness [44]. (It gives the COE, not the CUE.)

By process of elimination we conclude that the polynomials we seek must arise by having their zeros randomly generated, and on the unit circle. Such polynomials have arisen in mathematical physics, which we describe next.

Note 6.2 It would be interesting to define a natural-looking ensemble of random self-reciprocal polynomials which are unitary most of time, but not always. If such an ensemble was consistent with RH being true within the realm of current computation, and it predicted a height at which RH would start to fail, that could throw doubt on RH. Indeed, one of the great shortcomings of RH skepticism is the lack of conjectures for: the height at which RH fails, how often zeros would be off the line, and how their real parts would be distributed. Principle 2.2 implies that even if RH fails, 100% of the zeros are on the critical line.

6.3 The circular β-ensembles

The **circular β-ensemble** consists of N random points $e^{i\theta_1}, \dots, e^{i\theta_N}$ on the unit circle, equipped with the joint probability density function

$$\frac{\Gamma(1+\beta/2)^N}{\Gamma(1+N\beta/2)(2\pi)^N} \prod_{1\le j<k\le N} |e^{i\theta_j} - e^{i\theta_k}|^\beta. \tag{6.6}$$

This probability space is denoted CβE(N). We use $\langle f(\theta_1, \dots, \theta_N)\rangle$ for the **expected value** of f averaged over CβE(N).

The CβE(N) is defined for any $\beta \ge 0$ and any positive integer N. We can obtain a random unitary (hence, self-reciprocal) polynomial by taking the points to be the zeros of a polynomial with constant term 1, as in (6.1).

If we write the θ_j in increasing order, then it is clear that $\langle \theta_{j+1} - \theta_j\rangle = 2\pi/N$. Almost equally clear is that the PDF of $\theta_{j+1} - \theta_j$ vanishes to order β at 0. For that reason β is referred to as the **order of repulsion** between the points.

As $\beta \to 0$ the repulsion disappears and the points become independently uniformly distributed on the circle. As $\beta \to \infty$ the points approach equal spacing (known as the **picket fence** distribution), with the only remaining randomness being the angle of an initial point which determines the location of all the others. Thus, as $\beta \to 0$ the particles behave like a gas, and as $\beta \to \infty$ they crystallize. For this reason, β is sometimes referred to as **inverse temperature**.

Between those temperature extremes there are three values which have been studied extensively in the mathematical physics literature: $\beta = 1$, 2, and 4. Those are more commonly known as the COE, CUE, and CSE, where the

O/U/S stand respectively for Orthogonal/Unitary/Symplectic. Examples of random polynomials $\mathcal{Z}(\theta)$ from each of those ensembles are shown in Figure 6.3. Note that the graphs all have the same vertical scale, ranging from -6 to 6. The graphs show the effect of β on the frequency of small zero gaps, large zero gaps, and large values. Those examples have $N = 68$, that choice ensuring the polynomials have the same number of zeros in a span of 2π as the Z-function shown in Figure 6.1. Our next goal is to persuade that one of those ensembles provides a good model for the ζ-function.

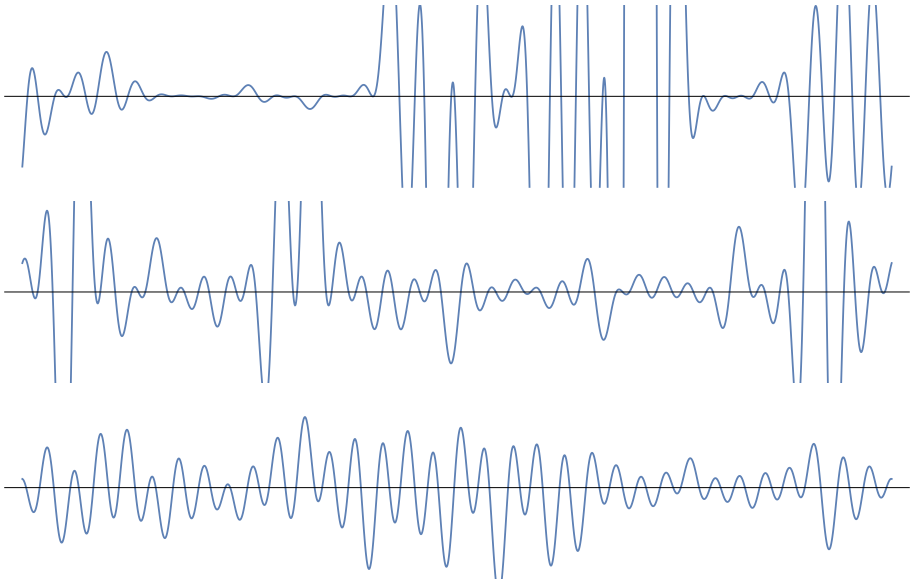

Figure 6.3 Example random polynomials from CβE(68) with, reading top to bottom, $\beta = 1, 2$, and 4.

Hopefully one of the plots in Figure 6.3 looks like it has similar randomness to the Z-function in Figure 6.1, because otherwise we are out of ideas. The bottom plot, $\beta = 4$, is ruled out in multiple ways. The 4th order repulsion causes the zeros to be too rigidly spaced, which in turn prevents the function from having sufficiently many large or small values. Polynomials from the CSE are pleasant but rather boring, lacking the pizzazz of the Riemann ζ-function.

The top plot in Figure 6.3 can be ruled out by carefully examining the features of the function. The most obvious property is that the COE example is large much more often: consider the width of the cutoff at the large maxima. The COE example also has more small zero gaps, both individually (small local extrema) and in clusters.

The CSE is too rigid to serve as a model for the ζ-function, and the COE is too flexible. The Goldilocks zone is occupied by the CUE which is conjectured to provide, with appropriate adjustments and caveats, a model for the zeros of the ζ-function. We clarify the connection, and provide more terminology and historical details, in Section 7.

Warning 6.4 The CUE has the same distribution as the eigenvalues of Haar-random matrices from the compact unitary group $U(N)$. But the COE and CSE, and the GOE, GUE, and GSE described in Subsection 7.1, do not correspond to Haar measure on classical matrix groups with similar names. The eigenvalues of Haar-random matrices from the unitary symplectic and unitary orthogonal groups all have quadratic repulsion for the bulk of their eigenvalues, with different behavior for the eigenvalues close to 1. These play a role in modeling the low-lying zeros of certain families of L-functions, see [70, 87].

6.4 Where does the CUE come from?

To do calculations involving the CUE, such as computing an expectation $\langle f \rangle$, all one needs is the joint law of the zeros: (6.6) with $\beta = 2$. That is, it does not matter how the points are generated. But to construct an example, such as the plots in Figure 6.3, one needs a way to produce actual sets of points with the appropriate distribution. Here are some ways:

1. *Brownian motion.*

   Have N points start at the origin and undergo non-intersecting Brownian motion until they hit the unit circle. Those points will be distributed according to (6.6). (There are Brownian motion models for the CUE and other β-ensembles. See [38].)

2. *A product of tridiagonal matrices.*

   Let Ξ_k be certain independent random 2×2 matrices, depending on a parameter β, defined in [72]. Let

   $$L = \mathrm{diag}(\Xi_0, \Xi_2, \dots, \Xi_{\lfloor N/2 \rfloor}) \quad \text{and} \quad M = \mathrm{diag}(\Xi_{-1}, \Xi_1, \dots, \Xi_{\lfloor N/2 \rfloor - 1}).$$

   The eigenvalues of the tridiagonal matrix LM, equivalently ML, will be distributed according to the CβE(N).

3. *The unitary group $U(N)$.*

   Let $A \in U(N)$ be chosen randomly with respect to Haar measure. The eigenvalues of A will be distributed according to (6.6) with $\beta = 2$.

The final option is the easiest to implement, but see Mezzadri's paper [76] which describes possible pitfalls.

To summarize:

Principle 6.5 *The connection between random matrices and the local statistics of zeros of the ζ-function is not actually about random matrices: all that really matters is the $\beta = 2$ repulsion (6.6) on the zeros/eigenvalues. Random matrices, either Haar-random from $U(N)$ or products of tridiagonal matrices [72], just happen to be convenient ways to generate such distributions.*

However, for historical reasons, and for lack of a good alternate term, it is common to refer to the N points in the CβE(N) as "eigenvalues", even if they do not arise from a matrix.

The description we have given is incomplete and our claim of a connection between the CUE and the ζ-function is perhaps not persuasive, so in Section 7 we provide more details about random matrix eigenvalues, random unitary polynomials, and the connection with L-functions. Then in Section 8 we return to our main theme and explore carrier waves in random polynomials from the CUE.

7 RMT and L-functions: terminology and history

The CβE were not how Random Matrix Theory (RMT) first arose in mathematical physics. RMT came about in the 1950s as a way to understand complicated quantum mechanical systems. For example: the energy levels of a large atomic nucleus. The justification is that the Hamiltonian of the system is a large and complicated matrix with certain symmetries, and so the eigenvalues of the Hamiltonian should have similar statistical properties to a large random matrix with the same symmetries.

The connection with L-functions arose in three phases.

1. *The* GUE *era.*

   The original formulation in the early 1970's concerned the limiting local statistics of eigenvalues/zeros. By *limiting* we refer to the rescaled eigenvalues in the large matrix limit. The matrices are Hermitian so the eigenvalues are real. See Subsection 7.1 for more details.

2. *The classical compact group era.*

   In the late 1990's it was realized that certain compact unitary groups of matrices could be associated with families of L-functions. The size of the matrices is a natural function of the conductor of the L-functions (so one could model at finite height), the particular matrix group is determined from properties of the L-functions, and the characteristic polynomial can be used to model the values of the L-functions. See Subsection 7.3 for more details.

3. *The recipe era.*

   In the early 2000's a heuristic, inspired by RMT but not actually making use of random matrices, was developed which reproduces virtually every conjecture arising from RMT. In particular, the heuristic for ratios can be used to obtain all correlation functions of the zeros. See [33]. Furthermore, the heuristic produces the complete main term and not just the leading order behavior. The heuristic goes by the name "the recipe". See the end of Subsection 7.5 for a brief mention.

Despite the fact that the recipe has largely replaced RMT as a tool for making conjectures about L-functions, random matrices are essential for our discussion of carrier waves, because characteristic polynomials provide concrete examples which we can visualize.

7.1 The GUE era

The matrices in this setting are Hermitian, so the eigenvalues are real. The random entries are real, complex, or quaternionic depending on the type of physical system being modeled, and the randomness is Gaussian. As in the circular ensembles, the three Gaussian ensembles are named GOE, GUE, and GSE, where again O/U/S stands for Orthogonal/Unitary/Symplectic. Again there is a single parameter β which describes the eigenvalue repulsion, with $\beta = 1$, 2, or 4 for each of those cases, respectively. And again the way the eigenvalues (or energy levels) are generated could be Brownian motion [38], or tridiagonal matrices [39], or the original formulation of random Hermitian matrices with a given symmetry. The only concern for our purposes is that one has N points on the real line with joint probability density function:

$$\frac{\Gamma(1+\beta/2)^N}{(2\pi)^{N/2}\prod_{j=1}^{N}\Gamma(1+j\beta/2)}\exp\bigl(-\tfrac{1}{2}\sum_{j=1}^{N}\lambda_j^2\bigr)\prod_{1\le j<k\le N}|\lambda_j-\lambda_k|^{\beta}. \tag{7.1}$$

A similarity with the CβE is the order β repulsion between points. A difference is that the points are on a line, with the factor $\exp(-\frac{1}{2}\lambda_j^2)$ keeping more of the points near the origin and preventing the points from wandering too far away. There is a tension between the factor keeping the eigenvalues close to the origin and the order β repulsion which tries to keep them apart. The result is that $|\lambda_j| < \sqrt{2\beta N}$ for most j and the rescaled points $\lambda_j/\sqrt{\beta N}$ are distributed according to the **Wigner semicircle law** $\pi^{-1}\sqrt{2-x^2}$.

The similarities outweigh the differences. The GβE and CβE belong to the same **universality class** which is characterized only by order β repulsion between points. Let's be specific about which properties are universal. Define the rescaled eigenvalues $\tilde{\lambda}$ so that $\langle \tilde{\lambda}_{j+k} - \tilde{\lambda}_j \rangle = k$ for all fixed k. Consider any **local statistic**, that is, a statistic only involving a finite number of differences of rescaled eigenvalues. Examples of local statistics are the normalized nearest neighbor spacing

$$p_2(x) = \lim_{\Delta x \to 0} \frac{1}{\Delta x} \left\langle \frac{\#\{\tilde{\lambda}_{j+1} - \tilde{\lambda}_j \in [x, x + \Delta x]\}}{N} \right\rangle, \tag{7.2}$$

the pair correlation function

$$R_2(x) = \lim_{\Delta x \to 0} \frac{1}{\Delta x} \left\langle \frac{\#\{\tilde{\lambda}' - \tilde{\lambda} \in [x, x + \Delta x]\}}{N} \right\rangle, \tag{7.3}$$

or more generally the joint distribution of $(\tilde{\lambda}_{j_1} - \tilde{\lambda}_{j_0}, \ldots, \tilde{\lambda}_{j_k} - \tilde{\lambda}_{j_0})$ for any j_0 and $j_1 < \ldots < j_k$. Note that the focus is on the relationship between an eigenvalue and its neighbors, not on their absolute location.

The quantities above depend on N, but we suppress that notation because the leading order behavior is independent of N.

Principle 7.1 Universality. *For each β, all β-ensembles have the same limiting local eigenvalue statistics as $N \to \infty$.*

Principle 7.1 comes from the fact that the β-repulsion term is the most important feature of the measure. Here is a less hand-wavy explanation in the case of the CUE(N) and the GUE(N). Both ensembles are examples of a **determinental point process**. This means there is a **kernel function** $K_N(x, y)$ such that the n-correlation function of the eigenvalues can be expressed as the determinant of an $n \times n$ matrix with entries involving $K_N(x, y)$. For example, the pair correlation function is given by

$$R_{2,N}(x - y) = \begin{vmatrix} 1 & K_N(x, y) \\ K_N(y, x) & 1 \end{vmatrix}. \tag{7.4}$$

Thus, the kernel function determines all the correlation functions, so it also determines all the local statistics (because those can be expressed in terms of the correlation functions, possibly involving a complicated inclusion-exclusion argument).

For GUE(N) the kernel function is

$$K^{\mathrm{GUE}(N)}(x, y) = \sqrt{N} \frac{\psi_N(x)\psi_{N-1}(y) - \psi_{N-1}(x)\psi_N(y)}{x - y}, \tag{7.5}$$

where $x, y \in \mathbb{R}$. Here $\psi_n(x) = He_n(x)e^{-x^2/4}/\sqrt{\sqrt{2\pi}n!}$, where $He_n(x)$ is a Hermite polynomial. Change variables $(x, y) \mapsto (4x/\sqrt{N}, 4y/\sqrt{N})$ for the normalized neighbor spacing.

For CUE(N) the kernel function is

$$K^{\mathrm{CUE}(N)}(\theta, \phi) = \frac{1}{N} \frac{\sin(N(\theta - \phi)/2)}{\sin((\theta - \phi)/2)}, \tag{7.6}$$

where $\theta, \phi \in \mathbb{R} \mod 2\pi$. Change variables $(\theta, \phi) \mapsto (2\pi\theta/N, 2\pi\phi/N)$ for the normalized neighbor spacing. To leading order those kernels are the same, because by properties of Hermite polynomials, both (normalized) large N limits are $\sin(\pi x)/(\pi x)$.

The difference between the kernel functions of the GUE(N) and CUE(N) is $O(1/N)$, so the same holds for all correlation functions. Figure 7.2 shows the pair correlation for each ensemble for $N = 50$, and their difference (with the GUE(N) pair correlation being larger in a neighborhood of 0).

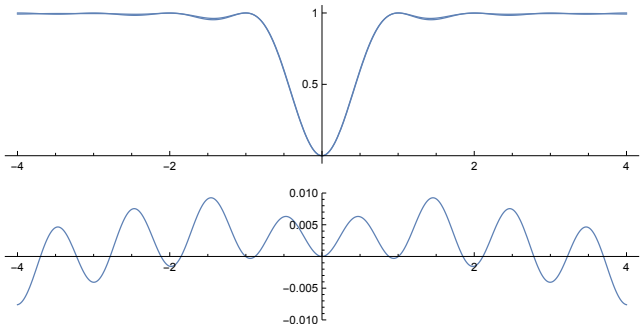

Figure 7.2 Pair correlation for GUE(50) and CUE(50), and their difference.

Note that by (7.4) and the limiting values of the kernel functions for the GUE and CUE, we find that both have the limiting normalized pair correlation function:

$$R_2(x) = 1 - \frac{\sin^2(\pi x)}{(\pi x)^2}. \tag{7.7}$$

That pair correlation function was the key to discovering the connection between random matrix eigenvalues and the zeros of the ζ-function.

7.2 Zero statistics for L-functions

The RMT revolution in number theory began when Montgomery [77] determined partial information about the pair correlation of the zeros of the ζ-function. The theorem he proved, combined with some heuristics about the prime numbers, led him to conjecture that asymptotically the pair correlation function was

$$R_{2,\zeta}(x) = 1 - \frac{\sin^2(\pi x)}{(\pi x)^2}. \tag{7.8}$$

Here

$$R_{2,\zeta}(x) := \lim_{\Delta x \to 0} \frac{1}{\Delta x} \lim_{T \to \infty} \frac{1}{N(T)} \sum_{0 < \gamma, \gamma' \le T} \#\{\tilde{\gamma}' - \tilde{\gamma} \in [x, x + \Delta x]\} \tag{7.9}$$

is the ζ-function analogue of (7.3). Montgomery's conjecture became more significant when he met Freeman Dyson, who informed him that (7.8) was the limiting pair correlation of eigenvalues of the GUE, as shown in (7.7). Combined with extensive computations by Odlyzko [81], this established:

Principle 7.3 The GUE **Hypothesis, aka the Montgomery-Odlyzko law.** *The limiting local statistics of the zeros of the ζ-function are the same as the limiting local statistics of the* GUE.

The "limiting statistics" above refer to $T \to \infty$ and $N \to \infty$.

Consequences of the GUE **Hypothesis.** The GUE Hypothesis has many useful consequences:

1. Montgomery's pair correlation conjecture (7.8).
2. 100% of the zeros of the ζ-function are simple.

3. The PDF $p_2(x)$, see (7.2), of the nearest neighbor spacing $\tilde{\gamma}_{j+1} - \tilde{\gamma}_j$ is given by an explicit expression, which is well-approximated in the bulk by the **Wigner surmise**

$$\tfrac{32}{\pi^2} x^2 e^{-\frac{4}{\pi}x^2}, \tag{7.10}$$

which is the exact expression for GUE(2). The limiting behavior of $p_2(x)$ for GUE(n) as $n \to \infty$ is $\pi^2 x^2/3$ as $x \to 0$ and $\exp(-\pi^2 x^2/8)$ as $x \to \infty$.

The next two items are consequences of those limiting behaviors.

4. The smallest gaps between consecutive zeros satisfy

$$\tilde{\gamma}_{j+1} - \tilde{\gamma}_j \sim \gamma_j^{-\frac{1}{3}+o(1)}, \tag{7.11}$$

that is,

$$\liminf_{j\to\infty} \frac{\log(\gamma_{j+1} - \gamma_j)}{\log \gamma_j} = -\frac{1}{3}. \tag{7.12}$$

It is possible to make a more precise statement. See Subsection 8.7.

5. The largest gaps satisfy

$$\tilde{\gamma}_{j+1} - \tilde{\gamma}_j \sim c_{large} \sqrt{\log \gamma_j}, \tag{7.13}$$

where perhaps $c_{large} = 1/\sqrt{32}$ [8]. Or perhaps c_{large} is a bit smaller. See Subsection 8.7.

Information about small gaps between zeros has implications for the class number problem [32], which was Montgomery's original motivation for studying the pair correlation function [77].

Shortcomings of the GUE **Hypothesis.** The predictions of the GUE Hypothesis are spectacular, but it was clear from the beginning that only part of the story was being revealed. Odlyzko's computations showed a discrepancy with the predictions. The discrepancy decreased at larger heights, as would be expected since the GUE Hypothesis only refers to the limiting behavior, but clearly there was a need for a more precise model.

Another shortcoming is that the prime numbers are nowhere to be seen in the GUE Hypothesis. How can it be that the truth about the ζ-function has nothing to do with the prime numbers?

Finally, and most relevant to this present work, the GUE Hypothesis says little or nothing about the actual values of the ζ function. This appeared in Odlyzko's calculations (which computed the values of the ζ-function and not just the zeros), which found a large discrepancy between the computed value distribution and the limiting Gaussian distribution from Selberg's central limit theorem. The convergence to Gaussian is slow, see Figure 7.5, so it was not surprising to see a discrepancy. But the nature of the discrepancy is not explained by the GUE Hypothesis. It took until the year 2000 to address that shortcoming.

7.3 After 30 years of GUE: classical compact groups and the Keating-Snaith Law

In the late 1990's two innovations increased the influence of RMT on number theory. The first was due to Katz and Sarnak [68] who found that naturally arising collections of L-functions have a **symmetry type** which is given by a classical compact matrix group. The possible symmetry types are Unitary,

Symplectic, and Orthogonal, with Orthogonal splitting into two cases depending on whether the sign of the functional equation is always $+1$ or equally likely $+1$ or -1. The random matrices in this context are elements of the compact unitary groups $U(N)$, $Sp(N)$, and $O(N)$, where the randomness is uniform with respect to Haar measure. Note that $U(N)$ with Haar measure is the same as the CUE(N), but the others are completely different than their mathematical physics counterparts having similar names. In fact, all of those classical compact groups have the same bulk eigenvalue statistics: the only difference arises with the eigenvalues near 1. Numerical evidence [87], and much subsequent work, supports the conjecture that the symmetry type of the family of L-functions determines the distribution of zeros near the critical point.

Keating and Snaith [69][70] made the second key leap when they recognized that the characteristic polynomial could be used to model the L-function itself. Here the **characteristic polynomial** of the $N \times N$ matrix A is written in the slightly nonstandard form

$$\begin{aligned} \Lambda_A(z) &= \det(I - zA) \\ &= \prod_{1 \le j \le N} (1 - z e^{i\theta_j}), \end{aligned} \tag{7.14}$$

which is in keeping with our normalization (6.1) for self-reciprocal polynomials. Here the $e^{i\theta_j}$ are the eigenvalues of A. The analogue of the Z-function is

$$\begin{aligned} \mathcal{Z}_A(\theta) &= ((-1)^N \det A)^{-\frac{1}{2}} e^{-iN\theta/2} \det(I - e^{i\theta} A) \\ &= ((-1)^N \det A\, e^{iN\theta})^{-\frac{1}{2}} \Lambda_A(e^{i\theta}), \end{aligned} \tag{7.15}$$

which is real for $\theta \in \mathbb{R}$. We write $\mathcal{S}_A$ for the error term in the zero counting function:

$$\#\{\theta_j \in [0, X]\} = \frac{1}{2\pi} X + C_A + \mathcal{S}_A(X)$$

where C_A is chosen so that $\mathcal{S}_A$ is 0 on average. The number C_A is analogous to the constant $7/8$ in the zero counting function of the ζ-function, see (3.9).

We confine our discussion to the case of the ζ-function, which constitutes a unitary family, meaning that $\zeta(\frac{1}{2} + it)$ is modeled by the characteristic polynomial of Haar-random matrices from $U(N)$.

Principle 7.4 The Keating-Snaith Law. *Model $\zeta(\frac{1}{2} + it)$ for $t \approx T$ by $\Lambda_A(e^{i\theta})$ for Haar-random $A \in U(N)$, where $N = \log(T/2\pi)$.*

A key component of Principle 7.4 is the identification $N = \log(T/2\pi)$, where the equality means "is an integer close to". This is justified by "equating the density of zeros". When calculating asymptotics, where the only realistic hope is to make leading-order predictions, one can take $N = \log T$. But for numerical comparison involving small numbers, the more precise choice of N is helpful. See Subsection 8.5 for further discussion of $N = \log T$.

The first great success of Principle 7.4 was explaining the apparent discrepancy between numerical data and Selberg's central limit theorem. Keating and Snaith proved a central limit theorem for $\log \Lambda_A(e^{i\theta})$, exactly analogous to Selberg's result. Odlyzko's data, taken at height 1.52×10^{19}, does not look Gaussian, and is not even symmetric. But analogous data for $\Lambda_A(e^{i\theta})$ for Haar-random $A \in U(42)$ gives a close fit with the ζ-function data. See Figure 1 in [69]. The "42" comes from $\log(1.52 \times 10^{19}/(2\pi)) \approx 42.33$. (Analogous to Principle 5.2, it is the carrier wave which obeys the Keating-Snaith central limit theorem.)

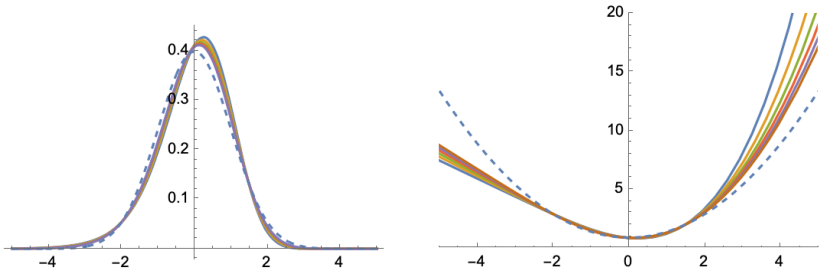

Figure 7.5 On the left is the PDF of $\log|\Lambda_A(e^{i\theta})|$ for random $A \in U(N)$ for $N = 25, 50, 100, 200, 400$, and 800, each normalized to have unit variance. On the right is the negative logarithm of those PDFs. In both cases the dashed line is the limiting Gaussian distribution.

The second great success came from predicting the mysterious factor g_k in the conjectured moments (13.2) of the ζ-function. Previously it was known that $g_1 = 1$ and $g_2 = 2$. It had been conjectured [29, 31] that $g_3 = 42$ and $g_4 = 24024$. What Keating and Snaith found was (they compute an exact expression, but only the asymptotic is relevant for this discussion):

$$\left\langle |\Lambda_A(e^{i\theta})|^{2k} \right\rangle_{U(N)} \sim g_{U,k} N^{k^2}, \tag{7.16}$$

as $N \to \infty$, where

$$g_{U,k} = k^2! \prod_{j=0}^{k-1} \frac{j!}{(j+k)!}. \tag{7.17}$$

They found the beautiful equality $g_{U,k} = g_k$ for the known and conjectured values for $k = 1, 2, 3, 4$.

We make explicit some of the assumptions in Principle 7.4.

Principle 7.6 The Keating-Snaith Law, part 2. *To model $\zeta(\frac{1}{2} + it)$ throughout the interval $[T, 2T]$, choose e^N Haar-random characteristic polynomials from $U(N)$, where $N \approx \log T/2\pi$.*

In other words, the ζ-function is treated as being independent on each separate interval of length 2π. One could argue that $e^N/2\pi$ random matrices is a better choice, but that makes no difference in practice. Principle 7.6 is one of the many ways to obtain the conjectured maximum values (3.13) and (3.12) of $Z(t)$ and $S(t)$ [43].

The success of the Keating-Snaith Law is our justification to use characteristic polynomials from $U(N)$, equivalently CUE(N), to explore the behavior of $Z(t)$ beyond the realm accessible by direct computation. We address some natural questions concerning the connection between RMT and the ζ-function before examining carrier waves in characteristic polynomials in Section 8.

Limitations of the Keating-Snaith law. The GUE Hypothesis is a statement about limiting behavior, and the Keating-Snaith law can be interpreted as a first order correction. Here we discuss the prospects of making a more precise version of the Keating-Snaith law.

It is obvious that random matrices cannot capture the subtle behavior of the ζ-function because the matrices do not know anything about the prime numbers. For example, when characteristic polynomials were first used to conjecture moments of the ζ-function, see (13.1), the arithmetic factor was inserted in an ad-hoc manner. A way to make the moment conjecture more natural is the "hybrid model" [52], which expresses the ζ-function as a product with two components: one involving the zeros and one involving the primes. See

Subsection 10.2. Assuming a statistical independence of those two components, the leading term in conjectured moments arises in a natural way.

It is natural to wonder if a more precise hybrid model could produce lower order terms in the conjectured moments? Since the zero statistics of the ζ-function and of characteristic polynomials agree only to leading order, a more precise model would either have to produce eigenvalue statistics which incorporate arithmetic effects, or the arithmetic factor would somehow have to compensate for the incorrect lower order terms in the factor involving the zeros. Neither option seems plausible. The difficulty is further compounded by the fact that the two pieces cannot be treated as independent — because by inspection the main terms in the conjectured moments do not factor as a product. We say a bit more about this in Subsection 10.2.

A more detailed hybrid model incorporating the interactions between the primes and the zeros would be interesting, but it seems unlikely that such a construction exists. Even without that, the recipe already provides precise information about the moments and the zeros of the ζ-function. In particular, we know a lot about the lower order corrections to the Keating-Snaith law. In every case we find that the lower order corrections have a dampening effect. For example, the arithmetic factor in the conjectured moments, a_k, is very small. Another example is that the spacing of zeros of the ζ-function is more rigid than the eigenvalues of random matrices, see Subsection 8.6. This has the effect of making extreme gaps, or dense clusters of zeros, slightly less likely. To summarize:

Principle 7.7 *Lower-order corrections to the Keating-Snaith law have a dampening effect: the statistical behavior of the ζ-function and its zeros tend to be slightly less extreme than the analogous quantities for random unitary matrices.*

7.4 Modeling $Z(t)$ vs. modeling $\zeta(s)$

Our original motivation was to model $Z(t)$, but our description of the Keating-Snaith law referred to modeling $\zeta(s)$. To clarify:

Principle 7.8 *Model $Z(t)$ for $t \in \mathbb{R}$ by $\mathcal{Z}_A(\theta)$. Model $\zeta(\frac{1}{2} + a + it)$ for $a \in \mathbb{C}$ by $\Lambda(e^{i\theta - a})$.*

Some consequences of this model arise from the symmetries in the CβE measure (6.6):

Lemma 7.9 *The CβE(N) is invariant under the operations*

$$\{\theta_j\} \leftrightarrow \{\theta_j + \theta\} \tag{7.18}$$

$$\{\theta_j\} \leftrightarrow \{-\theta_j\}. \tag{7.19}$$

Note that the GβE, (7.1), has a symmetry analogous to (7.19), but no translation invariance as in (7.18).

By (7.18), the expected value of expressions like $\mathcal{Z}_A(\theta)\mathcal{Z}_A(\theta+\alpha)$ or $|\Lambda_A(e^{i\theta})|^k$ are independent of θ, so one typically sets $\theta = 0$. See for example (7.20) and (7.21). Also, since $\mathcal{Z}_A(\theta + 2\pi/N) = -\mathcal{Z}_{A'}(\theta)$ with A and A' equally likely, the distribution of $\mathcal{Z}_A(\theta)$ is symmetric (but the distribution of $\Lambda_A(\theta)$ is not: it has average value 1.) Thus, Principle 3.2 is a theorem for characteristic polynomials.

By (7.19), any (finite) sequence of gaps between eigenvalues is as likely as that sequence in reverse. Another way to say it is that $\mathcal{Z}_A(-\theta) = \mathcal{Z}_{A'}(\theta)$ with A and A' equally likely. This explains some conjectures from [96], see Subsection 14.5.

Modeling $Z(t)$ by $\mathcal{Z}_A(\theta)$ should seem natural: both are real-valued functions on $\mathbb{R}$, and (with $N \approx \log t/2\pi$) have the same average spacing between their zeros.

Modeling $\zeta(s)$ by $\Lambda_A(z)$ is slightly more subtle: one is a function which (conjecturally) has its zeros on the line $\sigma = \frac{1}{2}$, and the other has its zeros on the circle $|z| = 1$. How to translate between those worlds? The answer is to identify the half-plane $\sigma > \frac{1}{2}$ with the interior of the unit disc $|z| < 1$. That choice makes sense for several reasons; two of the author's favorites are: $\lim_{s\to+\infty} \zeta(s) = 1$ and $\lim_{z\to 0} \Lambda_A(z) = 1$; and all zeros of $\zeta(s)$ lying on $\sigma = \frac{1}{2}$ implies [98] all zeros of $\zeta'(s)$ lie in $\sigma > \frac{1}{2}$, while all zeros of $\Lambda_A(z)$ lying on $|z| = 1$ implies (by Gauss-Lucas) all zeros of $\Lambda'_A(z)$ lie in $|z| < 1$.

The most natural mapping from $\sigma > \frac{1}{2}$ to $|z| < 1$ is $s \mapsto e^{\frac{1}{2}-s}$, as in Principle 7.8.

7.5 Analogies, and lack thereof

We have seen that, conjecturally, to leading order the zeros of the ζ-function have the same limiting local statistics as the GUE or CUE. When modeling the ζ-function at height T we use Haar-random characteristic polynomials from $U(N)$ under the matching $N \leftrightarrow \log T/2\pi$. One of three things can happen.

Case 1. Agreement to leading order. For example, the author conjectured [40]

$$\frac{1}{T}\int_0^T \frac{\zeta(\frac{1}{2}+a+it)\zeta(\frac{1}{2}+b-it)}{\zeta(\frac{1}{2}+c+it)\zeta(\frac{1}{2}+d-it)}\,dt \sim \frac{(a+d)(b+c)}{(a+b)(c+d)} - T^{-a-b}\frac{(c-a)(d-b)}{(a+b)(c+d)}, \tag{7.20}$$

for $a, b, c, d \ll 1/\log T$, as $T \to \infty$. For characteristic polynomials we have the theorem [61][27]

$$\left\langle \frac{\Lambda(e^{-a})\overline{\Lambda(e^{-b})}}{\Lambda(e^{-c})\overline{\Lambda(e^{-d})}} \right\rangle_{U(N)} \sim \frac{(a+d)(b+c)}{(a+b)(c+d)} - e^{-N(a+b)}\frac{(c-a)(d-b)}{(a+b)(c+d)}, \tag{7.21}$$

as $N \to \infty$. Those expressions fit the analogy in Principle 7.8 and are identical under $N \leftrightarrow \log T$. Both expressions can be made more precise; they have different lower order terms, with the primes playing an important role for the ζ-function. (See [33], Section 4.)

In Section 5 we described the result of Bombieri and Hejhal [20] that if a set of Z-functions satisfy RH, then a linear combination has 100% of its zeros on the critical line. The same holds for characteristic polynomials [7]: if $a_1, \dots, a_n$ are nonzero real numbers and $A_1, \dots, A_n$ are chosen independently and Haar-random from $U(N)$, then the expected number of zeros of $\sum a_n \mathcal{Z}_{A_n}$ on the unit circle is $N - o(N)$.

Case 2. Agreement to leading order, after inserting an arithmetic factor. This situation appears in the moments of L-functions, as illustrated in (13.2) and (7.16)

Case 3. Failure of the analogy. For L-functions, RH implies the Lindelöf Hypothesis (LH), see Conjecture 9.7. But for unitary polynomials the analogue of RH is true but the analogue of LH can fail. For example, the unitary polynomial $(z+1)^N$ achieves the value 2^N on the unit circle, which is not $\ll e^{\varepsilon N}$ for all $\varepsilon > 0$.