and the two new terms add up to $2^{-1/2}\cos(\pi/8)$ which is the old term. Thus to a first approximation the Riemann–Siegel formula is continuous at $t = 2^2 \cdot 2\pi$ and, more generally, at all points $t = k^2 2\pi$, where N changes. To a second approximation the C_2 term changes sign [the C_1 term does not change sign because C_1 is an odd function of $(1 - 2p)$ and the two sign changes cancel] from $\pm k^{-1/2}k^{-2}C_2(0)$ to $\mp k^{-1/2}k^{-2}C_2(0)$, whereas the new term of the sum is $\pm 2k^{-1/2}\cos[-(\pi/8) + (1/k^2 96\pi)]$ which consists of the term already accounted for plus $\pm 2k^{-1/2}\sin(\pi/8)\sin(1/k^2 96\pi) \sim \pm 2k^{-1/2}k^{-2}\sin(\pi/8)/96\pi$; hence there is no discontinuity at this level of approximation provided $C_2(0) = \sin(\pi/8)/96$, which is in fact the case as can be shown by straightforward evaluation of $C_2(0)$.

Although there is no serious discontinuity near $t = 2^2 \cdot 2\pi \sim 25.133$, it is still quite conceivable that the Riemann–Siegel formula would not be as accurate in this region as it proved to be in the examples above. To find a root of Z corresponding to the root $t \sim 25.5$ of $\cos \vartheta(t) = 0$, one would observe first that the terms after $2\cos\vartheta(t)$ are about $2^{-1/2}\cos(\pi/8) \sim 0.65$ in this range of t, as was just seen; hence t should be changed to make $2\cos\vartheta(t) \sim -0.65$. Since the derivative of $2\cos\vartheta(t)$ is about $-2 \cdot \frac{1}{2}\log(t/2\pi)\sin\vartheta(t) \sim \log(t/2\pi) \sim 1.4$, this suggests $t = 25.5 - (0.65)/(1.4) \sim 25.0$. Now for $t = 25$, N is 1 and p is 0.994 711 4, so $1 - 2p$ is −0.989 422 8. Careful computations for $t = 25$ (which are impeded by the fact that p is nearly 1 and that the series for C_t consequently converge slowly) give $Z(25) \sim -0.014\,873\,455$ which agrees with Haselgrove's six-place value −0.014 872 except in the last place. Thus even here, in the neighborhood of an apparent discontinuity, the Riemann–Siegel formula is astonishingly accurate and there is no conclusive empirical evidence of *any* inherent error in the formula, much less an inherent error of the order of magnitude allowed by the crude error estimates above.

The next approximation to the root near $t = 25$ would be obtained by increasing t enough to increase $2\cos\vartheta(t)$ by 0.0149. The derivative at $t = 25$ is about $-\log(25/2\pi)\sin[(3\pi/2) - 0.34)] = \cos(0.34)\log(25/2\pi) \sim (0.94) \times (1.38) \sim 1.30$, so t should be increased by $(0.0149)/(1.30) \sim 0.011$ to about

main term	−0.670 310 810
C_0 term	0.645 191 368
C_1 term	0.010 011 009
C_2 term	0.000 216 855
C_3 term	0.000 017 159
C_4 term	0.000 000 964
$Z(25)$	−0.014 873 455

Computation of the approximation to $Z(25)$.

25.011. The actual root is, of course, at 25.01085.... *Riemann also computed this root*, but the value 25.31 which he obtained is very far off—so far off, in fact, that it must surely indicate a computational error because even the rough calculation at the beginning of the preceding paragraph yielded the better value 25.0. I have not been able to follow the details of Riemann's computation, but I have followed enough of it to see that after starting with $t = 4 \cdot 2\pi$ he goes to the *larger* value $(4.030\ldots) \cdot 2\pi$, which means he has gone in the wrong direction. This error should have revealed itself when he computed the new value of Z, but presumably the error in the first computation, whatever it was, was repeated in the subsequent computations.

Although the above computations of $Z(t)$ [from which $\zeta(\frac{1}{2} + it)$ can be computed immediately] are considerably shorter than the corresponding computations using Euler–Maclaurin summation would be, the real superiority of the Riemann–Siegel formula is for larger values of t, both because the number of terms it requires increases slowly and because the inherent error decreases. For example, to compute $Z(1000)$ using the Riemann–Siegel formula requires the evaluation of $[(t/2\pi)^{1/2}] = 12$ terms in the main sum. The inherent error is, judging by the above calculations, much smaller than the C_4 term which is of the order of magnitude of $(t/2\pi)^{-9/4}(0.001) \sim 12^{-9/2} \times 10^{-3} < 2 \times 10^{-8}$. To achieve comparable accuracy with the Euler–Maclaurin formula would require hundreds of terms as opposed to just 12, and would require so much arithmetic that computations would have to be carried out with great accuracy to counteract the accumulation of roundoff error.

7.7 ERROR ESTIMATES

The computational examples of the preceding section suggest that the usual rule of thumb for asymptotic series (see Section 6.2) applies to the Riemann–Siegel formula; that is, *as long as the terms are decreasing rapidly the bulk of the error is in the first term omitted.* Moreover, the first four remainder terms are rapidly decreasing in size even when t is only 14, and the C_4 term is already less than 10^{-4} when t is in this range and is very much smaller for larger t. This suggests that even though the Riemann–Siegel formula has an inherent error, this error is extremely small and the formula in fact makes possible the computation of $Z(t)$ with an accuracy of several decimal places.

Unfortunately none of the estimates of the error in the Riemann–Siegel formula come anywhere near to justifying these conjectures about its accuracy. At the present time the only published error estimate seems to be Titchmarsh's [T5], which shows that *if* $t > 125 \cdot 2\pi \sim 786$ *and if the* C_1 *term is*

the first term omitted, then the error is less than $(3/2)(t/2\pi)^{-3/4}$ *in magnitude.* (Actually the estimates of Titchmarsh [T5] are a good deal more complicated and cover a wider range of t. The simplified version given here is taken from Titchmarsh's book [T8, p. 331].) In proving the existence of zeros of $Z(t)$, and hence of roots ρ on $\operatorname{Re} s = \frac{1}{2}$, the principal requirement is to be able to determine the *sign* of $Z(t)$ with certainty, and in most cases it has been possible to do this by finding values of t where $Z(t)$ is far enough from zero that its sign is rigorously determined by Titchmarsh's estimate of the error. However, cases do arise in which $Z(t)$ changes sign but remains very small in absolute value (Lehmer's phenomenon—see Section 8.3) and in these cases the presence of zeros of Z cannot be rigorously established without a stronger error estimate and, in particular, without one which uses more than one term of the Riemann–Siegel formula.

Rosser, Yohe, and Shoenfeld [R3] have announced that *if the* C_3 *term is the first term omitted, then the error is less than* $(2.88)(t/2\pi)^{-7/4}$ *provided* $t > 2000 \cdot 2\pi \sim 12{,}567$. Their proof of this result has not yet appeared. They have also established rigorous estimates of the error in their procedures for computing the C_0, C_1, and C_2 terms and in this way have been able to determine with certainty the sign of $Z(t)$ in the range covered by their calculations (see Section 8.4). In fact, they report that in five million evaluations of $Z(t)$ they found only four values of t where $|Z(t)|$ was so small that their program was unable to determine its sign with certainty.

Siegel himself proved that *all terms of the Riemann–Siegel formula are significant* in the sense that for every j there exist constants t_0, K such that if the C_j term is the first term omitted, then the error is less than $K(t/2\pi)^{-(2j+1)/4}$ provided $t > t_0$. Thus Titchmarsh's estimate gives specific values for K, t_0 in the case $j = 1$, and the estimate of Rosser *et al.* does the same for $j = 3$. Siegel's theorem can also be stated in the form: *The Riemann–Siegel formula is an asymptotic expansion of* $Z(t)$ in the sense of "asymptotic" defined in Section 5.4 provided the "order of magnitude" of the C_{j-1} term is interpreted† as meaning $(t/2\pi)^{-(2j-1)/4}$. The actual values of t_0 which Siegel's proof provides for various values of j are extremely large and are of no use in actual computation.

Since the location of the roots ρ has been reliably carried out by several computer programs up to the level where the estimate of Rosser *et al.* applies, and since it would seem that this estimate is sufficient for locating the roots beyond this level (it is conceivable, however, that there might be occurrences of Lehmer's phenomenon so extreme that the C_3 term is needed to determine

†The hitch here is that the C_{j-1} term for some values of t might be *zero*. However, it is of the order of magnitude of $(t/2\pi)^{-(2j-1)/4}$ in the sense that it is always less than a constant times this amount but not always less than a constant times any higher power of t^{-1}.

the sign of Z), the known estimates appear to be sufficient for the location of the roots ρ. Nonetheless, it would be of interest to have an estimate which comes closer to the "rule of thumb" estimate which seems justified by the computations, even for $t \sim 15$. Also of interest would be an answer to the question raised by Siegel as to whether the Riemann–Siegel formula *converges* for fixed t as more and more terms are used. Presumably it does not—by the analogy with Stirling's formula—but this has never been proved.

7.8 SPECULATIONS ON THE GENESIS OF THE RIEMANN HYPOTHESIS

We can never know what led Riemann to say it was "probable" that the roots ρ all lie on the line $\operatorname{Re} s = \frac{1}{2}$, and the contents of this chapter show very clearly how foolhardy it would be to try to say what mathematical ideas may have lain behind this statement. Nonetheless it is natural to *try* to guess what might have led him to it, and I believe that the Riemann–Siegel formula gives some grounds for a plausible guess.

Even today, more than a hundred years later, one cannot really give any solid reasons for saying that the truth of the Riemann hypothesis is "probable." The theorem of Bohr and Landau (Section 9.6) stating that *for any* $\delta > 0$ *all but an infinitesimal proportion of the roots* ρ *lie within* δ *of* $\operatorname{Re} s = \frac{1}{2}$ is the only positive result which lends real credence to the hypothesis. Also the verification of the hypothesis for the first three and a half million roots above the real axis (Section 8.4) perhaps makes it more "probable." However, any real *reason*, any plausibility argument or heuristic basis for the statement, seems entirely lacking. Siegel states quite positively that the Riemann papers contain† no steps toward a proof of the Riemann hypothesis, and therefore one is safe in assuming that they do not contain any plausibility arguments for the Riemann hypothesis either. Thus the question remains: Why did Riemann think it was "probable"?

My guess is simply that Riemann used the method followed in Section 7.6 to locate roots and that he observed that normally—as long as t is not so large that the Riemann–Siegel formula contains too many terms and as long as the terms do not exhibit too much reinforcement—this method allows one to go from a zero of the first term $2\cos\vartheta(t)$ to a nearby zero of $Z(t)$. This heuristic argument implies there are "about" as many zeros of $Z(t)$ as there are of $2\cos\vartheta(t)$, that is, about $\pi^{-1}\vartheta(t)$ zeros. But Riemann already knew (we do not know how) that this was the approximate formula for the total num-

†Of course there is no indication that these one hundred or so pages contain all of Riemann's studies of the zeta function, and they almost certainly do not.

ber of roots, on the line or off. Would it not be natural to explain this approximate equality by the hypothesis that, for some reason which might become clear on further investigation, the roots are all on the line? And would it not be natural to express this hypothesis in exactly the words which Riemann uses?

This guess, if it is correct, has two important consequences. In the first place it implies that when Riemann says that the number of roots on the line is "about" equal to the total number of roots, he does *not* imply asymptotic equality, as has often been assumed, but simply that when t is not too large and the terms do not reinforce too much, there is a one-to-one correspondence between zeros of $Z(t)$ and zeros of $\cos \vartheta(t)$. To go from a zero $\hat{t}$ of $\cos \vartheta(t)$ to one of $Z(t)$, one would evaluate the terms of the Riemann–Siegel formula at $\hat{t}$. Since the terms other than the first are individually small, since they would normally cancel each other to some extent, and since they change more slowly than the first term, it would seem likely that in most cases one could move to a point near $\hat{t}$ where the first term $2 \cos \vartheta(t)$ had a value equal to the negative of the value of the remaining terms at $\hat{t}$ and that the value of the remaining terms would not change too much in the process. If so, then the value of Z at the new point is small and can be made still smaller by successive approximations based on the assumption that for a small change in t the bulk of the change in $Z(t)$ occurs in the first term, hence converging to a root of Z. The method can certainly fail. To see how completely it can fail, it suffices to consider an extreme failure of Gram's law, for example the failure between g_{6708} and g_{6709} in Lehmer's graph (Fig. 3, Section 8.3). If $\hat{t}$ is the zero of $\cos \vartheta(t)$ in this interval, then $Z(\hat{t})$ is about -2, perhaps even less; so it is not possible to increase the first term $2 \cos \vartheta(t)$ enough to make up for the deficit in the other terms. If one increases $2 \cos \vartheta(t)$ the full amount by moving to g_{6708} where it is $+2$, the total value of Z is still negative, approximately $-\frac{1}{2}$. To reach the zero of Z corresponding to $\hat{t}$, one must move even further to the left and hence one must *decrease* $2 \cos \vartheta(t)$ in the hope that the other terms will increase and increase enough to make up for the decrease in $2 \cos \vartheta(t)$ and the remaining deficit of $-\frac{1}{2}$. Thus one must abandon the proposed method entirely and hope that the desired zero is there anyway. Viewed in this way, the heuristic argument which I impute to Riemann is virtually identical with Gram's law but with the very important difference that its rationale is based on the Riemann–Siegel formula rather than the Euler–Maclaurin formula, so that, unlike Gram's rationale, it is not at all absurd to expect the main term to dominate the sign of the series for t into the hundreds or even the thousands because the formula for $Z(t)$ has only ten or twenty terms in this range, not hundreds of terms. It seems entirely possible that Riemann would have been able to judge that the failures in this range would be relatively rare, as has now been verified by computation, and to conclude

that the number of zeros of $Z(t)$ between 0 and T *for* T *in this medium range* is about $\pi^{-1}\vartheta(T)$, that is, about† $(T/2\pi)\log(T/2\pi) - (T/2\pi)$. In support of this interpretation of Riemann's use of the word "about" (*etwas*) in this place, one might observe that he prefaces it with the phrase "one finds in fact" (meaning computationally?) and that, unlike his use of "about" in the previous sentence, he gives no estimate of the error in the approximation.

The second consequence of my guess is that it implies that Riemann based his hypothesis on no insights about the function ξ which are not available to us today (now that we have the Riemann–Siegel formula) and that, on the contrary, had he known some of the facts which have since been discovered, he might well have been led to reconsider. After all, he had just discovered the extension of the zeta function to the entire complex plane, the functional equation, and an effective numerical technique for locating many roots on the line, so it would be perfectly natural for him to be looking for regularities and perfectly natural for him to expect that an observed regularity of this sort would hold and would yield to the power of his function-theoretic concepts and techniques. However, it did not yield, and Riemann lived for several years after he made the hypothesis. Moreover, the discoveries of Lehmer's phenomenon (Section 8.3) and of the fact that $Z(t)$ is unbounded (Section 9.2) completely vitiate any argument based on the Riemann–Siegel formula and suggest that, unless some basic cause is operating which has eluded mathematicians for 110 years, occasional roots ρ off the line are altogether possible. In short, although Riemann's insight was stupendous it was not supernatural, and what seemed "probable" to him in 1859 might seem less so today.

7.9 THE RIEMANN–SIEGEL INTEGRAL FORMULA

In 1926 Bessel–Hagen found (according to Siegel [S4]) in the Riemann papers a new representation of the zeta function in terms of definite integrals. Naturally Siegel included an exposition of this formula in his 1932 account of the portions of Riemann's *Nachlass* relating to analytic number theory. As stated by Siegel, the formula is essentially

$$\frac{2\xi(s)}{s(s-1)} = F(s) + \overline{F(1-\bar{s})} \tag{1}$$

where F is defined by the formula

$$F(s) = \Pi\left(\frac{s}{2} - 1\right)\pi^{-s/2}\int_{0\searrow 1}\frac{e^{-i\pi x^2}x^{-s}\,dx}{e^{i\pi x} - e^{-i\pi x}}$$

†Of course, if the roots are all on the line, which Riemann thought was the probable explanation of the near equality of the two estimates, then the same estimate applies for all T and the slight ambiguity in the range of T does no harm.

in which the symbol $0\searrow 1$ means that the path of integration is a line of slope -1 crossing the real axis between 0 and 1 and directed from upper left to lower right, and in which x^{-s} is defined on the slit plane (excluding 0 and negative real numbers) in the usual way by taking $\log x$ to be real on the positive real axis and setting $x^{-s} = e^{-s\log x}$. Because $\exp(-i\pi x^2)$ approaches zero very rapidly as $|x| \to \infty$ along any line of the form $0\searrow 1$ and because the integrand is nonsingular on the slit plane except for simple poles at the positive integers, it is easily seen that $F(s)$ is an analytic function of s defined for all s except possibly for $s = 0, -2, -4, \ldots$, where the factor in front has simple poles. [Formula (1), once it is proved, implies that $F(s)$ is analytic at $-2, -4, \ldots$ and has a simple pole at 0.]

Siegel deduces formula (1) from an alternative form of the identity

$$(2) \qquad e^{i\pi/8}e^{-2\pi ip^2}\frac{1}{2\pi i}\int_\Gamma \frac{e^{iu^2/4\pi}e^{2pu}\,du}{e^u - 1} = \frac{\cos 2\pi(p^2 - p - \frac{1}{16})}{\cos 2\pi p}$$

[formula (5) of Section 7.4]. The change of variable $u = 2\pi i w$ puts the path of integration Γ in the form $0\searrow 1$ but with the orientation reversed, and puts the identity itself in the form

$$\int_{0\searrow 1} \frac{e^{-i\pi w^2}e^{4\pi ipw}\,dw}{e^{2\pi iw} - 1} = -e^{-i\pi/8}e^{2\pi ip^2}\frac{\cos 2\pi(p^2 - p - \frac{1}{16})}{\cos 2\pi p}$$

which with $p = \frac{1}{2}(v + \frac{1}{2})$ can be simplified using $2p = v + \frac{1}{2}$, $4p^2 - 4p - \frac{1}{4} = (v + \frac{1}{2})^2 - 2(v + \frac{1}{2}) - \frac{1}{4} = v^2 - v - 1$ to be

$$\begin{aligned}\int_{0\searrow 1} \frac{e^{-i\pi w^2}e^{2\pi i[v+(1/2)]w}\,dw}{e^{i\pi w}(e^{i\pi w} - e^{-i\pi w})} &= -e^{-i\pi/8}e^{i\pi[v+(1/2)]^2/2}\frac{\cos[\pi(v^2 - v - 1)/2]}{\cos \pi(v + \frac{1}{2})} \\ &= -e^{i\pi v^2/2}e^{i\pi v/2}\frac{e^{i\pi(v^2-v-1)/2} + e^{-i\pi(v^2-v-1)/2}}{e^{i\pi[v+(1/2)]} + e^{-i\pi[v+(1/2)]}} \\ &= -\frac{e^{i\pi v^2}e^{-i\pi/2} + e^{i\pi v}e^{i\pi/2}}{e^{i\pi v}e^{i\pi/2} + e^{-i\pi v}e^{-i\pi/2}}\end{aligned}$$

and finally

$$(3) \qquad \int_{0\searrow 1} \frac{e^{-i\pi w^2}e^{2\pi ivw}\,dw}{e^{i\pi w} - e^{-i\pi w}} = \frac{e^{i\pi v^2}}{e^{i\pi v} - e^{-i\pi v}} - \frac{1}{1 - e^{-2\pi iv}}$$

which is the alternative form of (2). Let s be a negative real number, multiply both sides of this equation by $v^{-s}\,dv$, and integrate along the ray from $v = 0$ to $v = \infty\, i^{1/2}$. The double integral on the left converges absolutely, so the order of integration can be interchanged. Since by elementary manipulation of definite integrals

$$\begin{aligned}\int_0^{\infty i^{1/2}} v^{-s}e^{2\pi ivw}\,dv &= \int_0^{\infty w i^{-1/2}} \left(\frac{ix}{2\pi w}\right)^{1-s} e^{-x}\,d\log x \\ &= \left(\frac{i}{2\pi w}\right)^{1-s}\int_0^\infty x^{-s}e^{-x}\,dx \\ &= ie^{-i\pi s/2}(2\pi)^{s-1}w^{s-1}\Pi(-s)\end{aligned}$$

(for w on $0 \searrow 1$), it follows that the left side becomes

$$\int_0^{\infty i^{1/2}} \int_{0\searrow 1} \frac{e^{-i\pi w^2} e^{2\pi i v w} v^{-s}\, dw\, dv}{e^{i\pi w} - e^{-i\pi w}}$$
$$= ie^{-i\pi s/2}(2\pi)^{s-1}\Pi(-s) \int_{0\searrow 1} \frac{e^{-i\pi w^2} w^{s-1}\, dw}{e^{i\pi w} - e^{-i\pi w}}.$$

The second term on the right becomes†

$$\int_0^{\infty i^{1/2}} \left(\frac{-1}{1 - e^{-2\pi i v}}\right) v^{-s}\, dv = \int_0^{\infty i^{1/2}} \sum_{n=1}^{\infty} e^{2\pi i n v} v^{-s}\, dv$$
$$= \sum_{n=1}^{\infty} \int_0^{\infty i^{1/2}} e^{2\pi i n v} v^{-s}\, dv$$
$$= \sum_{n=1}^{\infty} \int_0^{\infty i^{1/2}} e^{2\pi i w} \left(\frac{w}{n}\right)^{1-s} d \log w$$
$$= \zeta(1 - s) \int_0^{\infty i^{1/2}} w^{-s} e^{2\pi i w}\, dw$$
$$= \zeta(1 - s) i e^{-i\pi s/2}(2\pi)^{s-1} 1^{s-1}\Pi(-s),$$

by the same calculation. The first term on the right can be expressed in terms of the definite integral

$$\int_{0\nearrow 1} \frac{e^{i\pi u^2} u^{-s}\, du}{e^{i\pi u} - e^{-i\pi u}} \tag{4}$$

(where $0 \nearrow 1$ denotes the complex conjugate of a path $0 \searrow 1$) because for negative real s the path $0 \nearrow 1$ can be moved over to the line of slope 1 through the origin so that (4) can be expressed as

$$\int_{-\infty i^{1/2}}^{0} \frac{e^{i\pi u^2} u^{-s}\, du}{e^{i\pi u} - e^{-i\pi u}} + \int_0^{\infty i^{1/2}} \frac{e^{i\pi u^2} u^{-s}\, du}{e^{i\pi u} - e^{-i\pi u}}$$
$$= \int_{\infty i^{1/2}}^{0} \frac{e^{i\pi(-u)^2}(-u)^{-s}\, d(-u)}{e^{-i\pi u} - e^{i\pi u}} + \int_0^{\infty i^{1/2}} \frac{e^{i\pi u^2} u^{-s}\, du}{e^{i\pi u} - e^{-i\pi u}}$$
$$= \int_0^{\infty i^{1/2}} \frac{e^{i\pi u^2}[u^{-s} - (-u)^{-s}]\, du}{e^{i\pi u} - e^{-i\pi u}}$$
$$= (1 - e^{i\pi s}) \int_0^{\infty i^{1/2}} \frac{e^{i\pi v^2} v^{-s}\, dv}{e^{i\pi v} - e^{-i\pi v}}$$

because $\log(-u) = \log u - i\pi$ for u on the ray $\operatorname{Im} \log u = \pi/4$, and hence $(-u)^{-s} = u^{-s} e^{-s(-i\pi)}$. Thus (3) becomes

$$ie^{-i\pi s/2}(2\pi)^{s-1}\Pi(-s) \int_{0\searrow 1} \frac{e^{-i\pi w^2} w^{s-1}\, dw}{e^{i\pi w} - e^{-i\pi w}}$$
$$= \frac{1}{1 - e^{i\pi s}} \int_{0\nearrow 1} \frac{e^{i\pi u^2} u^{-s}\, du}{e^{i\pi u} - e^{-i\pi u}} + ie^{-i\pi s/2}(2\pi)^{s-1}\Pi(-s)\zeta(1 - s).$$

†Justification of the interchange of summation and integration is not altogether elementary. One method is to observe that $\lim_{N\to\infty} \int_0^{\infty i^{1/2}} (e^{2\pi i N v}/e^{-2\pi i v} - 1) v^{-s}\, dv = 0$ by the Riemann–Lebesgue lemma.

Now

$$(1 - e^{i\pi s})ie^{-i\pi s/2}(2\pi)^{s-1}\Pi(-s) = 2[\sin(s\pi/2)](2\pi)^{s-1}\Pi(-s)$$

is the factor which appears in the functional of ζ [formula (4) of Section 1.6], and therefore, as in Section 1.6, it can be written as

$$\frac{\Pi[\frac{1}{2}(1-s)-1]\pi^{-(1-s)/2}}{\Pi(\frac{1}{2}s-1)\pi^{-s/2}}.$$

Therefore, when the above formula is multiplied first by $(1 - e^{i\pi s})$ and then by $\Pi\left(\frac{s}{2} - 1\right)\pi^{-s/2}$, it becomes

$$\Pi\left(\frac{1-s}{2} - 1\right)\pi^{-(1-s)/2}\int_{0\searrow 1}\frac{e^{-i\pi w^2}w^{s-1}\,dw}{e^{i\pi w} - e^{-i\pi w}}$$
$$= \Pi\left(\frac{s}{2} - 1\right)\pi^{-s/2}\int_{0\nearrow 1}\frac{e^{i\pi u^2}u^{-s}\,du}{e^{i\pi u} - e^{-i\pi u}} + \frac{2\xi(1-s)}{(1-s)(-s)}$$

because by definition

$$\frac{2\xi(s)}{s(s-1)} = \Pi\left(\frac{s}{2} - 1\right)\pi^{-s/2}\zeta(s). \tag{5}$$

The left side of this equation is $F(1 - s)$ and the first term on the right is $-\overline{F(\bar{s})}$; hence

$$\frac{2\xi(1-s)}{(1-s)(-s)} = F(1-s) + \overline{F(\bar{s})}$$

and the desired formula (1) is proved by substituting $1 - s$ for s.

The Riemann–Siegel formula puts in evidence the fact that ξ satisfies the functional equation $\xi(s) = \overline{\xi(1-\bar{s})}$ because it shows that

$$\frac{2\xi(s)}{s(s-1)} = F(s) + \overline{F(1-\bar{s})} = \overline{F(1-\bar{s})} + F(s)$$
$$= \text{complex conjugate of } \frac{2\xi(1-\bar{s})}{(1-\bar{s})(-\bar{s})} = \frac{2\overline{\xi(1-\bar{s})}}{s(s-1)}.$$

On the other hand, ξ is real on the real axis by (5), and therefore, by the reflection principle, $\overline{\xi(s)} = \xi(\bar{s})$. The Riemann–Siegel integral formula therefore gives a new proof of the functional equation $\xi(s) = \xi(1 - s)$. This proof differs from Riemann's first proof in that it uses s and $1 - s$ more symmetrically, and it differs from his second proof in that it does not depend on the identity $1 + 2\psi(x) = x^{-1/2}[1 + 2\psi(x^{-1})]$ from the theory of theta functions.

Since the theta function identity $1 + 2\psi(x) = x^{-1/2}[1 + 2\psi(x^{-1})]$ can be deduced from $\xi(s) = \xi(1 - s)$ fairly easily (by Fourier inversion—see Chapter 10), the proof of this section gives an alternative proof of the theta function identity based on the evaluation of the definite integral (2). More generally, Siegel states that Riemann in his unpublished lectures derived the

transformation theory of theta functions from a study of the integral

$$\Phi(\tau, u) = \int_{0\searrow 1} \frac{e^{i\pi\tau x^2} e^{2\pi i u x}\, dx}{e^{-i\pi x} - e^{i\pi x}}$$

of which the special case $\tau = -1$ was considered above.

Chapter 8

Large-Scale Computations

8.1 INTRODUCTION

The discovery of the Riemann–Siegel formula made it quite feasible to extend the program begun by Gram and Hutchinson (see Chapter 6) well beyond the Gram point g_{137} reached by Hutchinson in 1925. Since the Gram points g_n are easily computed, this extension is simply a matter of using the Riemann–Siegel formula to evaluate $Z(g_n)$ and of finding points g_n' near g_n for which $(-1)^n Z(g_n') > 0$ in those presumably rare cases when Gram's law† $\operatorname{Re} \zeta(\frac{1}{2} + ig_n) = Z(g_n) \cos \vartheta(g_n) = (-1)^n Z(g_n) > 0$ fails. In this way it should be possible, unless the failures of Gram's law become too frequent, to locate many more roots ρ on the line $\operatorname{Re} s = \frac{1}{2}$. Such computations were carried out by Titchmarsh and Comrie [*T5*, *T6*] in 1935–1936 extending up to the Gram point g_{1040} and thus locating 1041 roots on the line. Moreover, by a suitable generalization of the techniques of Backlund and Hutchinson using the Riemann–Siegel formula (in a more general form which is applicable for $\operatorname{Re} s \neq \frac{1}{2}$) in place of the Euler–Maclaurin formula, Titchmarsh and Comrie were able to show that $N(g_{1040}) = 1041$ and to conclude therefore that *the roots ρ in the range $\{0 \leq \operatorname{Im} s \leq g_{1040}\}$ are all simple zeros on the line* $\operatorname{Re} s = \frac{1}{2}$.

No doubt this program of computation would have been carried further if World War II had not intervened. By the time the war was over, the computer revolution was well under way and automatic electronic digital com-

†Strictly speaking the statement $(-1)^n Z(g_n) > 0$ should perhaps be called the *weak* Gram's law, since, as defined by Hutchinson, "Gram's law" is the stronger statement that there are precisely $n + 1$ zeros of $Z(t)$ between 0 and g_n. This distinction is not very important and is ignored in what follows.

puters were rapidly being developed. These new tools made it feasible to extend the computations to cover tens of thousands and even hundreds of thousands of Gram points, but, apparently because computer technology was changing so rapidly that no single computer remained in operation long enough for such a low-priority project to be programmed and run on it before a new computer requiring a new program would replace it, it was not until 1955–1956 that the computations were carried significantly past the level reached by Titchmarsh and Comrie 20 years before.

These computations, carried out by D.H. Lehmer [L7, L8], showed that for the first 25,000 Gram points g_n the exceptions to Gram's law $(-1)^n Z(g_n) > 0$ are not great and *all roots ρ in the range* $\{0 \leq \operatorname{Im} s \leq g_{25000}\}$ *are simple zeros on the line* $\operatorname{Re} s = \frac{1}{2}$. Lehmer had at his disposal, in addition to the Riemann–Siegel formula and the new electronic computers, a new method introduced by Turing [T9] in 1953 for determining the number of roots in a given range. Turing's method is much easier to apply in practice than is the method of Backlund (see Section 6.6) which it supplants. Turing's method is described in the next section.

Although Lehmer's computations confirmed the Riemann hypothesis as far as they went, they disclosed certain irregularities in the behavior of $Z(t)$ which made it seem altogether possible that further computations might actually produce a counterexample to the Riemann hypothesis (see Section 8.3). Further computations were carried out—by Lehman [L5] to the two hundred and fifty thousandth zero and by Rosser *et al.* [R3] to the *three and a half millionth zero*—without, however, producing a counterexample and in fact proving that all roots ρ in the range $\{0 \leq \operatorname{Im} s \leq g_{3500000}\}$ are simple zeros on the line $\operatorname{Re} s = \frac{1}{2}$ (see Section 8.4).

8.2 TURING'S METHOD

As before, let $N(T)$ denote the number of roots ρ in the range $\{0 \leq \operatorname{Im} s \leq T\}$ (counted with multiplicities). Recall that Backlund's method of evaluating $N(T)$ is to prove if possible that $\operatorname{Re} \zeta(\sigma + iT)$ is never zero for $\frac{1}{2} \leq \sigma \leq 1\frac{1}{2}$, from which it follows that $N(T)$ is the integer nearest $\pi^{-1}\vartheta(T) + 1$ (see Section 6.6). The disadvantage of this method is that it requires the evaluation of $\zeta(s)$ at points not on the line $\operatorname{Re} s = \frac{1}{2}$. By contrast, Turing's method not only does not require the evaluation of $\zeta(s)$ at points not on $\operatorname{Re} s = \frac{1}{2}$, but in fact it requires only the information which is naturally acquired in looking for changes of sign in $Z(t)$, namely, a list of those Gram points g_n for which Gram's law $(-1)^n Z(g_n) > 0$ fails and a determination of how far each of them must be moved in order to give Z the desired sign $(-1)^n$.

More precisely, assume that for all integers n in a certain range numbers h_n have been found such that $(-1)^n Z(g_n + h_n) > 0$, such that the sequence $\ldots, g_{n-1} + h_{n-1}, g_n + h_n, g_{n+1} + h_{n+1}, \ldots$ is strictly increasing, and such that h_n is small and is zero whenever possible. Such a list of numbers h_n would naturally be generated in using Gram's law to locate changes of sign of Z. Turing showed that *if* $h_m = 0$ *and if the values of* h_n *for* n *near* m *are not too large, then* $N(g_m)$ *must have the value predicted by Gram's law*, namely, $N(g_m) = m + 1$.

Turing's method is based on the following theorem of Littlewood (see Section 9.5). Let $S(T)$ denote† the error in the approximation $N(T) \sim \pi^{-1}\vartheta(T) + 1$, that is, let $S(T) = N(T) - \pi^{-1}\vartheta(T) - 1$. Von Mangoldt proved (see Section 6.7) that the absolute value of $S(T)$ grows no faster than a constant times $\log T$ as $T \longrightarrow \infty$. Littlewood in 1924 proved a different kind of estimate of $S(T)$, namely, that $\int_0^T S(t)\,dt$ grows no faster than a constant times $\log T$ as $T \longrightarrow \infty$. Although this does not imply von Mangoldt's result— it is possible for $S(t)$ to have arbitrarily large absolute values without its integral necessarily being large—Littlewood's theorem is in a sense much stronger in that it shows that on the average $S(T)$ approaches zero,‡ $\lim_{T\to\infty} 1/T \int_0^T S(t)\,dt = 0$, whereas von Mangoldt's result only shows that $S(T)$ does not become large too fast.

Now suppose that $h_m = 0$ and that the h_n for n near m are small. Then $S(g_m)$ must be an integer because $S(g_m) = N(g_m) - m - 1$ and this integer must be even because the parity of the number of roots *on* the line segment from $\frac{1}{2}$ to $\frac{1}{2} + ig_m$ (counted with multiplicities) is determined by the sign of $Z(g_m)$ which is $(-1)^m$ by assumption, and because the roots *off* the line $\operatorname{Re} s = \frac{1}{2}$, if any, occur in pairs. Thus to prove $S(g_m) = 0$ it suffices to prove $S(g_m) < 2$ and $S(g_m) > -2$. Assume first that $S(g_m) \geq 2$. If $h_{m+1} = 0$, then $(-1)^{m+1} Z(g_{m+1}) > 0$, so there must be a zero of Z between g_m and g_{m+1}, and N must increase by at least one as t goes from g_m to g_{m+1}; on the other hand $\pi^{-1}\vartheta(t) + 1$ increases by exactly one as t goes from g_m to g_{m+1}, so $S(g_{m+1}) \geq 2$ and $S(t)$ never falls below one for $g_m \leq t \leq g_{m+1}$. If $h_{m+1} < 0$, then it is true *a fortiori* that $S(g_{m+1}) \geq 2$ and that $S(t)$ never falls below one for $g_m \leq t \leq g_{m+1}$, whereas if h_{m+1} is a small positive number, then t must pass g_{m+1} and go to $g_{m+1} + h_{m+1}$ in order to bring $S(t)$ back up to a value near two and in the process $S(t)$ falls only slightly below one. If h_{m+2} and the succeeding h's are zero or are small this argument can be continued to show that $S(g_n + h_n)$ is nearly two for $n = m + 1, m + 2, m + 3, \ldots$ and that $S(t)$

†The formula $N(T) = \pi^{-1}\vartheta(T) + 1 + \pi^{-1} \operatorname{Im} \int_C [\zeta'(s)/\zeta(s)]\,ds$ of Section 6.6 shows that $S(T)$ can also be defined as $\pi^{-1} \operatorname{Im} \log \zeta(\frac{1}{2} + iT)$ when $\log \zeta$ is defined by analytic continuation from the positive real axis as in Section 6.6.

‡In particular, Littlewood's theorem shows that the constant 1 in the definition of $S(T)$ has significance.

falls only slightly below one to the right of g_m. Since by Littlewood's theorem this can not continue indefinitely [because then the average value of $S(t)$ would not approach zero] either some of the values $h_{m+1}, h_{m+2}, h_{m+3}, \ldots$ must be large and positive or the original assumption $S(g_m) \geq 2$ must be false. Turing's idea was to prove a *quantitative* version of Littlewood's theorem in order to obtain a quantitative description of the sizes of the h_n implied by $S(g_m) \geq 2$; then to prove $S(g_m) \leq 0$ it suffices to show that the h_n are in fact less than this amount.

Specifically, Turing showed that Littlewood's proof (see Section 9.5) can be made to yield the inequality

$$\left| \int_{t_1}^{t_2} S(t)\, dt \right| \leq 2.30 + 0.128 \log\left(\frac{t_2}{2\pi}\right) \tag{1}$$

for all $t_2 > t_1 > 168\pi$. This can be used to derive a relationship between $S(g_m)$ and the sizes of $h_{m+1}, h_{m+2}, h_{m+3}, \ldots$ as follows. Assume as before that $h_m = 0$ and let $L(t)$ denote the step function which is zero at g_m and which has jumps of one at $g_{m+1} + h_{m+1}, g_{m+2} + h_{m+2}, \ldots$. Then since Z changes sign between successive points $g_n + h_n$, the number of roots N must go up by at least one and $N(t) \geq N(g_m) + L(t)$ for $t \geq g_m$. On the other hand let $L_1(t)$ denote the step function which is zero at g_m and which has jumps of one at $g_{m+1}, g_{m+2}, \ldots$. Then $L_1(t)$ increases by one when $\pi^{-1}\vartheta(t)$ increases by one, from which it follows that $\pi^{-1}\vartheta(t) + 1 \leq \pi^{-1}\vartheta(g_m) + 1 + L_1(t) + 1$ for $t \geq g_m$ and hence $S(t) \geq S(g_m) + L(t) - L_1(t) - 1$. Now $L(t) - L_1(t)$ is normally zero, but if h_n is positive, then $L_1(t)$ jumps before $L(t)$ does and $L(t) - L_1(t)$ is -1 on the interval from g_n to $g_n + h_n$. Similarly if $h_n < 0$, then $L(t) - L_1(t)$ is $+1$ on the interval from $g_n + h_n$ to g_n. Assume for the sake of convenience that $h_{m+k} = 0$. Then

$$\begin{aligned}
\int_{g_m}^{g_{m+k}} S(t)\, dt &\geq \int_{g_m}^{g_{m+k}} [S(g_m) - 1]\, dt + \int_{g_m}^{g_{m+k}} [L(t) - L_1(t)]\, dt \\
&= (g_{m+k} - g_m)[S(g_m) - 1] - \sum_{j=m+1}^{m+k-1} h_j, \\
S(g_m) - 1 &\leq \frac{\int_{g_m}^{g_{m+k}} S(t)\, dt + \sum_{j=m+1}^{m+k-1} h_j}{g_{m+k} - g_m}, \\
S(g_m) &\leq 1 + \frac{2.30 + 0.128 \log(g_{m+k}/2\pi) + \sum_{j=m+1}^{m+k-1} h_j}{g_{m+k} - g_m}. \qquad (2)
\end{aligned}$$

Since $k\pi = \vartheta(g_{m+k}) - \vartheta(g_m) = \int_{g_m}^{g_{m+k}} \vartheta'(t)\, dt$ is approximately $(g_{m+k} - g_m) \cdot \frac{1}{2}\log(g_m/2\pi)$ (see Section 6.5), the second term on the right is about

$$\frac{1}{2k\pi}\left[2.30 \log \frac{g_m}{2\pi} + 0.128\left(\log \frac{g_m}{2\pi}\right)^2 + (\textstyle\sum h_j) \log \frac{g_m}{2\pi}\right].$$

As k increases, this term rapidly becomes less than one unless $\sum h_j$ is rather large and positive. In actual fact, in the range of the calculations of Rosser *et al.* (in which g_m is comfortably less than two million so $\log(g_m/2\pi)$ is comfortably less than 14), it was always possible to prove in this way that $S(g_m) < 2$ for all values of m such that $h_m = 0$, and for this purpose it was never necessary to use any value of k larger than 15 (see Section 8.4). In this way it was proved that $N(g_m) \leq m + 1$ for Gram points g_m in the vicinity of $m =$ 3,500,000. Since h's had been found all the way up to this level, no lower bound of N was necessary and it followed that all $m + 1$ of these roots are simple zeros on $\operatorname{Re} s = \frac{1}{2}$.

In the same way one can obtain a lower bound for $S(g_m)$, namely,

$$S(g_m) \geq -1 - \frac{2.30 + 0.128 \log(g_m/2\pi) - \sum_{j=1}^{k-1} h_{m-j}}{g_m - g_{m-k}}$$

where m, k are such that $h_m = 0$, $h_{m-k} = 0$. Using a bound similar to this one, Rosser *et al.* were able to prove that $N(g_m) = m + 1$ for certain Gram points g_m in the vicinity of $m =$ 13,400,000 and then to prove that 41,000 consecutive roots in this range are all simple zeros on the line $\operatorname{Re} s = \frac{1}{2}$.

8.3 LEHMER'S PHENOMENON

Lehmer's 1955–1956 computations followed Hutchinson's scheme of determining the sign of $Z(g_n)$ for the Gram points g_n and, in the cases where $(-1)^n Z(g_n) > 0$ fails to hold, of finding small numbers h_n such that $(-1)^n Z(g_n + h_n) > 0$ holds. As long as this is possible it shows that there are at least $m + 1$ zeros of $Z(t)$ (not counting multiplicities) in the range $0 < t < g_m$ and, using Turing's method, it should be possible to show that there are no more than $m + 1$ roots ρ in the range $\{0 \leq \operatorname{Im} s \leq g_m\}$ (counting multiplicities) and hence that all roots ρ in this range are simple zeros on the line $\operatorname{Re} s = \frac{1}{2}$.

The first step in carrying out this scheme is naturally to write a program for computing $Z(t)$ economically and not necessarily very accurately, with a view simply to determining the sign of $Z(t)$ for a given t. Lehmer's program computed the main sum in the Riemann–Siegel formula with an accuracy of several decimal places and computed the C_0 term very roughly. If the absolute value of the resulting estimate of $Z(t)$ was safely larger than Titchmarsh's estimate $\frac{3}{2}(t/2\pi)^{-3/4}$ of the error in the Riemann–Siegel formula, then its sign was considered to be the sign of $Z(t)$. Lehmer then began running through the list of Gram points (previously computed by a method similar to the

method outlined in Section 6.5) and in those few cases† where $Z(g_n)$ could not be shown to have the sign $(-1)^n$, found small h_n for which $Z(g_n + h_n)$ could. No difficulty was encountered until the Gram point g_{4763} was reached, but here, despite the fact that the Riemann–Siegel formula is very accurate for such large t, the values of Z were too small to find an h, although there was a range in which the sign of $Z(g + h)$ was not conclusively determined. More exact computations‡ did show that Z has the required number of sign changes in this region, but the appearance of this phenomenon of a region in which Z just *barely* changes sign was of very great interest.

If there were a point at which the graph of $Z(t)$ came near to $Z = 0$ but did not actually cross it—that is, if Z had a small positive local minimum or a small negative local maximum—*then the Riemann hypothesis would be contradicted.* This theorem can be proved as follows. It will suffice to show that the Riemann hypothesis implies Z'/Z is monotone because this, in turn, implies that Z'/Z has at most one zero between successive zeros of Z and therefore that Z' has at most one zero between successive zeros of Z. Now $\xi(\frac{1}{2} + it) = -f(t)Z(t)$, where

$$f(t) = |\Pi[(s/2) - 1]|\pi^{-1/4}\tfrac{1}{2}(t^2 + \tfrac{1}{4})$$

(see Section 6.5). Thus

$$\frac{(d/dt)\xi(\frac{1}{2} + it)}{\xi(\frac{1}{2} + it)} = \frac{f'(t)}{f(t)} + \frac{Z'(t)}{Z(t)}$$

$$-\frac{Z'(t)}{Z(t)} = -i\frac{\xi'(\frac{1}{2} + it)}{\xi(\frac{1}{2} + it)} + \frac{f'(t)}{f(t)}.$$

Logarithmic differentiation of the product formula for ξ (justified in Chapter 3) puts the first term on the right in the form

$$-i\sum_{\rho}\frac{1}{(\frac{1}{2} + it) - \rho} = \sum_{\alpha}\frac{-i}{it - i\alpha} = \sum_{\alpha}\frac{1}{\alpha - t},$$

where, as before, $\rho = \frac{1}{2} + i\alpha$ runs over all roots ρ. Thus the derivative of the first term on the right is $\sum(\alpha - t)^{-2}$ and, if the Riemann hypothesis is true, this derivative is not only positive (because all terms are positive) but also very large [because by von Mangoldt's estimate of $N(t)$ the α's must be quite dense] between successive zeros of Z. The second term on the right can be rewritten

$$\frac{f'(t)}{f(t)} = \frac{d}{dt}\operatorname{Re}\log\Pi\left(\frac{\frac{1}{2} + it}{2} - 1\right) + \frac{2t}{t^2 + \frac{1}{4}}.$$

†There were about 360 such cases altogether in the first 5000 Gram points, and about 100 such in each succeeding 1000 up to 25,000.

‡Lehmer used the Euler–Maclaurin formula for these in order to have a rigorous error estimate.

The second term in this formula does decrease as t increases, but its decrease is obviously insignificant compared to the increase of the term already considered. The derivative of the first term can be estimated using Stirling's series

$$\begin{aligned}\frac{d^2}{dt^2}\operatorname{Re}\log\Pi\Big(-\frac{3}{4}+\frac{it}{2}\Big) &= \operatorname{Re}\frac{d^2}{dt^2}\log\Pi\Big(-\frac{3}{4}+\frac{it}{2}\Big)\\ &= \Big(\frac{i}{2}\Big)^2\operatorname{Re}\frac{d^2}{ds^2}\Big|_{s=-(3/4)+(it/2)}\log\Pi(s)\\ &= -\frac{1}{4}\operatorname{Re}\Big\{\frac{1}{s}-\frac{1}{2s^2}+\frac{B_2}{s^3}+\frac{B_4}{s^5}+\cdots\Big\}\Big|_{s=-(3/4)+(it/2)}\end{aligned}$$

The first two terms give

$$\begin{aligned}\operatorname{Re}\Big\{\frac{1-2s}{8s^2}\Big\} &= \operatorname{Re}\Big\{\frac{(1-2s)\bar{s}^2}{8|s|^4}\Big\}\\ &= \frac{\operatorname{Re}\Big\{\Big(\frac{5}{2}-it\Big)\Big(\frac{9}{16}+\frac{3it}{4}-\frac{t^2}{4}\Big)\Big\}}{8|s|^4}\\ &= \frac{-\frac{45}{32}-\frac{5}{8}t^2+\frac{3}{4}t^2}{8|s|^4}\end{aligned}$$

which is positive and of the order of magnitude of t^{-2} for large t. The other terms of Stirling's series are insignificant in comparison with this one, and it follows that the first term in the formula for f'/f increases with t. Therefore $-Z'/Z$ is an increasing function of t (for t at all large) and the theorem is proved. This shows that the place near g_{4763} where Z just barely crosses the axis is "almost" a counterexample to the Riemann hypothesis.

Pursuing the calculations on up to $g_{25,000}$, Lehmer found more of these "near counterexamples" to the Riemann hypothesis. One of these on which he gives very complete information occurs for g_{6708} (see Fig. 3†). There are three zeros between g_{6704} and g_{6705}, and to achieve the correct sign for $Z(g_n + h_n)$ in these cases, a positive value of h_{6704} and a negative value of h_{6705} are necessary. Then $(-1)^n Z(g_n) > 0$ for $n = 6706,6707$, but to obtain the correct sign for $Z(g_n + h_n)$ when $n = 6708$, h_n must be negative and must be chosen so that $g_n + h_n$ lies in the very short range of the "near counterexample" where the graph of Z just barely crosses over to positive values between g_{6707} and g_{6708}. This range where Z is positive has length 0.0377 (difference of t values of the two zeros), and the largest value of Z which

†One question which naturally presents itself when one examines the diagram is the question of the size of $Z(t)$. Will it always remain this small? A proof that $|Z(t)| = |\zeta(\frac{1}{2}+it)|$ is in fact *unbounded* as $t \longrightarrow \infty$ is outlined in Section 9.2. The *Lindelöf hypothesis* is that $|Z(t)|$ grows more slowly than any positive power of t—that is, for every $\epsilon > 0$ it is true that $Z(t)/t^\epsilon \longrightarrow 0$ as $t \longrightarrow \infty$—but this has never been proved or disproved.

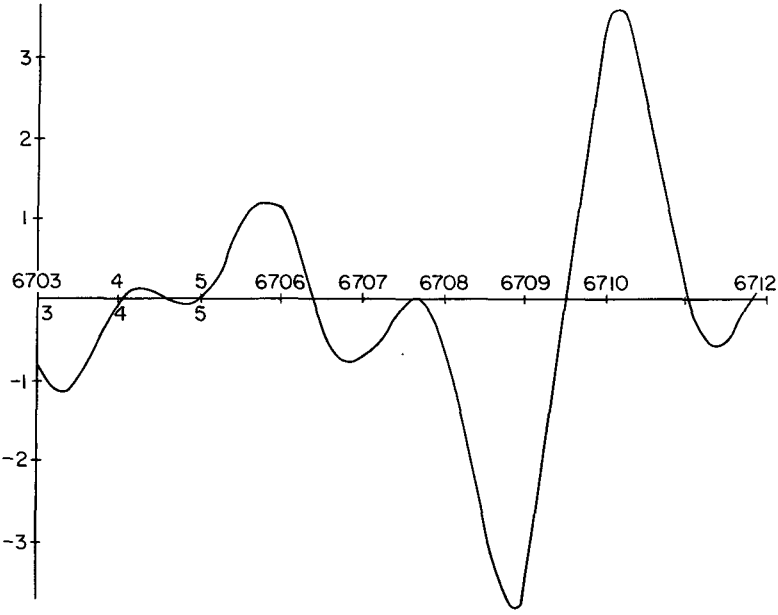

Fig. 3 The scale on the vertical axis is the value of $Z(t)$. The scale on the horizontal axis shows the location of the Gram points g_{6703}, g_{6704}, etc. (From D. H. Lehmer, On the roots of the Riemann zeta-function. *Acta Mathematica* **95**, 291–298, 1956, with the permission of the author).

occurs in it is only about 0.00397, so the choice of h_{6708} is a delicate task requiring quite accurate values of Z. Note that this "near counterexample" is followed by a very strong oscillation of Z as the terms go from a high degree of cancellation to a high degree of reinforcement. The degree of irregularity of Z shown by this graph of Lehmer, and especially the low maximum value between g_{6707} and g_{6708}, must give pause to even the most convinced believer in the Riemann hypothesis.

The extremum of lowest absolute value in the range of the first 25,000 Gram points is reported by Lehmer to be the value $+0.002$ at $t = 17143.803905$. Here the Riemann–Siegel formula gives $Z(t) \sim 0.002\,153\,336$. This low local maximum occurs between the two most nearly coincident zeros of Z found by Lehmer, namely, the zeros at† $t = 17143.786\,536$ and $t = 17143.821\,844$. To see how very close this low maximum comes to being a counterexample to the Riemann hypothesis, note that it completely de-

†The values given by Lehmer [L8] are incorrect. The values above are taken from Haselgrove's tables [H8].

main sum	0.073 478 610
C_0 term	−0.071 297 360
C_1 term	−0.000 027 686
C_2 term	−0.000 000 227
C_3 term	+0.000 000 001
$Z(17143.803905)$	+0.002 153 336.

Computation of the approximation to $Z(17143.803905)$.

stroys the rationale of Gram's law, according to which $Z(g_n)$ should have a tendency to have the sign $(-1)^n$ because the first term of the Riemann–Siegel formula is $2\cdot(-1)^n$ and because the terms decrease in absolute value. In the example above, whether because the terms of the main sum tend to be zero or whether because there is large-scale cancellation of terms, the *entire* main sum amounts to only 0.073; this is of the same order of magnitude as the C_0 term which, as was seen in Section 7.6, is of the same order of magnitude as the last terms of the main sum. In short, the determination of the sign of $Z(t)$ can be a very delicate matter involving even the smallest terms in the main sum of the Riemann–Siegel formula, and although *on the average* one can expect the sign to be determined by the largest terms, there is no obvious reason why the exceptions to this statement could not include a counterexample to the Riemann hypothesis.

Subsequent calculations have so far fulfilled Lehmer's prediction that this phenomenon would recur infinitely often. For example, Rosser *et al.* found a pair of zeros in the vicinity of the 13,400,000th zero which were separated by just 4.4×10^{-4} (whereas the distance between successive Gram points in this region is about $0.07 = 700 \times 10^{-4}$) and between which $|Z|$ is less than 7.1×10^{-5}. It would be interesting to have a graph, such as Lehmer's above, of the behavior of Z in this region.

8.4 COMPUTATIONS OF ROSSER, YOHE, AND SCHOENFELD

At the Mathematics Research Center in Madison, Wisconsin, there are three reels of magnetic tape containing three and a half million triples of numbers (G_n, Z_n, ϵ_n) such that $G_{n+1} > G_n$, $(-1)^n Z_n > 0$, $|Z_n| > \epsilon_n > 0$, and such that, according to a rigorous analysis of the error in the Riemann–Siegel formula,† the value of $Z(G_n)$ differs from Z_n by less than ϵ_n. These tapes, unless they contain an error‡ prove the existence of three and a half million

†In the computations the C_3 term was the first term omitted.

‡This is a proviso which applies, after all, to any "proof" (see Lakatos [L1]).

roots ρ on the line $\operatorname{Re} s = \frac{1}{2}$ and locate them between $\frac{1}{2} + iG_n$ and $\frac{1}{2} + iG_{n+1}$. Moreover, by applying Turing's method to the last 15 or so of the G_n ($= g_n + h_n$ in the notation of Section 8.2) these data prove that there are *only* three and a half million roots with imaginary parts in this range (counting multiplicities), hence that for each n there is precisely one root ρ in the range $\{G_n \leq \operatorname{Im} s \leq G_{n+1}\}$ and it is a simple zero on the line $\operatorname{Re} s = \frac{1}{2}$. There is a fourth reel of tape which proves in similar fashion that 41,600 consecutive roots ρ beginning with the 13,400,000th are simple zeros on the line.

In the course of the computations of which these tapes are the record, Rosser, Yohe, and Schoenfeld discovered the following interesting phenomenon. Let a Gram point be called "good" if $(-1)^n Z(g_n) > 0$ and "bad" otherwise. Rosser *et al.* called the interval between two consecutive good Gram points a "Gram block"—that is, a Gram block is an interval $\{g_n \leq t \leq g_{n+k}\}$, where g_n, g_{n+k} are good Gram points but $g_{n+1}, g_{n+2}, \ldots, g_{n+k-1}$ are bad—and they found somewhat to their surprise that in the range of their computations *every Gram block contains the expected number of roots.* Let this be called "Rosser's rule." This phenomenon, as long as it continues, is of obvious usefulness in locating roots. However, Rosser *et al.* express a belief that it will not continue forever and this belief can be proved† to be correct as follows.

To be specific, let Rosser's rule be the statement that in any Gram block $\{g_n \leq t \leq g_{n+k}\}$ there are at least k changes of sign of Z. This implies, since $\pi^{-1}\vartheta(t) + 1$ increases by exactly k on the block, that $S(g_{n+k}) \geq S(g_n)$. Therefore, by induction it implies $S(g_n) \geq 0$ for all good Gram points g_n. On the other hand, if g_m, g_{m+1} are both bad, then Z changes sign on $\{g_m \leq t \leq g_{m+1}\}$, from which $S(g_{m+1}) \geq S(g_m)$. Thus after a good Gram point the value of S could drop by one at the next Gram point (assuming it is bad), but thereafter it could drop no further until a good Gram point were reached, at which time it would have to return at least to its former value by Rosser's rule. Thus, in particular, Rosser's rule implies $S(g_n) \geq -1$ for all Gram points g_n. Now this implies that the Riemann hypothesis is true because, if it were false, then there would be an increase of 2 in $N(T)$ not counted in the above estimates (which counted only sign changes of Z, hence roots on the line) which together with the above estimates gives $S(g_n) \geq 1$ for all Gram points g_n past the supposed counterexample to the Riemann hypothesis and hence $S(t) \geq 0$ beyond this point. This pretty clearly contradicts Littlewood's theorem that the average of S is zero and a rigorous proof is not hard to give. However, the actual estimates can be avoided by the method given below. For the moment assume it has been shown that Rosser's rule implies there

†This proof is in essence the one given by Titchmarsh [T5] to prove that Gram's law $\operatorname{Re} \zeta(\frac{1}{2} + ig_n) > 0$ fails infinitely often.

are no counterexamples to the Riemann hypothesis. Then by a 1913 theorem of Bohr and Landau (see Section 9.8) it follows that $S(t)$ is neither bounded above nor bounded below. In particular $S(t) \geq -2$ is false and this contradiction proves that Rosser's rule cannot hold.

More generally this argument can be used to show that there are *infinitely many* exceptions to Rosser's rule, that is, infinitely many Gram blocks with fewer than the expected number of sign changes. One need only observe that otherwise the argument above shows that once one is past all exceptions to Rosser's rule, the value of $S(g_n)$ and hence of $S(t)$ is bounded below. Then there are at most finitely many exceptions to the Riemann hypothesis, since otherwise $S(t)$ would eventually be large and positive, contradicting Littlewood's theorem. However, as Titchmarsh observes, the theorem of Bohr and Landau remains true if there are only finitely many exceptions to the Riemann hypothesis and it implies that $S(t)$ cannot then be bounded below. (Thus the proof above that Rosser's rule implies *no* exceptions to the Riemann hypothesis can be omitted.) This contradiction proves that there must be infinitely many exceptions to Rosser's rule.

From the fact that not a single exception to Rosser's rule has yet been found it is tempting to conclude that the computations have not yet reached the real irregularities of $Z(t)$. But actually Rosser's rule is not in any way a measure of the "regularity" of $Z(t)$. On the contrary, it measures only the success of a rather crude attempt to predict the oscillations in sign of $Z(t)$ [and hence of $\xi(\frac{1}{2} + it)$], an attempt which in fact has proved far more successful than in all likelihood Gram imagined it would be when he first proposed it.

Note added in second printing: This fact, that Rosser's rule fails infinitely often, was proved by R. Sherman Lehman (On the distribution of zeros of the Riemann zeta function, *Proc. Lon. Math. Soc.* (3) XX (1970) 303–320). In this same paper Lehman points out errors in Turing's proof of the main estimate (1) of Section 8.2. However, he replaces Turing's proof with his own proof of a slightly stronger inequality.

Chapter 9

The Growth of Zeta as $t \to \infty$ and the Location of Its Zeros

9.1 INTRODUCTION

The problem of locating the roots ρ of ζ, and consequently the problem of estimating the error in the prime number theorem, is closely related to the problem of estimating the growth of ζ in the critical strip $\{0 \leq \operatorname{Re} s \leq 1\}$ as $\operatorname{Im} s \to \infty$. Evidence of the relation between these two problems can be seen in Section 4.2 where the main step in the proof of the prime number theorem depended on estimates of $\operatorname{Re} \log \zeta(\sigma + it) = \log|\zeta(\sigma + it)|$ for σ near 1 and for all t, in Section 5.2 where the main step in de la Vallée Poussin's estimate of the error in the prime number theorem depended on estimates of $\zeta'(\sigma + it)/\zeta(\sigma + it)$ for large t, and in Section 6.7 where Backlund's proof of Riemann's estimate of $N(T)$ depended on estimates of the growth of $|\xi(s)|$ in the strip $0 \leq \operatorname{Re} s \leq 4$.

A major landmark in the study of ζ in the critical strip is Lindelöf's 1908 paper [L11] in which he not only proved some estimates which were far stronger than those that had been established previously and introduced new techniques and theorems basic to subsequent studies, but in which he also enunciated the famous "Lindelöf hypothesis." This paper is the subject of Section 9.2. In Section 9.3 a brief but important note [L12] written by Littlewood in 1912 is discussed; this note is important because it introduced the use of the three circles theorem and showed that the Riemann hypothesis implies the Lindelöf hypothesis. In 1918 Backlund proved a more exact result to the effect that the Lindelöf hypothesis implies and is implied by a certain statement about the location of the roots ρ which is much weaker than the Riemann hypothesis; this result is proved in Section 9.4. The following section, 9.5, is devoted to Littlewood's theorem, mentioned in Section 8.2, that

$\int_0^T S(t)\,dt$ grows no faster than a constant times $\log T$ as $T \to \infty$ [where $S(T)$ is the error in the approximation $N(T) \sim \pi^{-1}\vartheta(t) + 1$ or, what was seen in Section 6.7 to be the same, is π^{-1} Im $\log \zeta(\frac{1}{2} + iT)$]. The difficult step in this proof is the estimation of Re $\log \zeta(s)$ for s in the critical strip with large imaginary part. In Section 9.6 the theorem of Bohr and Landau is proved which states (see Section 1.9) that the relative error in the approximation "the number of roots ρ with imaginary parts between 0 and T which lie within δ of Re $s = \frac{1}{2}$ is, for every $\delta > 0$, approximately equal to the total number of roots ρ with imaginary parts in this range" approaches zero as $T \to \infty$ for fixed δ. Here the estimate of $|\zeta(s)|$ which is needed is an estimate of the *average* value of $|\zeta(s)|^2$ on lines Re s = const. These averages are evaluated in Section 9.7. Finally, Section 9.8 is devoted to the enunciation, without proof, of various other theorems of this sort on the growth of ζ in the critical strip and the location of the roots ρ.

9.2 LINDELÖF'S ESTIMATES AND HIS HYPOTHESIS

The estimate $|\zeta(\sigma + it)| = |\Sigma n^{-\sigma-it}| \leq \Sigma n^{-\sigma} = \zeta(\sigma)$ shows that for $\sigma > 1$ the modulus of $\zeta(\sigma + it)$ is *bounded* as $t \to \infty$. On the other hand, it is not difficult to show† that $\zeta(\sigma)$ is the *least* upper bound of $|\zeta(\sigma + it)|$ as $t \to \infty$ because values of t can be chosen to make $n^{-\sigma-it} = n^{-\sigma}[\cos(t \log n) - i \sin(t \log n)]$ nearly equal to $n^{-\sigma}$ for as many values of n as desired so that for any $\epsilon > 0$ arbitrarily large t can be chosen to make $|\zeta(\sigma + it)| > \zeta(\sigma) - \epsilon$. Since $\zeta(\sigma) \to \infty$ as $\sigma \downarrow 1$, this shows that $|\zeta(s)|$ is *not* bounded on the quarterplane $\{\text{Re } s > 1, \text{Im } s > 1\}$. As for the line $\{\text{Re } s = 1\}$, Mellin [M4] showed in 1900 that on it the growth of $|\zeta(s)|$ is no more rapid than the growth of $\log t$ as $t \to \infty$, an estimate which Lindelöf proves very simply by using Euler–Maclaurin summation to write

$$\zeta(s) = \sum_{n=1}^{N-1} n^{-s} + \frac{N^{1-s}}{s-1} + \frac{1}{2}N^{-s} - s\int_N^\infty \bar{B}_1(x)x^{-s-1}\,dx,$$

$$|\zeta(1 + it)| \leq 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{N-1} + \frac{1}{t} + \frac{1}{2N} + |1 + it|\int_N^\infty \frac{1}{2}x^{-2}\,dx$$

which with N equal to the greatest integer less than or equal to t gives easily $|\zeta(1 + it)| < \log t +$ const as desired. This makes it reasonable to expect

†See, for example, Titchmarsh [T3, pp. 6–7].

that although $|\zeta(s)|$ is unbounded on $\{\text{Re } s \geq 1\}$, its growth is less rapid than $\log t$. This is an immediate consequence of the following important generalization of the maximum modulus theorem to a particular type of noncompact domain.

Lindelöf's Theorem Let $f(s)$ be defined and analytic in a halfstrip $D = \{s: \sigma_1 \leq \text{Re } s \leq \sigma_2, \text{Im } s \geq t_0 > 0\}$. If the modulus of f is less than or equal to M on the boundary ∂D of D *and* if there is a constant A such that $|f(\sigma + it)|t^{-A}$ is bounded on D, then the modulus of f is less than or equal to M throughout D.

Proof Consider the function $\log|f(s)|$ which is a real-valued harmonic function defined throughout D except for singularities at the zeros of $f(s)$ where it approaches $-\infty$. The additional growth condition on f states that $\log|f(\sigma + it)| < A \log t + \text{const}$ on D for A sufficiently large. This implies that for any $\epsilon > 0$ the harmonic function $\log|f(s)| - \epsilon t$ is less than any given constant on the line segment $\{\text{Im } s = T, \sigma_1 \leq \text{Re } s \leq \sigma_2\}$ provided T is large enough to make ϵT much larger than $A \log T$. In particular, for T sufficiently large $\log|f(s)| - \epsilon t$ is less than $\log M$ on the boundary of the rectangle $\{\sigma_1 \leq \text{Re } s \leq \sigma_2, t_0 \leq \text{Im } s \leq T\}$; hence the same inequality holds throughout the rectangle and therefore throughout the half-strip, and it follows that $|f(s)| \leq e^{\epsilon t}M$ throughout the half-strip. Since ϵ was arbitrary, this implies Lindelöf's theorem.

Corollary 1 $|\zeta(\sigma + it)|/\log t$ is bounded for $\sigma \geq 1$, $t \geq 2$.

Proof Since $|\zeta(\sigma + it)|$ is bounded for $\sigma \geq 1 + \delta$ by $|\zeta(\sigma + it)| \leq \zeta(\sigma) \leq \zeta(1 + \delta)$, it suffices to consider the half-strip $\{1 \leq \text{Re } s \leq 1 + \delta, \text{Im } s \geq 2\}$. Within this half-strip $\log s$ differs by a bounded amount from $\log t$. Moreover, $|\zeta(\sigma + it)| \leq \text{const } t^2$ in the half-strip, as was proved in Section 6.7 using Euler–Maclaurin summation. Combining these observations with Mellin's estimate of $|\zeta(1 + it)|$ shows that Lindelöf's theorem applies to $\zeta(s)/\log s$ on the half-strip and the corollary follows.

Corollary 2 $|\zeta(s)|$ is not bounded on any line $\text{Re } s = \sigma$ for $\sigma \leq 1$.

Proof If it were bounded, then Lindelof's theorem would show that $|\zeta(s)|$ was bounded on a half-strip which included $\{1 \leq \sigma \leq 2, t \geq 1\}$, contrary to the fact that $|\zeta(s)|$ is unbounded on $\{\sigma > 1, t \geq 1\}$.

Thus the general pattern is for the values of $|\zeta(s)|$ on $\text{Re } s = \sigma$ to be greater as σ decreases, at least in the range considered above. In the range

$\operatorname{Re} s \leq 0$ the functional equation combined with Stirling's formula can be used to estimate the growth of $|\zeta(s)|$ as $\operatorname{Im} s \to \infty$ as follows:

$$\begin{aligned}
\log|\xi(\sigma+it)| &= \operatorname{Re}\log\xi(\sigma+it)\\
&= \operatorname{Re}\log\Pi\Big(\frac{\sigma+it}{2}\Big) - \frac{\sigma}{2}\log\pi + \log|\sigma-1+it|\\
&\quad + \log|\zeta(\sigma+it)|\\
&\sim \frac{\sigma+1}{2}\log\Big|\frac{s}{2}\Big| - \frac{t}{2}\operatorname{Im}\log\frac{s}{2} - \frac{\sigma}{2} + \frac{1}{2}\log 2\pi\\
&\quad - \frac{\sigma}{2}\log\pi + \log|t| + \log|\zeta(\sigma+it)|\\
&\sim \frac{\sigma}{2}\Big(\log\frac{t}{2} - 1 - \log\pi\Big) - \frac{t}{2}\cdot\frac{\pi}{2} + \frac{3}{2}\log t\\
&\quad - \frac{1}{2}\log 2 + \frac{1}{2}\log 2\pi + \log|\zeta(\sigma+it)|\\
&\sim \frac{\sigma}{2}\log\frac{t}{2\pi e} - \frac{t\pi}{4} + \frac{3}{2}\log t + \frac{1}{2}\log\pi\\
&\quad + \log|\zeta(\sigma+it)|
\end{aligned}$$

where the error in the approximation approaches zero as $t \to \infty$ for fixed σ. Thus

$$\begin{aligned}
0 &= \log|\xi(\sigma+it)| - \log|\xi(1-\sigma+it)|\\
&\sim \frac{\sigma}{2}\log\frac{t}{2\pi e} - \frac{1-\sigma}{2}\log\frac{t}{2\pi e} + \log|\zeta(\sigma+it)|\\
&\quad - \log|\zeta(1-\sigma+it)|
\end{aligned}$$

and finally

$$1 \sim \Big(\frac{t}{2\pi e}\Big)^{\sigma-(1/2)}\Big|\frac{\zeta(\sigma+it)}{\zeta(1-\sigma+it)}\Big|, \tag{1}$$

where the error in the approximation approaches zero as $t \to \infty$ for fixed σ. Thus for $\operatorname{Re} s = 0$ the modulus of $\zeta(s)$ is less than a constant times $t^{1/2}\log t$ and on any line $\operatorname{Re} s = \sigma < 0$ it is less than a constant times $t^{(1/2)-\sigma}$. Moreover, the last estimate is a least upper bound. This gives a satisfactory description of the growth of $|\zeta(s)|$ on lines $\operatorname{Re} s = \sigma \leq 0$ and shows that this growth becomes more rapid as σ decreases.

For lines $\operatorname{Re} s = \sigma$ inside the critical strip $0 < \sigma < 1$ the above estimates do not apply. However, Lindelöf observed that an upper bound for the growth of $|\zeta(s)|$ on such lines can be obtained by linear interpolation of the estimates for $\sigma = 0$ and $\sigma = 1$. More precisely, he observed that there is a constant K such that $|\zeta(\sigma+it)| < Kt^{(1/2)-(1/2)\sigma}\log t$ throughout the half-strip

$\{0 \leq \operatorname{Re} s \leq 1, \operatorname{Im} t \geq 1\}$. (The exponent $\frac{1}{2} - \frac{1}{2}\sigma$ is the affine function which is $\frac{1}{2}$ for $\sigma = 0$ and 0 for $\sigma = 1$.) This is a consequence of the following general theorem.

Modified Lindelöf's Theorem Let $f(s)$ be defined and analytic in a half-strip $D = \{s: \sigma_1 \leq \operatorname{Re} s \leq \sigma_2, \operatorname{Im} s \geq t_0 > 0\}$. If p, q are such that the modulus of f is less than a constant times t^p on $\operatorname{Re} s = \sigma_1$ and less than a constant times t^q on $\operatorname{Re} s = \sigma_2$ *and* if there is a constant A such that $|f(\sigma + it)|t^{-A}$ is bounded on D, then there is a constant K such that $|f(\sigma + it)| \leq Kt^{k(\sigma)}$ throughout D, where $k(\sigma) = [(q - p)/(\sigma_2 - \sigma_1)](\sigma - \sigma_1) + p$ is the affine function which is p at σ_1 and q at σ_2.

Proof Apply the previous argument using the harmonic function $\log|f(s)| - k(\sigma)t - \epsilon t$ instead of $\log|f(s)| - \epsilon t$.

Lindelöf denotes† by $\mu(\sigma)$ the least upper bound of the numbers A such that $|\zeta(\sigma + it)|t^{-A}$ is bounded as $t \to \infty$. Otherwise stated, $\mu(\sigma)$ is characterized by the condition that $|\zeta(\sigma + it)|$ divided by $t^{\mu(\sigma)+\epsilon}$ is bounded as $t \to \infty$ if $\epsilon > 0$ but unbounded if $\epsilon < 0$. The above estimates show that $\mu(\sigma) = 0$ for $\sigma \geq 1$ and that $\mu(\sigma) = \frac{1}{2} - \sigma$ for $\sigma \leq 0$. Formula (1) shows that $\mu(\sigma)$ satisfies the functional equation $\mu(\sigma) = \mu(1 - \sigma) + \frac{1}{2} - \sigma$. The modified Lindelöf's theorem shows that $\mu(\sigma) \leq \frac{1}{2} - \frac{1}{2}\sigma$ for $0 \leq \sigma \leq 1$ and, more generally, it shows that $\mu(\sigma)$ is convex downward in the sense that any segment of the graph of μ lies below the line joining its endpoints. This implies that $\mu(\sigma) \geq 0$ for $\sigma < 1$ and that if $\mu(\sigma_0) = 0$ for some $\sigma_0 < 1$, then necessarily $\mu(\sigma) = 0$ for $\sigma_0 < \sigma < 1$.

The so-called‡ *Lindelöf hypothesis* is that $\mu(\sigma)$ is the simplest§ function which has all the above properties, namely, the function which is zero for $\sigma \geq \frac{1}{2}$ and $\frac{1}{2} - \sigma$ for $\sigma \leq \frac{1}{2}$. By the convexity of μ the Lindelöf hypothesis is equivalent to the hypothesis that $\mu(\frac{1}{2}) = 0$. It is shown in Section 9.4 that the Lindelöf hypothesis implies and is implied by a condition on the location of the roots ρ which is weaker than the Riemann hypothesis, so the Lindelöf hypothesis is, in Titchmarsh's phrase, less drastic than the Riemann hypothesis. Nonetheless, its proof appears to be no easier and it has never been proved or disproved.

†This μ is not, of course, related in any way to the Möbius function $\mu(n)$ defined in Section 5.6.

‡Actually Lindelöf conjectured that $|\zeta(s)|$ is *bounded* on $\operatorname{Re} s = \sigma$ for $\sigma > \frac{1}{2}$, a conjecture which was shown above to be false.

§Another possibility is $\mu(\sigma) = \frac{1}{2} - \frac{1}{2}\sigma$ on $0 \leq \sigma \leq 1$. However, it is known (see Section 9.8) that $\mu(\frac{1}{2}) < \frac{1}{4}$, so this possibility is excluded.

9.3 THE THREE CIRCLES THEOREM

In 1912 Littlewood introduced a new technique into the study of the growth of ζ when he published in a brief note [L12] some new estimates obtained using the theorem now known as the "three circles" theorem. Littlewood attributes the theorem to no one, saying it was discovered independently by several authors, but Bohr and Landau [B7] say the theorem was first published by Hadamard in 1896 (although Hadamard published no proof) and it is now commonly known as "Hadamard's three circles theorem." In any case, the theorem consists of the following simple observations.

Let $f(s)$ be defined and analytic on an annulus $D = \{r_1 \leq |s - s_0| \leq r_3\}$. Given an upper bound M_3 for the modulus of f on the outer circle $|s - s_0| = r_3$ and an upper bound M_1 for the modulus of f on the inner circle $|s - s_0| = r_1$ the problem is to find an upper bound M_2 for the modulus of f on a concentric circle $|s - s_0| = r_2$ inside the annulus. The method is to consider the harmonic function $\log|f(s)|$ on the annulus D and to compare it to a harmonic function which is identically $\log M_1$ on the inner circle and identically $\log M_3$ on the outer circle. Such a function is easily found by applying linear interpolation and the fact that $a \log|s - s_0| + b$ is a two-parameter family of harmonic functions constant on circles $|s - s_0| = \text{const}$ (note the analogy with the modified Lindelöf theorem in the preceding section). This leads to consideration of the harmonic function

$$H(s) = \frac{(\log M_3 - \log M_1)\log|s - s_0| + \log M_1 \log r_3 - \log M_3 \log r_1}{\log r_3 - \log r_1}.$$

Since $H(s) \geq \log|f(s)|$ on ∂D and since both are harmonic [except that $\log|f(s)|$ may have singularities at zeros of f where it is $-\infty$], the same inequality holds throughout D, which for $|s - s_0| = r_2$ gives

$$\frac{(\log M_3 - \log M_1)\log r_2 + \log M_1 \log r_3 - \log M_3 \log r_1}{\log r_3 - \log r_1} \geq \log M_2$$

which simplifies to

$$\log M_1 \log \frac{r_3}{r_2} + \log M_3 \log \frac{r_2}{r_1} \geq \log M_2 \log \frac{r_3}{r_1}$$

or

$$M_1^{\log(r_3/r_2)} M_3^{\log(r_2/r_1)} \geq M_2^{\log(r_3/r_1)}$$

which is the desired estimate. Otherwise stated, $M_2 \leq M_1^{\alpha} M_3^{\beta}$, where $\alpha + \beta = 1$ and $\alpha : \beta = \log(r_3/r_2) : \log(r_2/r_1)$, in which form M_2 appears as a sort of mean value between M_1 and M_3.

As a simple application of this theorem, Littlewood proved that *if the Riemann hypothesis is true, then for every* $\epsilon > 0$ *and* $\delta > 0$ *the function* $|\log \zeta(\sigma + it)|$ *is less than a constant times* $(\log t)^{2-2\sigma+\epsilon}$ *on the half-strip* $\{\frac{1}{2} + \delta \leq \sigma \leq 1, t \geq 2\}$, where $\log \zeta(s)$ is defined for $\operatorname{Re} s > \frac{1}{2}$ by virtue of the Riemann hypothesis. Since $\sigma > \frac{1}{2}$ implies $2 - 2\sigma + \epsilon < 1$ for ϵ sufficiently small, this shows that the Riemann hypothesis implies $\log|\zeta(\sigma + it)| = \operatorname{Re} \log \zeta(\sigma + it) \leq |\log \zeta(\sigma + it)| \leq K \log t (\log t)^{-\theta}$, where $\theta > 0$; hence for any $\epsilon' > 0$ it follows that $\log|\zeta(\sigma + it)| \leq \epsilon' \log t$ for all sufficiently large t; hence $|\zeta(s)| < t^{\epsilon'}$ on $\operatorname{Re} s = \sigma > \frac{1}{2}$ as $t \to \infty$—in short, *the Riemann hypothesis implies the Lindelöf hypothesis.* Since this is the consequence of Littlewood's theorem which is of principal† interest in this chapter and since it is subsumed in Backlund's proof (which took its inspiration from Littlewood) in the next section, the details of Littlewood's application of the three circles theorem will not be given here.

9.4 BACKLUND'S REFORMULATION OF THE LINDELÖF HYPOTHESIS

Backlund [B4] proved in 1918 that the *Lindelöf hypothesis is equivalent to the statement that for every* $\sigma > \frac{1}{2}$ *the number of roots in the rectangle* $\{T \leq \operatorname{Im} s \leq T + 1, \sigma \leq \operatorname{Re} s \leq 1\}$ *grows less rapidly than* $\log T$ *as* $T \to \infty$—more precisely, it is equivalent to the statement that for every $\epsilon > 0$ there is a T_0 such that the number of such zeros is less than $\epsilon \log T$ whenever $T \geq T_0$. It follows from this that the Riemann hypothesis implies the Lindelöf hypothesis because if the Riemann hypothesis is true, then there are never *any* zeros in the rectangle in question.

The implication in one direction is quite an easy consequence of Jensen's theorem. Consider a circle which passes through the points $\sigma + iT$ and $\sigma + i(T + 1)$ and which lies in $\operatorname{Re} s > \frac{1}{2}$. Let $\sigma_0 + i(T + \frac{1}{2}) = s_0$ be the center of this circle and let ρ be its radius. Finally, let r be the radius of the slightly larger circle concentric with this one and tangent to the line $\{\operatorname{Re} s = \frac{1}{2}\}$. Then by Jensen's theorem $\log|\zeta(s_0)| + \sum \log(r/|s_j - s_0|) \leq M$ where M is the maximum of $\log|\zeta(s)|$ on the larger circle and the sum on the left is over the zeros s_j of $\zeta(s)$ in the larger circle. If this sum is restricted to the zeros which lie in the rectangle $\{T \leq \operatorname{Im} s \leq T + 1, \sigma \leq \operatorname{Re} s \leq 1\}$, then it contains n terms, where n is the integer to be estimated, and each of them is at least $\log(r/\rho)$; hence

$$n \cdot \log(r/\rho) \leq M - \log|\zeta(s_0)|.$$

†The consequence which Littlewood was principally interested in, however, was that the Riemann hypothesis implies convergence throughout the halfplane $\operatorname{Re} s > \frac{1}{2}$ of the Dirichlet series $[\zeta(s)]^{-1} = \sum \mu(n) n^{-s}$ of Section 5.6. For a proof of this see Section 12.1.

Now $\log(r/\rho)$ is a positive number independent of T and†

$$-\log|\zeta(s_0)| = \log\frac{1}{|\zeta(s_0)|} = \log\left|\sum_n \mu(n)n^{-s_0}\right| \leq \log\sum_n n^{-\sigma_0} = \log\zeta(\sigma_0)$$

for all t. If the Lindelöf hypothesis is true, then for every $\epsilon > 0$ there is a K such that $|\zeta(\sigma + it)| \leq Kt^{\epsilon}$ for $\frac{1}{2} \leq \sigma \leq \sigma_0 + r$; hence $M \leq \log[K(T+\frac{1}{2}+r)^{\epsilon}] = \epsilon\log(T+\frac{1}{2}+r) + \log K \leq 2\epsilon\log T$ for T sufficiently large; hence $n\log(r/\rho) \leq 3\epsilon\log T$ for T sufficiently large. Since $\epsilon > 0$ is arbitrary, this shows that the Lindelöf hypothesis implies n grows more slowly than $\log T$.

Backlund's proof of the converse uses the actual function which is used in the proof of Jensen's theorem in Section 2.2. Let $\sigma_0 > 1$ be fixed, let $s_0 = \sigma_0 + iT$, where, as before, T will go to infinity, let σ be fixed in the range $\frac{1}{2} < \sigma < 1$, let C be the circle with center s_0 tangent to the line $\operatorname{Re} s = \sigma$ so that the radius R of C is $\sigma_0 - \sigma$, let $s_1, s_2, \ldots, s_n$ be the zeros of ζ (counted with multiplicities) inside C, and let

$$F(s) = \zeta(s)\prod_{\nu=1}^{n}\frac{R^2 - (\bar{s}_\nu - \bar{s}_0)(s - s_0)}{R(s - s_\nu)}. \tag{1}$$

(It will be assumed that ζ has no zeros on C, a condition which excludes at most a discrete set of T's.) The s_ν are contained in a finite number of rectangles of the form $\{T' \leq \operatorname{Im} s \leq T' + 1, \sigma \leq \operatorname{Re} s \leq 1\}$, and the number of these rectangles is independent of T; hence what is to be shown is that if for each σ the number n of these zeros grows more slowly than $\log T$, then the Lindelöf hypothesis must be true.

The first step is to consider $|\log F(s)|$ on a circle C_1 concentric with C lying entirely in the halfplane $\operatorname{Re} s > 1$, say the circle with center s_0 tangent to the line $\operatorname{Re} s = 1 + \Delta$, where $0 < \Delta < \sigma_0 - 1$. Let $\alpha_\nu(s)$ denote the νth factor in the product in (1) and consider $\log\alpha_\nu(s)$. Since α_ν is a fractional linear transformation which carries s_ν to ∞ and C to the unit circle, it carries C_1 to a circle which lies outside the unit circle and does not encompass it. Since this circle contains in its interior the point $\alpha_\nu(s_0) = R/(s_0 - s_\nu)$ which lies on the same ray from the origin as $\bar{s}_0 - \bar{s}_\nu$ and which therefore lies in the halfplane $\operatorname{Re}\alpha_\nu(s_0) > 0$, $\log\alpha_\nu(s)$ can be defined inside and on C_1 by the condition $|\operatorname{Im}\log\alpha_\nu(s_0)| < \frac{1}{2}\pi$. This gives a meaning to $\log F(s)$, namely, $\log F(s) = \log\zeta(s) + \sum_{\nu=1}^{n}\log\alpha_\nu(s)$, throughout the interior of C_1 and hence, by analytic continuation, throughout the interior of C where F is analytic and nonzero. Now $|\operatorname{Im}\log\alpha_\nu(s)| < 3\pi/2$ on C_1 because a circle which does not contain the origin cannot intersect both halves of the imaginary axis. On the other hand, $\operatorname{Re}\log\alpha_\nu(s) = \log|\alpha_\nu(s)|$ is positive on C_1 but less than $\log[(R^2 + RR_1)/R\Delta]$, where R_1 is the radius of C_1. Thus there is a bound $b \geq |\log\alpha_\nu(s)|$ for s on

†Note that σ_0 must be greater than one.

C_1 which is valid for all ν and all T. Since $|\log F(s)| \leq |\log \zeta(s)| + nb$ and since $|\log \zeta(s)|$ is bounded on C_1 independently of T [namely, by $\log \zeta(1 + \Delta)$], this shows that *the given assumption on n implies that for every* $\epsilon > 0$ *there is a* T_0 *such that* $|\log F(s)| < \epsilon \log T$ *on* C_1 *whenever* $T > T_0$.

Consider next $\log |F(s)| = \operatorname{Re} \log F(s) = \log |\zeta(s)| + \sum_{\nu=1}^{n} \log |\alpha_\nu(s)|$ on C. Since α_ν is chosen in such a way that $|\alpha_\nu(s)| \equiv 1$ on C (see Section 2.2), this is $\log |\zeta(s)|$ which, since $|\zeta(s)| \leq \text{const } t^2$, is less than a constant times $\log T$ as $T \longrightarrow \infty$. On the other hand, $\log |F(s_0)|$ is bounded below because $\log |\zeta(s_0)| \geq -\log \zeta(\sigma_0)$ and $\log |\alpha_\nu(s_0)| = \log (R/|s_\nu - s_0|) \geq 0$; hence $\operatorname{Re} \log [F(s)/F(s_0)]$ is zero at the center of C and less than a constant times $\log T$ on C. Now by the lemma of Section 2.7 this implies that the *modulus* of $\log [F(s)/F(s_0)]$ is less than a constant times $\log T$ on a smaller circle. Specifically this lemma shows that on the circle C_3 concentric with C but slightly smaller, say with radius $R_3 = R - \eta$, where $\eta > 0$ is small, the modulus of $\log [F(s)/F(s_0)]$ is at most $2R_3(R - R_3)^{-1}$ times the maximum of $\operatorname{Re} \log [F(s)/F(s_0)]$ on C. Thus there is a constant K such that $|\log F(s)| < K \log T$ on C_3 for all sufficiently large T.

Finally let C_2 be a circle concentric with C_3 but slightly smaller still, say with radius $R_2 = R_3 - \eta = R - 2\eta$, and consider the modulus of $\log F(s)$ on C_2. By the three circles theorem this modulus is at most $(\epsilon \log T)^\alpha (K \log T)^\beta = \epsilon^\alpha K^\beta \log T$, where α and β are positive numbers independent of T which satisfy $\alpha + \beta = 1$. Since ϵ is arbitrarily small, so is $\epsilon^\alpha K^\beta$, and it follows that for any given $\delta > 0$ there is a T_0 such that $|\log F(s)| < \delta \log T$ inside and on C_2 whenever $T \geq T_0$. Since $\log |F(s)| \leq |\log F(s)|$, this gives $|F(s)| < T^\delta$ and consequently, since $|\alpha_\nu(s)| > 1$ on C_2, $|\zeta(s)| < T^\delta$ inside and on C_2; hence $|\zeta(s)| < T^\delta$ throughout the strip $\sigma_0 - R_2 \leq \operatorname{Re} s \leq \sigma_0 + R_2$ once T is sufficiently large. Since $\sigma_0 - R_2 = \sigma + 2\eta$ is arbitrarily near $\frac{1}{2}$ and since δ is arbitrarily small, the Lindelöf hypothesis follows.

9.5 THE AVERAGE VALUE OF $S(t)$ IS ZERO

This section is devoted to the proof of Littlewood's theorem, mentioned in Section 8.2, that $\int_{t_1}^{t_2} S(t)\, dt$ grows no more rapidly than $\log t_2$ as $t_2 \longrightarrow \infty$. The essential step is to show that $\int_{t_1}^{t_2} S(t)\, dt$ can be rewritten in the form

$$\int_{t_1}^{t_2} S(t)\, dt = \int_{1/2}^{\infty} \pi^{-1} \log |\zeta(\sigma + it_2)|\, d\sigma - \int_{1/2}^{\infty} \pi^{-1} \log |\zeta(\sigma + it_1)|\, d\sigma. \tag{1}$$

To prove this consider the rectangle $D = \{s: \frac{1}{2} \leq \operatorname{Re} s \leq K, t_1 \leq \operatorname{Im} s \leq t_2\}$, where K is a large constant. The function $\log \zeta(s)$ is well defined on the portion

of D which lies to the right of $\operatorname{Re} s = 1$ and the function $S(t)$ is $\pi^{-1} \log \zeta(s)$ on the line $\operatorname{Re} s = \frac{1}{2}$ when $\log \zeta(s)$ is analytically continued along lines $\operatorname{Im} s = \text{const}$ which contain no zeros of $\zeta(s)$. Let D_ϵ denote the domain obtained by deleting from D all points whose imaginary parts lie within ϵ of the imaginary part of a zero of $\zeta(s)$. Then D_ϵ is a union of a finite number of rectangles and it contains no zeros of $\zeta(s)$, so $\pi^{-1} \log \zeta(s)$ is well defined throughout D_ϵ by analytic continuation and its imaginary part at points of ∂D_ϵ of the form $\frac{1}{2} + it$ is $S(t)$. Now by Cauchy's theorem

$$\int_{\partial D_\epsilon} \pi^{-1} \log \zeta(s)\, ds = 0.$$

Taking the real part of this equation gives

$$\int_{\partial D_\epsilon} \pi^{-1} \operatorname{Re} \log \zeta(s)\, d\sigma = \int_{\partial D_\epsilon} \pi^{-1} \operatorname{Im} \log \zeta(s)\, dt.$$

The integral on the left involves only the horizontal boundaries of D_ϵ. These consist of the top boundary $\operatorname{Im} s = t_2$, the bottom boundary $\operatorname{Im} s = t_1$, and the interior boundaries $\operatorname{Im} s = t \pm \epsilon$, where t is the imaginary part of a zero of $\zeta(s)$. Now the integral over a pair of interior boundaries can be written

$$\begin{aligned} &\int_{1/2}^{K} \pi^{-1} \operatorname{Re} \log \zeta(\sigma + it + i\epsilon)\, d\sigma \\ &\qquad - \int_{1/2}^{K} \pi^{-1} \operatorname{Re} \log \zeta(\sigma + it - i\epsilon)\, d\sigma \\ &\qquad = \int_{1/2}^{K} \pi^{-1} \log \left| \frac{\zeta(\sigma + it + i\epsilon)}{\zeta(\sigma + it - i\epsilon)} \right| d\sigma. \end{aligned}$$

Since $\log |\zeta(\sigma + it + i\epsilon)/\zeta(\sigma + it - i\epsilon)|$ approaches $\log 1 = 0$ uniformly as $\epsilon \downarrow 0$, the integrals over the interior boundaries cancel as $\epsilon \downarrow 0$. The integral on the right involves only the vertical boundaries of D_ϵ. These consist of the two line segments $\{\frac{1}{2} + it : t_1 \le t \le t_2\}$ and $\{K + it : t_1 \le t \le t_2\}$ with intervals of length 2ϵ deleted. Since deleting intervals of length 2ϵ from the domain of a convergent† integral and letting $\bullet \downarrow 0$ does not change the value of the integral, and since $S(t)$ is Riemann integrable on $\{t_1 \le t \le t_2\}$ (it is continuous except for a finite number of jump discontinuities), the integral on the right approaches $-\int_{t_1}^{t_2} S(t)\, dt + \int_{t_1}^{t_2} \pi^{-1} \operatorname{Im} \log [\zeta(K + it)]\, dt$ as $\epsilon \downarrow 0$ and the equation

$$\begin{aligned} &\int_{1/2}^{K} \pi^{-1} \log |\zeta(\sigma + it_1)|\, d\sigma - \int_{1/2}^{K} \pi^{-1} \log |\zeta(\sigma + it_2)|\, d\sigma \\ &\qquad = -\int_{t_1}^{t_2} S(t)\, dt + \int_{t_1}^{t_2} \pi^{-1} \operatorname{Im} \log \zeta(K + it)\, dt \end{aligned}$$

†$\operatorname{Im} \log \zeta$ is continuous on ∂D except for a finite number of jump discontinuities; hence it is Riemann integrable.

results. Thus to prove (1) it remains only to show that $\int_{t_1}^{t_2} \operatorname{Im} \log \zeta(K + it)\, dt$ approaches zero and $\int_{1/2}^{K} \operatorname{Re} \log \zeta(\sigma + it)\, d\sigma$ converges as $K \to \infty$. Since both the real and the imaginary part of a function are less in absolute value than its modulus, both of these statements follow from estimates of $|\log \zeta(K + it)| \leq \log \zeta(K)$ for large K. If, for given K, u is defined to be $u = 2^{-K} + 3^{-K} + 4^{-K} + \cdots$, then $0 < u \leq 2^{-K} + \int_2^\infty t^{-K}\, dt \leq 3 \cdot 2^{-K}$ for $K \geq 2$, and $\log \zeta(K) = \log(1 + u) < u \leq 3 \cdot 2^{-K}$, from which the desired conclusions follow immediately.

Thus Littlewood's estimate of $\int_{t_1}^{t_2} S(t)\, dt$ is equivalent to the statement that $\left| \int_{1/2}^{\infty} \log |\zeta(\sigma + it)|\, d\sigma \right|$ grows no faster than $\log t$ as $t \to \infty$. Since the estimate above gives $\left| \int_2^\infty \log |\zeta(\sigma + it)|\, d\sigma \right| \leq \int_2^\infty 3 \cdot 2^{-\sigma}\, d\sigma$, which is independent of t, it will suffice to show that $\left| \int_{1/2}^{2} \log |\zeta(\sigma + it)|\, d\sigma \right|$ grows no faster than $\log t$. The main step in the proof of this fact is an estimate of $|\log \zeta(s)|$ similar to Backlund's estimate in the preceding section.

Let C be the circle of radius $R = 2 + \delta$ ($\delta > 0$) with center at $s_0 = 2 + it$ (t large and variable), and let

$$F(s) = \zeta(s) \prod_{\nu=1}^{n} \frac{R^2 - (\bar{s}_\nu - \bar{s}_0)(s - s_0)}{R(s - s_\nu)} \tag{2}$$

where $s_1, s_2, \ldots, s_n$ are the zeros of ζ inside C. As before it will be assumed that there are no zeros on C, a condition which excludes only a discrete set of values of t and does not affect the validity of the conclusion of the argument. Then, as in the preceding section, $\log F(s)$ can be defined throughout the disk bounded by C, and at points s of C it satisfies $\operatorname{Re} \log [F(s)/F(s_0)] = \log |\zeta(s)| - \log |\zeta(s_0)| - \sum_{\nu=1}^{n} \log |R/(s_0 - s_\nu)| \leq \log |\zeta(s)| + \log \zeta(2)$ which is less than a constant times $\log t$ (see Section 6.7). Therefore by the lemma of Section 2.7 on the slightly smaller circle C_3 of radius $R_3 = 2$ about s_0, the modulus of $\log [F(s)/F(s_0)]$, and hence of $\log F(s)$, is less than a constant times $\log t$. Now

$$\begin{aligned} \left| \int_{1/2}^{2} \log |\zeta(\sigma + it)|\, d\sigma \right| &= \left| \operatorname{Re} \int_{1/2}^{2} \log \zeta(\sigma + it)\, d\sigma \right| \\ &\leq \int_{1/2}^{2} |\operatorname{Re} \log F(\sigma + it)|\, d\sigma \\ &\quad + \sum_{\nu=1}^{n} \left| \int_{1/2}^{2} \operatorname{Re} \log \alpha_\nu(\sigma + it)\, d\sigma \right|, \end{aligned}$$

where, as before, α_ν denotes the νth factor in definition (2) of $F(s)$. Now $|\operatorname{Re} \log F(\sigma + it)| \leq |\log F(\sigma + it)|$ has been shown to be less than a constant times $\log t$; hence the first term on the right is less than a constant times $\log t$. The second term on the right is a sum of n terms where, by von Mangoldt's theorem on the density of the roots (Section 3.4), n is less than a constant times $\log t$; so to prove the theorem, it will suffice to show that the terms of this sum $\left| \int_{1/2}^{2} \operatorname{Re} \log \alpha_\nu(\sigma + it)\, d\sigma \right|$ have an upper bound independent of

ν and t. Since the numerator of α_ν is bounded away from 0 and ∞ on the domain of integration, the log of its modulus is also; so this amounts essentially to finding a bound for $\int_{1/2}^{2} \log|\sigma + it - s_\nu|\, d\sigma$. Now $4 \geq |\sigma + it - s_\nu| \geq |\sigma - \sigma_\nu|$, where $\sigma_\nu = \operatorname{Re} s_\nu$; so the integral in question is at most $\frac{3}{2}\log 4$ and at least $\int_{1/2}^{2} \log|\sigma - \sigma_\nu|\, d\sigma$. Although the integrand in this last integral is unbounded at $\sigma = \sigma_\nu$, the integral is convergent as an improper integral (or as a Lebesgue integral) and is easily shown to be bounded. This completes the proof that for every $t_1 > 0$ there is a $K > 0$ and a $T_0' > t_1$ such that $\left|\int_{t_1}^{t_2} S(t)\, dt\right| < K \log t_2$ whenever $t_2 \geq T_0$.

9.6 THE BOHR–LANDAU THEOREM

In 1914 Bohr and Landau [B8] proved a different sort of relationship between the growth of ζ and the location of its zeros. Roughly what they proved was that the fact that the *average* value of $|\zeta(s)|^2$ on lines $\operatorname{Re} s = \sigma$ is bounded for $\sigma > \frac{1}{2}$, uniformly for $\sigma \geq \frac{1}{2} + \delta$, implies that *most* of the roots ρ lie in the range $\operatorname{Re} s < \frac{1}{2} + \delta$ for any $\delta > 0$. [Actually Bohr and Landau deduced their conclusions about the roots ρ not from properties of ζ but from properties of the related function $1 - 2^{-s} + 3^{-s} - 4^{-s} + \cdots = (1 + 2^{-s} + 3^{-s} + \cdots) - 2(2^{-s} + 4^{-s} + 6^{-s} + \cdots) = (1 - 2^{1-s})\zeta(s)$, for which the needed facts about averages on lines $\operatorname{Re} s = \sigma$ are easier to prove.] Specifically, it will be shown in the next section that *for every $\sigma_0 > \frac{1}{2}$ there exist K, T_0 such that* $(T-1)^{-1}\int_1^T |\zeta(\sigma + it)|^2\, dt < K$ *whenever* $\sigma \geq \sigma_0$ and $T \geq T_0$. This section is devoted to Bohr and Landau's method of concluding from this that *for every $\delta > 0$ there is a K' such that the number of roots ρ in the range* $\{\operatorname{Re} s \geq \frac{1}{2} + \delta, 0 \leq \operatorname{Im} s \leq T\}$ *is less than $K'T$ for all T.* Since the total number $N(T)$ of roots in the range $\{0 \leq \operatorname{Im} s \leq T\}$ is about $(T/2\pi)\log(T/2\pi)$ (see Section 6.7), this proves that the number of these roots to the right of the line $\operatorname{Re} s = \frac{1}{2} + \delta$ divided by their total number approaches zero as $T \to \infty$. In short, *for any $\delta > 0$ all but an infinitesimal proportion of the roots ρ lie within δ of the line* $\operatorname{Re} s = \frac{1}{2}$. In the sense that this is a statement about all but an infinitesimal proportion of the roots (one is tempted to say "almost all," but to avoid the misinterpretation that this might mean "all but a finite number," Littlewood's phrase "infinitesimal proportion" is better) it is to this day the strongest theorem on the location of the roots which substantiates the Riemann hypothesis.

Bohr and Landau's method of estimating the number of roots ρ to the right of $\operatorname{Re} s = \frac{1}{2} + \delta$ given the above fact about the average of $|\zeta(s)|^2$ on $\operatorname{Re} s = \sigma > \frac{1}{2}$ is as follows. Let $\delta > 0$ be given, let C be a circle through $\frac{1}{2} + \delta + it$ and $\frac{1}{2} + \delta + i(t+1)$ which lies in the halfplane $\operatorname{Re} s < \frac{1}{2}$, and

let C^+ be a circle concentric with C, slightly larger than C but still contained in the halfplane $\operatorname{Re} s > \frac{1}{2}$. As in the proofs of this type given above, the choice of C, C^+, which is different for different values of t, is to differ merely by a vertical translation. Let r denote the radius of C, R the radius of C^+, $s_0 = \sigma_0 + i(t + \frac{1}{2})$ their common center, and n the number of zeros of ζ in the rectangle $\{\frac{1}{2} + \delta \leq \operatorname{Re} s \leq 1, t \leq \operatorname{Im} s \leq t + 1\}$. Since this rectangle is contained in C, Jensen's theorem gives

$$\log|\zeta(s_0)| + n \log \frac{R}{r} \leq \frac{1}{2\pi} \int_0^{2\pi} \log|\zeta(s_0 + \operatorname{Re}^{i\theta})|\, d\theta,$$

$$2 \log|\zeta(s_0)| + 2n \log \frac{R}{r} \leq \frac{1}{2\pi} \int_0^{2\pi} \log|\zeta(s_0 + \operatorname{Re}^{i\theta})|^2\, d\theta.$$

Without loss of generality it can be assumed that $1 - 2^{-\sigma_0} - 3^{-\sigma_0} - \cdots > 0$ (increase σ_0 if necessary) so that there is an A with $|\zeta(s_0)| \geq A > 0$ for all t. The fact that the geometric mean of a function is less than or equal to its arithmetic mean gives

$$\frac{1}{2\pi} \int_0^{2\pi} \log|\zeta(s_0 + \operatorname{Re}^{i\theta})|^2\, d\theta \leq \log\left[\frac{1}{2\pi} \int_0^{2\pi} |\zeta(s_0 + \operatorname{Re}^{i\theta})|^2\, d\theta\right].$$

Hence

$$A^2 \left(\frac{R}{r}\right)^{2n} \leq \frac{1}{2\pi} \int_0^{2\pi} |\zeta(s_0 + \operatorname{Re}^{i\theta})|^2\, d\theta.$$

Moreover, the analogous inequality holds when R is replaced by any radius ρ between $(r + R)/2$ and R. Multiplying this inequality by $\rho\, d\rho$ and integrating then gives

$$A^2 r^{-2n} \int_{(r+R)/2}^{R} \rho^{2n+1}\, d\rho \leq \frac{1}{2\pi} \iint |\zeta(s_0 + \rho e^{i\theta})|^2\, \rho\, d\rho\, d\theta.$$

The right side is a constant times the integral of $|\zeta(s)|^2$ over the annulus $\{(r + R)/2 \leq |s - s_0| \leq R\}$ which is less than the integral of $|\zeta(s)|^2$ over the entire disk bounded by C^+. The left side is greater than

$$A^2 r^{-2n} \left(\frac{r + R}{2}\right)^{2n+1} \left(\frac{R - r}{2}\right) = A^2 \left(\frac{R^2 - r^2}{4}\right) \left(1 + \frac{R - r}{2r}\right)^{2n}$$

$$> A^2 \left(\frac{R^2 - r^2}{4}\right) \cdot 2n \left(\frac{R - r}{2r}\right),$$

that is, greater than a constant times n. Hence n is less than a constant times the integral of $|\zeta(s)|^2$ over the disk bounded by C^+, the constant depending on A, r, R and hence on δ but not on t. Ignoring the range of t for which C^+ includes the singularity $s = 1$ of ζ and adding the inequality just obtained over all integer values of t above this range and below $T + 1$ shows (since the overlapping C^+'s include any given point at most $2R$ times) that the total number of zeros with imaginary parts less than a given integer is less than a constant plus a constant times an integral of $|\zeta(s)|^2$ over a strip of the form

$\{\sigma_0 - R \leq \operatorname{Re} s \leq \sigma_0 + R, 1 \leq \operatorname{Im} s \leq T + R\}$ which, by the property of the averages of $|\zeta(s)|^2$ to be proved in the next section, is less than a constant times T, as was to be shown.

Of course the actual constant obtained by this method of estimation is absurdly large, particularly in view of the fact that according to the Riemann hypothesis it should be zero. Another technique for proving the same estimate with a much smaller constant was given by Littlewood in 1924 [L14].

9.7 THE AVERAGE OF $|\zeta(s)|^2$

For $\sigma > 1$ it is easy to prove that the average $(T-1)^{-1}\int_1^T |\zeta(\sigma + it)|^2\,dt$ of $|\zeta(s)|^2$ on $\operatorname{Re} s = \sigma$ approaches $\zeta(2\sigma)$ as $T \to \infty$ and that the approach to the limit is uniform for $\sigma \geq \sigma_0 \geq 1$. First of all, since $\sigma \neq 1$ and since $|\zeta(\sigma + it)| = |\zeta(\sigma - it)|$, the average can be written in the more natural form $\lim_{T\to\infty} (1/2T)\int_{-T}^{T} |\zeta(\sigma + it)|^2\,dt$. Then using

$$|\zeta(\sigma + it)|^2 = \zeta(\sigma + it)\zeta(\sigma - it) = \sum_m \sum_n \frac{1}{n^{\sigma+it}} \cdot \frac{1}{m^{\sigma+it}}$$

and noting that this double series is uniformly convergent (it is dominated by $\sum\sum n^{-\sigma}m^{-\sigma}$), so it can be integrated termwise, gives

$$\lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} |\zeta(\sigma + it)|^2\,dt = \lim_{T\to\infty} \sum_m \sum_n \frac{1}{n^\sigma} \cdot \frac{1}{m^\sigma} \frac{1}{2T}\int_{-T}^{T} \left(\frac{m}{n}\right)^{it} dt.$$

If $m = n$, the coefficient of $n^{-\sigma}m^{-\sigma} = n^{-2\sigma}$ is identically 1, whereas if $n \neq m$, it is $2 \sin [T \log (n/m)]/2T \log (n/m)$. Since $(\sin h)/h$ is bounded, the limit of the sums is the sum of the limits which is just $\sum n^{-2\sigma} = \zeta(2\sigma)$ and this is true uniformly in σ, provided σ is bounded away from 1.

Since $\zeta(2\sigma)$ makes sense all the way to $\sigma = \frac{1}{2}$, it is natural to ask whether it is not still true that the average of $|\zeta(s)|^2$ on $\operatorname{Re} s = \sigma$ is $\zeta(2\sigma)$ for $\sigma > \frac{1}{2}$. The method of proof will of course have to be drastically modified because $\sum n^{-\sigma}$ no longer converges when $\sigma \leq 1$, but the theorem is still true. This theorem appears in Landau's 1908 *Handbuch* [L3], but the central idea of the proof which follows is from Hardy and Littlewood [H6].

The proof for $\sigma > 1$ and the form of the theorem strongly suggest that the divergent series $\sum n^{-\sigma-it}$ will play a role in the proof. If, as in the estimation of $\zeta(1 + it)$ in Section 9.2, one uses just the terms of this series in which $n < t$, then the remainder $R(\sigma, t)$ in $\zeta(\sigma + it) = \sum_{n<t} n^{-\sigma-it} + R(\sigma, t)$ must be estimated. Now the Euler–Maclaurin formula for $\zeta(\sigma + it)$ gives

$$\begin{aligned} R(\sigma, t) &= \zeta(\sigma + it) - \sum_{n<t} n^{-\sigma-it} \\ &= \sum_{t\leq n<N} n^{-s} + \frac{N^{1-s}}{s-1} + \frac{1}{2}N^{-s} - s\int_N^\infty \bar{B}_1(x)x^{-s-1}\,dx, \end{aligned}$$