1. WILLIAM ADAMS & DANIEL SHANKS, "Strong primality tests that are not sufficient," *Math. Comp.*, v. 39, 1982, pp. 255–300.

2. WILLIAM ADAMS & DANIEL SHANKS, "Strong primality tests. II—Algebraic identification of the p-adic limits and their implications." (To appear.)

3. H. BEHNKE & F. SOMMER, *Theorie der analytischen Funktionen einer complexen Veränderlichen*, Springer, Berlin, 1965, viii + 603 pp.

4. MARVIN I. KNOPP, *Modular Functions in Analytic Number Theory*, Markham, Chicago, Ill., 1970, x + 150 pp.

5. DERRICK H. LEHMER & EMMA LEHMER, "Cyclotomy with short periods," *Math. Comp.*, v. 41, 1983, pp. 743–758.

6. DANIEL SHANKS, "Dihedral quartic approximations and series for π," *J. Number Theory*, v. 14, 1982, pp. 397–423.

7. DANIEL SHANKS, "Review of A. O. L. Atkin's table," *Math. Comp.*, v. 32, 1978, p. 315.

8. THOMAS R. PARKIN & DANIEL SHANKS, "On the distribution of parity in the partition function," *Math. Comp.*, v. 21, 1967, pp. 446–480.

9. DANIEL SHANKS & LARRY P. SCHMID, "Variations on a theorem of Landau," *Math. Comp.*, v. 20, 1966, pp. 551–569.

10. DANIEL SHANKS, "Review 112", *Math. Comp.*, v. 19, 1965, pp. 684–686.

L'Enseignement Mathématique, t. 30 (1984), p. 275-330

THE ARITHMETIC-GEOMETRIC MEAN OF GAUSS

by David A. Cox

INTRODUCTION

The arithmetic-geometric mean of two numbers a and b is defined to be the common limit of the two sequences $\{a_n\}_{n=0}^{\infty}$ and $\{b_n\}_{n=0}^{\infty}$ determined by the algorithm

$$a_0 = a\,, \qquad b_0 = b\,,$$

(0.1)

$$a_{n+1} = (a_n+b_n)/2\,, \qquad b_{n+1} = (a_nb_n)^{1/2}\,, \quad n = 0, 1, 2, \ldots\,.$$

Note that a_1 and b_1 are the respective arithmetic and geometric means of a and b, a_2 and b_2 the corresponding means of a_1 and b_1, etc. Thus the limit

$$M(a, b) = \lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n \tag{0.2}$$

really does deserve to be called the arithmetic-geometric mean of a and b. This algorithm first appeared in a paper of Lagrange, but it was Gauss who really discovered the amazing depth of this subject. Unfortunately, Gauss published little on the agM (his abbreviation for the arithmetic-geometric mean) during his lifetime. It was only with the publication of his collected works [12] between 1868 and 1927 that the full extent of his work became apparent. Immediately after the last volume appeared, several papers (see [15] and [35]) were written to bring this material to a wider mathematical audience. Since then, little has been done, and only the more elementary properties of the agM are widely known today.

In § 1 we review these elementary properties, where a and b are positive real numbers and the square root in (0.1) is also positive. The convergence of the algorithm is easy to see, though less obvious is the connection between the agM and certain elliptic integrals. As an application, we use $M(\sqrt{2}, 1)$ to determine the arc length of the lemniscate. In § 2, we allow a and b to be complex numbers, and the level of difficulty changes dramatically.

The convergence of the algorithm is no longer obvious, and as might be expected, the square root in (0.1) causes trouble. In fact, $M(a, b)$ becomes a multiple valued function, and in order to determine the relation between the various values, we will need to "uniformize" the agM using quotients of the classical Jacobian theta functions, which are modular functions for certain congruence subgroups of level four in $SL(2, \mathbf{Z})$. The amazing fact is that Gauss knew all of this! Hence in § 3 we explore some of the history of these ideas. The topics encountered will range from Bernoulli's study of elastic rods (the origin of the lemniscate) to Gauss' famous mathematical diary and his work on secular perturbations (the only article on the agM published in his lifetime).

I would like to thank my colleagues David Armacost and Robert Breusch for providing translations of numerous passages originally in Latin or German. Thanks also go to Don O'Shea for suggesting the wonderfully quick proof of (2.2) given in § 2.

1. The arithmetic-geometric mean of real numbers

When a and b are positive real numbers, the properties of the agM $M(a, b)$ are well known (see, for example, [5] and [26]). We will still give complete proofs of these properties so that the reader can fully appreciate the difficulties we encounter in § 2.

We will assume that $a \geqslant b > 0$, and we let $\{a_n\}_{n=0}^{\infty}$ and $\{b_n\}_{n=0}^{\infty}$ be as in (0.1), where b_{n+1} is always the positive square root of $a_n b_n$. The usual inequality between the arithmetic and geometric means,

$$(a+b)/2 \geqslant (ab)^{1/2},$$

immediately implies that $a_n \geqslant b_n$ for all $n \geqslant 0$. Actually, much more is true: we have

$$a \geqslant a_1 \geqslant \ldots \geqslant a_n \geqslant a_{n+1} \geqslant \ldots \geqslant b_{n+1} \geqslant b_n \geqslant \ldots \geqslant b_1 \geqslant b \tag{1.1}$$

$$0 \leqslant a_n - b_n \leqslant 2^{-n}(a-b). \tag{1.2}$$

To prove (1.1), note that $a_n \geqslant b_n$ and $a_{n+1} \geqslant b_{n+1}$ imply

$$a_n \geqslant (a_n+b_n)/2 = a_{n+1} \geqslant b_{n+1} = (a_n b_n)^{1/2} \geqslant b_n,$$

and (1.1) follows. From $b_{n+1} \geqslant b_n$ we obtain

$$a_{n+1} - b_{n+1} \leqslant a_{n+1} - b_n = 2^{-1}(a_n - b_n),$$

and (1.2) follows by induction. From (1.1) we see immediately that $\lim_{n\to\infty} a_n$ and $\lim_{n\to\infty} b_n$ exist, and (1.2) implies that the limits are equal. Thus, we can use (0.2) to define the arithmetic-geometric mean $M(a, b)$ of a and b.

Let us work out two examples.

Example 1. $M(a, a) = a$.

This is obvious because $a = b$ implies $a_n = b_n = a$ for all $n \geqslant 0$.

Example 2. $M(\sqrt{2}, 1) = 1.1981402347355922074...$

The accuracy is to 19 decimal places. To compute this, we use the fact that $a_n \geqslant M(a, b) \geqslant b_n$ for all $n \geqslant 0$ and the following table (all entries are rounded off to 21 decimal places).

n	a_n	b_n
0	1.414213562373905048802	1.000000000000000000000
1	1.207106781186547524401	1.189207115002721066717
2	1.198156948094634295559	1.198123521493120122607
3	1.198140234793877209083	1.198140234677307205798
4	1.198140234735592207441	1.198140234735592207439

Such computations are not too difficult these days, though some extra programming was required since we went beyond the usual 16 digits of double-precision. The surprising fact is that these calculations were done not by computer but rather by Gauss himself. The above table is one of four examples given in the manuscript "De origine proprietatibusque generalibus numerorum mediorum arithmetico-geometricorum" which Gauss wrote in 1800 (see [12, III, pp. 361-371]). As we shall see later, this is an especially important example.

Let us note two obvious properties of the agM:

$$M(a, b) = M(a_1, b_1) = M(a_2, b_2) = \ldots \tag{1.3}$$

$$M(\lambda a, \lambda b) = \lambda M(a, b).$$

Both of these follow easily from the definition of $M(a, b)$.

Our next result shows that the agM is not as simple as indicated by what we have done so far. We now get our first glimpse of the depth of this subject.

THEOREM 1.1. *If* $a \geqslant b > 0$, *then*

$$M(a, b) \cdot \int_0^{\pi/2} (a^2\cos^2\phi + b^2\sin^2\phi)^{-1/2} d\phi = \pi/2 .$$

Proof. Let $I(a, b)$ denote the above integral, and set $\mu = M(a, b)$. Thus we need to prove $I(a, b) = (\pi/2)\mu^{-1}$. The key step is to show that

$$I(a, b) = I(a_1, b_1) . \tag{1.4}$$

The shortest proof of (1.4) is due to Gauss. He introduces a new variable ϕ' such that

$$\sin\phi = \frac{2a \sin\phi'}{a + b + (a-b)\sin^2\phi'} . \tag{1.5}$$

Note that $0 \leqslant \phi' \leqslant \pi/2$ corresponds to $0 \leqslant \phi \leqslant \pi/2$. Gauss then asserts "after the development has been made correctly, it will be seen" that

$$(a^2\cos^2\phi + b^2\sin^2\phi)^{-1/2} d\phi = (a_1^2\cos^2\phi' + b_1^2\sin^2\phi')^{-1/2} d\phi' \tag{1.6}$$

(see [12, III, p. 352]). Given this, (1.4) follows easily. In "Fundamenta nova theoriae functionum ellipticorum," Jacobi fills in some of the details Gauss left out (see [20, I, p. 152]). Specifically, one first proves that

$$\cos\phi = \frac{2\cos\phi'(a_1^2\cos^2\phi' + b_1^2\sin^2\phi')^{1/2}}{a + b + (a-b)\sin^2\phi'}$$

$$(a^2\cos^2\phi + b^2\sin^2\phi)^{1/2} = a\frac{a + b - (a-b)\sin^2\phi'}{a + b + (a-b)\sin^2\phi'}$$

(these are straightforward manipulations), and then (1.6) follows from these formulas by taking the differential of (1.5).

Iterating (1.4) gives us

$$I(a, b) = I(a_1, b_1) = I(a_2, b_2) = \dots ,$$

so that $I(a, b) = \lim_{n\to\infty} I(a_n, b_n) = \pi/2\mu$ since the functions

$$(a_n^2\cos^2\phi + b_n^2\sin^2\phi)^{-1/2}$$

converge uniformly to the constant function μ^{-1}. QED

This theorem relates very nicely to the classical theory of complete elliptic integrals of the first kind, i.e., integrals of the form

$$F(k, \pi/2) = \int_0^{\pi/2} (1-k^2\sin^2\phi)^{-1/2} d\phi = \int_0^1 \left((1-z^2)(1-k^2z^2)\right)^{-1/2} dz .$$

To see this, we set $k = \dfrac{a-b}{a+b}$. Then one easily obtains

$$I(a, b) = a^{-1} F\left(\frac{2\sqrt{k}}{1+k}, \pi/2\right), \quad I(a_1, b_1) = a_1^{-1} F(k, \pi/2) ,$$

so that (1.4) is equivalent to the well-known formula

$$F\left(\frac{2\sqrt{k}}{1+k}, \pi/2\right) = (1+k) F(k, \pi/2)$$

(see [16, p. 250] or [17, p. 908]). Also, the substitution (1.5) can be written as

$$\sin\phi = \frac{(1+k)\sin\phi'}{1 + k\sin^2\phi'} ,$$

which is now called the Gauss transformation (see [32, p. 206]).

For someone well versed in these formulas, the derivation of (1.4) would not be difficult. In fact, a problem on the 1895 Mathematical Tripos was to prove (1.4), and the same problem appears as an exercise in Whittaker and Watson's *Modern Analysis* (see [36, p. 533]), though the agM is not mentioned. Some books on complex analysis do define $M(a, b)$ and state Theorem 1.1 (see, for example, [7, p. 417]).

There are several other ways to express Theorem 1.1. For example, if $0 \leqslant k < 1$, then one can restate the theorem as

$$\frac{1}{M(1+k, 1-k)} = \frac{2}{\pi}\int_0^{\pi/2} (1-k^2\sin^2\gamma)^{-1/2} d\gamma = \frac{2}{\pi} F(k, \pi/2) . \tag{1.7}$$

Furthermore, using the well-known power series expansion for $F(k, \pi/2)$ (see [16, p. 905]), we obtain

$$\frac{1}{M(1+k, 1-k)} = \sum_{n=0}^{\infty} \left[\frac{1\cdot 3\cdot \dots \cdot (2n-1)}{2^n n!}\right]^2 k^{2n} . \tag{1.8}$$

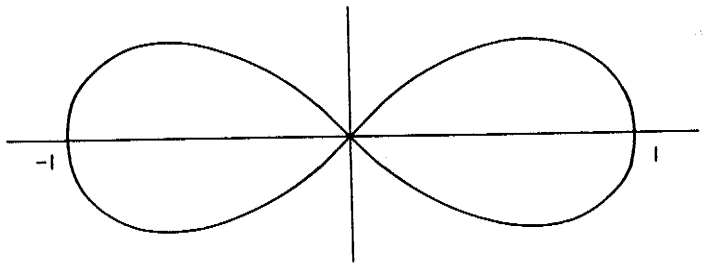
Finally, it is customary to set $k' = \sqrt{1-k^2}$. Then, using (1.3), we can rewrite (1.7) as

$$\frac{1}{M(1, k')} = \frac{2}{\pi}\int_0^{\pi/2} (1-k^2\sin^2\gamma)^{-1/2} d\gamma . \tag{1.9}$$

This last equation shows that the average value of the function $(1-k^2\sin^2\gamma)^{-1/2}$ on the interval $[0, \pi/2]$ is the reciprocal of the agM of the reciprocals of the minimum and maximum values of the function, a lovely interpretation due to Gauss — see [12, III, p. 371].

One application of Theorem 1.1, in the guise of (1.7), is that the algorithm for the agM now provides a very efficient method for approximating the elliptic integral $F(k, \pi/2)$. As we will see in § 3, it was just this problem that led Lagrange to independently discover the algorithm for the agM.

Another application of Theorem 1.1 concerns the arc length of the lemniscate $r^2 = \cos 2\theta$:



Using the formula for arc length in polar coordinates, we see that the total arc length is

$$4\int_0^{\pi/4}(r^2+(dr/d\theta)^2)^{1/2}\,d\theta = 4\int_0^{\pi/4}(\cos 2\theta)^{-1/2}\,d\theta\,.$$

The substitution $\cos 2\theta = \cos^2\phi$ transforms this to the integral

$$4\int_0^{\pi/2}(1+\cos^2\phi)^{-1/2}\,d\phi = 4\int_0^{\pi/2}(2\cos^2\phi+\sin^2\phi)^{-1/2}\,d\phi\,.$$

Using Theorem 1.1 to interpret this last integral in terms of $M(\sqrt{2}, 1)$, we see that the arc length of the lemniscate $r^2 = \cos 2\theta$ is $2\pi/M(\sqrt{2}, 1)$.

From Example 2 it follows that the arc length is approximately 5.244, and much better approximations can be easily obtained. (For more on the computation of the arc length of the lemniscate, the reader should consult [33].)

On the surface, this arc length computation seems rather harmless. However, from an historical point of view, it is of fundamental importance. If we set $z = \cos\phi$, then we obtain

$$\int_0^{\pi/2}(2\cos^2\phi+\sin^2\phi)^{-1/2}\,d\phi = \int_0^1(1-z^4)^{-1/2}\,dz\,.$$

The integral on the right appeared in 1691 in a paper of Jacob Bernoulli and was well known throughout the 18th century. Gauss even had a special notation for this integral, writing

$$\varpi = 2\int_0^1(1-z^4)^{-1/2}\,dz\,.$$

Then the relation between the arc length of the lemniscate and $M(\sqrt{2}, 1)$ can be written

$$M(\sqrt{2}, 1) = \frac{\pi}{\varpi}\,.$$

To see the significance of this equation, we turn to Gauss' mathematical diary. The 98th entry, dated May 30, 1799, reads as follows:

> We have established that the arithmetic-geometric mean between 1 and $\sqrt{2}$ is π/ϖ to the eleventh decimal place; the demonstration of this fact will surely open an entirely new field of analysis.

(See [12, X.1, p. 542].) The genesis of this entire subject lies in Gauss' observation that these two numbers are the same. It was in trying to understand the real meaning of this equality that several streams of Gauss' thought came together and produced the exceptionally rich mathematics which we will explore in § 2.

Let us first examine how Gauss actually showed that $M(\sqrt{2}, 1) = \pi/\varpi$. The proof of Theorem 1.1 given above appeared in 1818 in a paper on secular perturbations (see [12, III, pp. 331-355]), which is the only article on the agM Gauss published in his lifetime (though as we've seen, Jacobi knew this paper well). It is more difficult to tell precisely when he first proved Theorem 1.1, although his notes do reveal that he had two proofs by December 23, 1799.

Both proofs derive the power series version (1.8) of Theorem 1.1. Thus the goal is to show that $M(1+k, 1-k)^{-1}$ equals the function

$$y = \sum_{n=0}^{\infty}\left(\frac{1\cdot 3\cdot\ldots\cdot(2n-1)}{2^n n!}\right)^2 k^{2n}\,. \tag{1.10}$$

The first proof, very much in the spirit of Euler, proceeds as follows. Using (1.3), Gauss derives the identity

$$M\left(1+\frac{2t}{1+t^2},\ 1-\frac{2t}{1+t^2}\right) = \frac{1}{1+t^2}\,M(1+t^2, 1-t^2)\,. \tag{1.11}$$

He then assumes that there is a power series expansion of the form

$$\frac{1}{M(1+k,\,1-k)} = 1 + A\,k^2 + B\,k^4 + C\,k^6 + \ldots\,.$$

By letting $k = t^2$ and $2t/(1+t^2)$ in this series and using (1.11), Gauss obtains

$$1 + A\left(\frac{2t}{1+t^2}\right)^2 + B\left(\frac{2t}{1+t^2}\right)^4 + C\left(\frac{2t}{1+t^2}\right)^6 + \ldots$$
$$= (1+t^2)\,(1+At^4+Bt^8+Ct^{12}+\ldots)\,.$$

Multiplying by $2t/(1+t^2)$, this becomes

$$\frac{2t}{1+t^2} + A\left(\frac{2t}{1+t^2}\right)^3 + B\left(\frac{2t}{1+t^2}\right)^5 + \ldots = 2t(1+At^4+Bt^8+\ldots)\,.$$

A comparison of the coefficients of powers of t gives an infinite system of equations in $A, B, C, \ldots$. Gauss showed that this system is equivalent to the equations $0 = 1 - 4A = 9A - 16B = 25B - 36C = \ldots$, and (1.8) follows easily (see [12, III, pp. 367-369] for details). Gauss' second proof also uses the identity (1.11), but in a different way. Here, he first shows that the series y of (1.10) is a solution of the hypergeometric differential equation

$$(k^3-k)y'' + (3k^2-1)y' + ky = 0\,. \tag{1.12}$$

This enables him to show that y satisfies the identity

$$y\left(\frac{2t}{1+t^2}\right) = (1+t^2)y(t^2)\,,$$

so that by (1.11), $F(k) = M(1+k, 1-k)y(k)$ has the property that

$$F\left(\frac{2t}{1+t^2}\right) = F(t^2)\,.$$

Gauss then asserts that $F(k)$ is clearly constant. Since $F(0) = 1$, we obtain a second proof of (1.8) (see [12, X.1, pp. 181-183]). It is interesting to note that neither proof is rigorous from the modern point of view: the first assumes without proof that $M(1+k, 1-k)^{-1}$ has a power series expansion, and the second assumes without proof that $M(1+k, 1-k)$ is continuous (this is needed in order to show that $F(k)$ is constant).

We can be certain that Gauss knew both of these proofs by December 23, 1799. The evidence for this is the 102nd entry in Gauss' mathematical

diary. Dated as above, it states that "the arithmetic-geometric mean is itself an integral quantity" (see [12, X.1, p. 544]). However, this statement is not so easy to interpret. If we turn to Gauss' unpublished manuscript of 1800 (where we got the example $M(\sqrt{2}, 1)$), we find (1.7) and (1.8) as expected, but also the observation that a complete solution of the differential equation (1.12) is given by

$$\frac{A}{M(1+k,\,1-k)} + \frac{B}{M(1,\,k)}\,, \qquad A, B \in \mathbf{C} \tag{1.13}$$

(see [12, III, p. 370]). In eighteenth century terminology, this is the "complete integral" of (1.12) and thus may be the "integral quantity" that Gauss was referring to (see [12, X.1, pp. 544-545]). Even if this is so, the second proof must predate December 23, 1799 since it uses the same differential equation.

In § 3 we will study Gauss' early work on the agM in more detail. But one thing should be already clear: none of the three proofs of Theorem 1.1 discussed so far live up to Gauss' May 30, 1799 prediction of "an entirely new field of analysis." In order to see that his claim was justified, we will need to study his work on the agM of complex numbers.

2. The arithmetic-geometric mean of complex numbers

The arithmetic-geometric mean of two complex numbers a and b is not easy to define. The immediate problem is that in our algorithm

$$a_0 = a\,, \qquad b_0 = b\,,$$
$$a_{n+1} = (a_n+b_n)/2\,, \qquad b_{n+1} = (a_nb_n)^{1/2}\,, \quad n = 0, 1, 2, \ldots, \tag{2.1}$$

there is no longer an obvious choice for b_{n+1}. In fact, since we are presented with two choices for b_{n+1} for all $n \geqslant 0$, there are uncountably many sequences $\{a_n\}_{n=0}^{\infty}$ and $\{b_n\}_{n=0}^{\infty}$ for given a and b. Nor is it clear that any of these converge!

We will see below (Proposition 2.1) that in fact all of these sequences converge, but only countably many have a non-zero limit. The limits of these particular sequences then allow us to define $M(a, b)$ as a multiple valued function of a and b. Our main result (Theorem 2.2) gives the relationship between the various values of $M(a, b)$. This theorem was discovered

by Gauss in 1800, and we will follow his proof, which makes extensive use of theta functions and modular functions of level four.

We first restrict ourselves to consider only those a's and b's such that $a \neq 0$, $b \neq 0$ and $a \neq \pm b$. (If $a=0$, $b=0$ or $a=\pm b$, one easily sees that the sequences (2.1) converge to either 0 or a, and hence are not very interesting.) An easy induction argument shows that if a and b satisfy these restrictions, so do a_n and b_n for all $n \geqslant 0$ in (2.1).

We next give a way of distinguishing between the two possible choices for each b_{n+1}.

Definition. Let $a, b \in \mathbf{C}^*$ satisfy $a \neq \pm b$. Then a square root b_1 of ab is called the *right choice* if $|a_1 - b_1| \leqslant |a_1 + b_1|$ and, when $|a_1 - b_1| = |a_1 + b_1|$, we also have $\mathrm{Im}(b_1/a_1) > 0$.

To see that this definition makes sense, suppose that $\mathrm{Im}(b_1/a_1) = 0$. Then $b_1/a_1 = r \in \mathbf{R}$, and thus

$$|a_1 - b_1| = |a_1|\,|1 - r| \neq |a_1|\,|1 + r| = |a_1 + b_1|$$

since $r \neq 0$. Notice also that the right choice is unchanged if we switch a and b, and that if a and b are as in § 1, then the right choice for $(ab)^{1/2}$ is the positive one.

It thus seems natural that we should define the agM using (2.1) with b_{n+1} always the right choice for $(a_n b_n)^{1/2}$. However, this is not the only possibility: one can make some wrong choices for b_{n+1} and still get an interesting answer. For instance, in Gauss' notebooks, we find the following example:

n	a_n	b_n
0	3.0000000	1.0000000
1	2.0000000	−1.7320508
2	.1339746	1.8612098i
3	.0669873 + .9306049i	.3530969 + .3530969i
4	.2100421 + .6418509i	.2836903 + .6208239i
5	.2468676 + .6313374i	.2470649 + .6324002i
6	.2469962 + .6318688i	.2469962 + .6318685i

(see [12, III, p. 379]). Note that b_1 is the wrong choice but b_n is the right choice for $n \geqslant 2$. The algorithm appears to converge nicely.

Let us make this idea more precise with a definition.

Definition. Let $a, b \in \mathbf{C}^*$ satisfy $a \neq \pm b$. A pair of sequences $\{a_n\}_{n=0}^{\infty}$ and $\{b_n\}_{n=0}^{\infty}$ as in (2.1) is called *good* if b_{n+1} is the right choice for $(a_n b_n)^{1/2}$ for all but finitely many $n \geqslant 0$.

The following proposition shows the special role played by good sequences.

PROPOSITION 2.1. *If $a, b \in \mathbf{C}^*$ satisfy $a \neq \pm b$, then any pair of sequences $\{a_n\}_{n=0}^{\infty}$ and $\{b_n\}_{n=0}^{\infty}$ as in (2.1) converge to a common limit, and this common limit is non-zero if and only if $\{a_n\}_{n=0}^{\infty}$ and $\{b_n\}_{n=0}^{\infty}$ are good sequences.*

Proof. We first study the properties of the right choice b_1 of $(ab)^{1/2}$ in more detail. Let $0 \leqslant \mathrm{ang}(a, b) \leqslant \pi$ denote the unoriented angle between a and b.

Then we have:

$$|a_1 - b_1| \leqslant (1/2)\,|a - b| \tag{2.2}$$

$$\mathrm{ang}(a_1, b_1) \leqslant (1/2)\,\mathrm{ang}(a, b)\,. \tag{2.3}$$

To prove (2.2), note that

$$|a_1 - b_1|\,|a_1 + b_1| = (1/4)\,|a - b|^2\,.$$

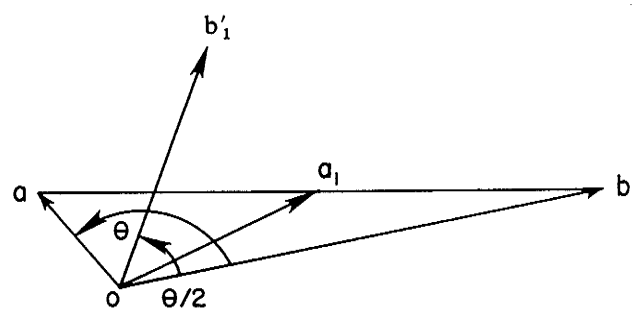
Since $|a_1 - b_1| \leqslant |a_1 + b_1|$, (2.2) follows immediately. To prove (2.3), let $\theta_1 = \mathrm{ang}(a_1, b_1)$ and $\theta = \mathrm{ang}(a, b)$. From the law of cosines

$$|a_1 \pm b_1|^2 = |a_1|^2 + |b_1|^2 \pm 2\,|a_1|\,|b_1|\cos\theta_1\,,$$

we see that $\theta_1 \leqslant \pi/2$ because $|a_1 - b_1| \leqslant |a_1 + b_1|$. Thus

$$\mathrm{ang}(a_1, b_1) = \theta_1 \leqslant \pi - 0_1 = \mathrm{ang}(a_1, -b_1)\,.$$

To compare this to θ, note that one of $\pm b_1$, say b_1', satisfies $\mathrm{ang}(a, b_1') = \mathrm{ang}(b_1', b) = \theta/2$. Then the following picture



shows that $\operatorname{ang}(a_1, b_1') \leqslant \theta/2$. Since $b_1' = \pm b_1$, the above inequalities imply that

$$\operatorname{ang}(a_1, b_1) \leqslant \operatorname{ang}(a_1, b_1') \leqslant (1/2)\operatorname{ang}(a, b),$$

proving (2.3).

Now, suppose that $\{a_n\}_{n=0}^{\infty}$ and $\{b_n\}_{n=0}^{\infty}$ are not good sequences. We set $M_n = \max\{|a_n|, |b_n|\}$, and it suffices to show that $\lim_{n\to\infty} M_n = 0$. Note that $M_{n+1} \leqslant M_n$ for $n \geqslant 0$. Suppose that for some n, b_{n+1} is not the right choice for $(a_n b_n)^{1/2}$. Then $-b_{n+1}$ is the right choice, and thus (2.2), applied to a_n and b_n, implies that

$$|a_{n+2}| = (1/2)|a_{n+1} - b_{n+1}| \leqslant (1/4)|a_n - b_n| \leqslant (1/2)M_n.$$

However, we also have $|b_{n+2}| \leqslant M_n$. It follows easily that

$$M_{n+3} \leqslant (3/4)M_n. \tag{2.4}$$

Since $\{a_n\}_{n=0}^{\infty}$ and $\{b_n\}_{n=0}^{\infty}$ are not good sequences, (2.4) must occur infinitely often, proving that $\lim_{n\to\infty} M_n = 0$.

Next, suppose that $\{a_n\}_{n=0}^{\infty}$ and $\{b_n\}_{n=0}^{\infty}$ are good sequences. By neglecting the first N terms for N sufficiently large, we may assume that b_{n+1} is the right choice for all $n \geqslant 0$ and that $\operatorname{ang}(a, b) < \pi$ (this is possible by (2.3)). We also set $\theta_n = \operatorname{ang}(a_n, b_n)$. From (2.2) and (2.3) we obtain

$$|a_n - b_n| \leqslant 2^{-n}|a - b|, \quad \theta_n \leqslant 2^{-n}\theta_0. \tag{2.5}$$

Note that $a_n - a_{n+1} = (1/2)(a_n - b_n)$, so that by (2.5),

$$|a_n - a_{n+1}| \leqslant 2^{-(n+1)}|a - b|$$

Hence, if $m > n$, we see that

$$|a_n - a_m| \leqslant \sum_{k=n}^{m-1} |a_k - a_{k+1}| \leqslant \left(\sum_{k=n}^{m-1} 2^{-(k+1)}\right)|a - b| < 2^{-n}|a - b|.$$

Thus $\{a_n\}_{n=0}^{\infty}$ converges because it is a Cauchy sequence, and then (2.5) implies that $\lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n$.

It remains to show that this common limit is nonzero. Let

$$m_n = \min\{|a_n|, |b_n|\}.$$

Clearly $|b_{n+1}| \geqslant m_n$. To relate $|a_{n+1}|$ and m_n, we use the law of cosines:

$$\begin{aligned}(2|a_{n+1}|)^2 &= |a_n|^2 + |b_n|^2 + 2|a_n||b_n|\cos\theta_n \\ &\geqslant 2m_n^2(1+\cos\theta_n) = 4m_n^2\cos^2(\theta_n/2).\end{aligned}$$

It follows that $m_{n+1} \geqslant \cos(\theta_n/2)m_n$ since $0 \leqslant \theta_n < \pi$ (this uses (2.5) and the fact that $\theta_0 = \operatorname{ang}(a, b) < \pi$). Using (2.5) again, we obtain

$$m_n \geqslant \left(\prod_{k=1}^{n} \cos(\theta_0/2^k)\right) m_0.$$

However, it is well known that

$$\prod_{k=1}^{\infty} \cos(\theta_0/2^k) = \frac{\sin\theta_0}{\theta_0}.$$

(See [16, p. 38]. When $\theta_0 = 0$, the right hand side is interpreted to be 1.) We thus have

$$m_n \geqslant \left(\frac{\sin\theta_0}{\theta_0}\right) m_0$$

for all $n \geqslant 1$. Since $0 \leqslant \theta_0 < \pi$, it follows that $\lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n \neq 0$. QED

We now define the agM of two complex numbers.

Definition. Let $a, b \in \mathbf{C}^*$ satisfy $a \neq \pm b$. A nonzero complex number μ is a value of the *arithmetic-geometric mean* $M(a, b)$ of a and b if there are good sequences $\{a_n\}_{n=0}^{\infty}$ and $\{b_n\}_{n=0}^{\infty}$ as in (2.1) such that

$$\mu = \lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n.$$

Thus $M(a, b)$ is a multiple valued function of a and b and there are a countable number of values. Note, however, that there is a distinguished value of $M(a, b)$, namely the common limit of $\{a_n\}_{n=0}^{\infty}$ and $\{b_n\}_{n=0}^{\infty}$ where b_{n+1} is the right choice for $(a_n b_n)^{1/2}$ for *all* $n \geqslant 0$. We will call this the *simplest value* of $M(a, b)$. When a and b are positive real numbers, this simplest value is just the agM as defined in § 1.

We now come to the major result of this paper, which determines how the various values of $M(a, b)$ are related for fixed a and b.

THEOREM 2.2. *Fix $a, b \in \mathbf{C}^*$ which satisfy $a \neq \pm b$ and $|a| \geqslant |b|$, and let μ and λ denote the simplest values of $M(a, b)$ and $M(a+b, a-b)$ respectively. Then all values μ' of $M(a, b)$ are given by the formula*

$$\frac{1}{\mu'} = \frac{d}{\mu} + \frac{ic}{\lambda},$$

where d *and* c *are arbitrary relatively prime integers satisfying* $d \equiv 1 \bmod 4$ *and* $c \equiv 0 \bmod 4$.

Proof. Our treatment of the agM of complex numbers thus far has been fairly elementary. The proof of this theorem, however, will be quite different; we will finally discover the "entirely new field of analysis" predicted by Gauss in the diary entry quoted in § 1. In the proof we will follow Gauss' ideas and even some of his notations, though sometimes translating them to a modern setting and of course filling in the details he omitted (Gauss' notes are extremely sketchy and incomplete — see [12, III, pp. 467-468 and 477-478]).

The proof will be broken up into four steps. In order to avoid writing a treatise on modular functions, we will quote certain classical facts without proof.

Step 1. Theta Functions

Let $\mathfrak{H} = \{\tau \in \mathbf{C} : \operatorname{Im}\tau > 0\}$ and set $q = e^{\pi i \tau}$. The Jacobi theta functions are defined as follows:

$$p(\tau) = 1 + 2\sum_{n=1}^{\infty} q^{n^2} = \Theta_3(\tau, 0),$$

$$q(\tau) = 1 + 2\sum_{n=1}^{\infty} (-1)^n q^{n^2} = \Theta_4(\tau, 0),$$

$$r(\tau) = 2\sum_{n=1}^{\infty} q^{(2n-1)^2/4} = \Theta_2(\tau, 0).$$

Since $|q| < 1$ for $\tau \in \mathfrak{H}$, these are holomorphic functions of τ. The notation p, q and r is due to Gauss, though he wrote them as power series in $e^{-\pi t}$, $\operatorname{Re} t > 0$ (thus he used the right half plane rather than the upper half plane $\mathfrak{H}$ — see [12, III, pp. 383-386]). The more common notation Θ_3, Θ_4 and Θ_2 is from [36, p. 464] and [32, p. 27].

A wealth of formulas are associated with these functions, including the product expansions:

$$p(\tau) = \prod_{n=1}^{\infty} (1-q^{2n})(1+q^{2n-1})^2,$$

$$q(\tau) = \prod_{n=1}^{\infty} (1-q^{2n})(1-q^{2n-1})^2, \tag{2.6}$$

$$r(\tau) = 2q^{1/4}\prod_{n=1}^{\infty} (1-q^{2n})(1+q^{2n})^2,$$

(which show that $p(\tau)$, $q(\tau)$ and $r(\tau)$ are nonvanishing on $\mathfrak{H}$), the transformations:

$$\begin{aligned} p(\tau+1) &= q(\tau), & p(-1/\tau) &= (-i\tau)^{1/2}p(\tau), \\ q(\tau+1) &= p(\tau), & q(-1/\tau) &= (-i\tau)^{1/2}r(\tau), \\ r(\tau+1) &= e^{\pi i/4}r(\tau), & r(-1/\tau) &= (-i\tau)^{1/2}q(\tau), \end{aligned} \tag{2.7}$$

(where we assume that $\operatorname{Re}(-i\tau)^{1/2} > 0$), and finally the identities

$$\begin{aligned} p(\tau)^2 + q(\tau)^2 &= 2p(2\tau)^2, \\ p(\tau)^2 - q(\tau)^2 &= 2r(2\tau)^2, \\ p(\tau)q(\tau) &= q(2\tau)^2, \end{aligned} \tag{2.8}$$

and

$$\begin{aligned} p(2\tau)^2 + r(2\tau)^2 &= p(\tau)^2, \\ p(2\tau)^2 - r(2\tau)^2 &= q(\tau)^2, \\ q(\tau)^4 + r(\tau)^4 &= p(\tau)^4. \end{aligned} \tag{2.9}$$

Proofs of (2.6) and (2.7) can be found in [36, p. 469 and p. 475], while one must turn to more complete works like [32, pp. 118-119] for proofs of (2.8). (For a modern proof of (2.8), consult [34].) Finally, (2.9) follows easily from (2.8). Of course, Gauss knew all of these formulas (see [12, III, pp. 386 and 466-467]).

What do these formulas have to do with the agM? The key lies in (2.8): one sees that $p(2\tau)^2$ and $q(2\tau)^2$ are the respective arithmetic and geometric means of $p(\tau)^2$ and $q(\tau)^2$! To make the best use of this observation, we need to introduce the function $k'(\tau) = q(\tau)^2/p(\tau)^2$.

Then we have:

LEMMA 2.3. *Let* $a, b \in \mathbf{C}^*$ *satisfy* $a \neq \pm b$, *and suppose there is* $\tau \in \mathfrak{H}$ *such that* $k'(\tau) = b/a$. *Set* $\mu = a/p(\tau)^2$ *and, for* $n \geqslant 0$, $a_n = \mu\, p(2^n\tau)^2$ *and* $b_n = \mu\, q(2^n\tau)^2$. *Then*

(i) $\{a_n\}_{n=0}^{\infty}$ *and* $\{b_n\}_{n=0}^{\infty}$ *are good sequences satisfying (2.1),*

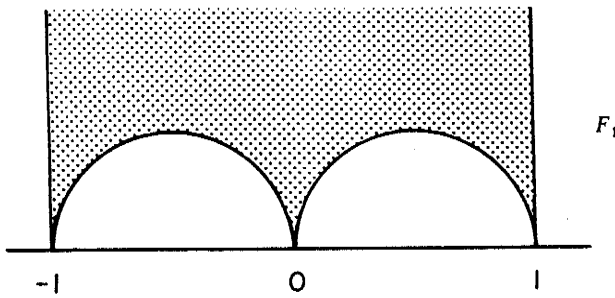
(ii) $\lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n = \mu$.

Proof. We have $a_0 = a$ by definition, and $b_0 = b$ follows easily from $k'(\tau) = b/a$. As we observed above, the other conditions of (2.1) are clearly

satisfied. Finally, note that $\exp(\pi i 2^n\tau) \to 0$ as $n \to \infty$, so that $\lim_{n\to\infty} p(2^n\tau)^2 = \lim_{n\to\infty} q(2^n\tau)^2 = 1$, and (ii) follows. Since $\mu \neq 0$, Proposition 2.1 shows that $\{a_n\}_{n=0}^{\infty}$ and $\{b_n\}_{n=0}^{\infty}$ are good sequences. QED

Thus every solution τ of $k'(\tau) = b/a$ gives us a value $\mu = a/p(\tau)^2$ of $M(a, b)$. As a first step toward understanding all solutions of $k'(\tau) = b/a$, we introduce the region $F_1 \subseteq \mathfrak{H}$:

$$F_1 = \{\tau \in \mathfrak{H} : |\operatorname{Re}\tau| \leqslant 1\,,\ |\operatorname{Re}(1/\tau)| \leqslant 1\}$$

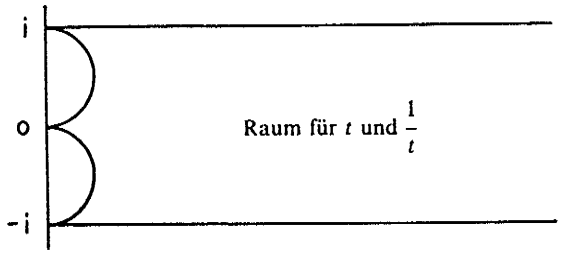


F_1 shaded

The following result is well known.

LEMMA 2.4. *k'^2 assumes every value in $\mathbf{C} - \{0, 1\}$ exactly once in $F_1' = F_1 - (\partial F_1 \cap \{\tau \in \mathfrak{H} : \operatorname{Re}\tau < 0\})$.*

A proof can be found in [36, pp. 481-484]. Gauss was aware of similar results which we will discuss below. He drew F_1 as follows (see [12, III, p. 478]).



Note that our restrictions on a and b ensure that $(b/a)^2 \in \mathbf{C} - \{0, 1\}$. Thus, by Lemma 2.4, we can always solve $k'(\tau)^2 = (b/a)^2$, i.e., $k'(\tau) = \pm b/a$. We will prove below that

$$k'\left(\frac{\tau}{2\tau+1}\right) = -k'(\tau)\,, \tag{2.10}$$

which shows that we can always solve $k'(\tau) = b/a$. Thus, for every a and b as above, $M(a, b)$ has at least one value of the form $a/p(\tau)^2$, where $k'(\tau) = b/a$.

Three tasks now remain. We need to find *all* solutions τ of $k'(\tau) = b/a$, we need to see how the values $a/p(\tau)^2$ are related for these τ's, and we need to prove that *all* values of $M(a, b)$ arise in this way. To accomplish these goals, we must first recast the properties of $k'(\tau)$ and $p(\tau)^2$ into more modern terms.

Step 2. Modular Forms of Weight One.

The four lemmas proved here are well known to experts, but we include their proofs in order to show how easily one can move from the classical facts of Step 1 to their modern interpretations. We will also discuss what Gauss had to say about these facts.

We will use the transformation properties (2.7) by way of the group

$$SL(2, \mathbf{Z}) = \left\{\begin{pmatrix} a & b \\ c & d \end{pmatrix} : a, b, c, d \in \mathbf{Z},\ ad - bc = 1\right\}$$

which acts on $\mathfrak{H}$ by linear fractional transformations as follows: if $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL(2, \mathbf{Z})$ and $\tau \in \mathfrak{H}$, then $\gamma\tau = \dfrac{a\tau + b}{c\tau + d}$.

For example, if

$$S = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \quad \text{and} \quad T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad \text{then} \quad S\tau = \frac{-1}{\tau}, \quad T\tau = \tau + 1\,,$$

which are the transformations in (2.7). It can be shown that S and T generate $SL(2, \mathbf{Z})$ (see [29, Ch. VII, Thm. 2]), a fact we do not need here.

We will consider several subgroups of $SL(2, \mathbf{Z})$. The first of these is $\Gamma(2)$, the principal congruence subgroup of level 2:

$$\Gamma(2) = \{\gamma \in SL(2, \mathbf{Z}) : \gamma \equiv \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \bmod 2\}\,.$$

Note that $-1 \in \Gamma(2)$ and that $\Gamma(2)/\{\pm 1\}$ acts on $\mathfrak{H}$.

LEMMA 2.5.

(i) *$\Gamma(2)/\{\pm 1\}$ acts freely on $\mathfrak{H}$.*

(ii) *$\Gamma(2)$ is generated by -1, $U = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}$ and $V = \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix}$.*

(iii) *Given $\tau \in \mathfrak{H}$, there is $\gamma \in \Gamma(2)$ such that $\gamma\tau \in F_1$.*

Proof. Let $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ be an element of $\Gamma(2)$.

(i) If $\tau \in \mathfrak{H}$ and $\gamma\tau = \tau$, then we obtain $c\tau^2 + (d-a)\tau - b = 0$. If $c = 0$, then $\gamma = \pm 1$ follows immediately. If $c \neq 0$, then $(d-a)^2 + 4bc < 0$ because $\tau \in \mathfrak{H}$. Using $ad - bc = 1$, this becomes $(a+d)^2 < 4$, and thus $a + d = 0$ since a and d are odd. However, b and c are even so that

$$1 \equiv ad - bc \equiv ad \equiv -a^2 \bmod 4$$

This contradiction proves (i).

(ii) We start with a variation of the Euclidean algorithm. Given γ as above, let $r_1 = a - 2a_1c$, where $a_1 \in \mathbf{Z}$ is chosen so that $|r_1|$ is minimal. Then $|r_1| \leqslant |c|$, and hence $|r_1| < |c|$ since a and c have different parity. Thus

$$a = 2a_1c + r_1, \quad a_1, r_1 \in \mathbf{Z}, \quad |r_1| < |c|.$$

Note that c and r_1 also have different parity. Continuing this process, we obtain

$$\begin{aligned} c &= 2a_2 r_1 + r_2, && |r_2| < |r_1|, \\ r_1 &= 2a_3 r_2 + r_3, && |r_3| < |r_2|, \\ &\vdots \\ r_{2n-1} &= 2a_{2n+1} r_{2n} + r_{2n+1}, && r_{2n+1} = \pm 1, \\ r_{2n} &= 2a_{2n+2} r_{2n+1} + 0, \end{aligned}$$

since $\mathrm{GCD}(a, c) = 1$. Then one easily computes that

$$V^{-a_{2n+2}} U^{-a_{2n+1}} \ldots V^{-a_2} U^{-a_1} \gamma = \begin{pmatrix} \pm 1 & * \\ 0 & * \end{pmatrix}.$$

Since the left-hand side is in $\Gamma(2)$, the right-hand side must be of the form $\pm U^m$, and we thus obtain

$$\gamma = \pm U^{a_1} V^{a_2} \ldots U^{a_{2n+1}} V^{a_{2n+2}} U^m.$$

(iii) Fix $\tau \in \mathfrak{H}$. The quadratic form $|x\tau + y|^2$ is positive definite for $x, y \in \mathbf{R}$, so that for any $S \subseteq \mathbf{Z}^2$, $|x\tau + y|^2$ assumes a minimum value at some $(x, y) \in S$. In particular, $|c\tau + d|^2$, where $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma(2)$, assumes a minimum value at some $\gamma_0 \in \Gamma(2)$. Since $\mathrm{Im}\,\gamma\tau = \mathrm{Im}\,\tau\,|c\tau + d|^{-2}$, we see

that $\tau' = \gamma_0\tau$ has maximal imaginary part, i.e., $\mathrm{Im}\,\tau' \geqslant \mathrm{Im}\,\gamma\tau'$ for $\gamma \in \Gamma(2)$. Since $\mathrm{Im}\,\tau' = \mathrm{Im}\,U\tau'$, we may assume that $|\mathrm{Re}\,\tau'| \leqslant 1$. Applying the above inequality to $V^{\pm 1} \in \Gamma(2)$, we obtain

$$\mathrm{Im}\,\tau' \geqslant \mathrm{Im}\,V^{\pm 1}\tau' = \mathrm{Im}\,\tau'\,|2\tau' \pm 1|^{-2}.$$

Thus $|2\tau \pm 1| \geqslant 1$, or $|\tau \pm (1/2)| \geqslant 1/2$. This is equivalent to $|\mathrm{Re}\,1/\tau'| \leqslant 1$, and hence $\tau' \in F_1$. QED

We next study how $p(\tau)$ and $q(\tau)$ transform under elements of $\Gamma(2)$.

LEMMA 2.6. *Let* $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma(2)$, *and assume that* $a \equiv d \equiv 1 \bmod 4$. *Then*

(i) $p(\gamma\tau)^2 = (c\tau + d)\,p(\tau)^2$,

(ii) $q(\gamma\tau)^2 = i^c(c\tau + d)\,q(\tau)^2$.

Proof. From (2.7) and $V = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} U^{-1} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$ we obtain

$$\begin{aligned} p(U\tau)^2 &= p(\tau)^2, \quad p(V\tau)^2 = (2\tau + 1)\,p(\tau)^2, \\ q(U\tau)^2 &= q(\tau)^2, \quad q(V\tau)^2 = -(2\tau + 1)\,q(\tau)^2. \end{aligned} \tag{2.11}$$

Thus (i) and (ii) hold for U and V. The proof of the previous lemma shows that γ is in the subgroup of $\Gamma(2)$ generated by U and V. We now proceed by induction on the length of γ as a word in U and V.

(i) If $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ and $p(\gamma\tau)^2 = (c\tau + d)\,p(\tau)^2$ then (2.11) implies that

$$\begin{aligned} p(U\gamma\tau)^2 &= p(\gamma\tau)^2 = (c\tau + d)\,p(\tau)^2, \\ p(V\gamma\tau)^2 &= (2\gamma\tau + 1)\,p(\gamma\tau)^2 = (2\gamma\tau + 1)(c\tau + d)\,p(\tau)^2 \\ &= ((2a + c)\tau + (2b + d))\,p(\tau)^2. \end{aligned}$$

However $U\gamma = \begin{pmatrix} * & * \\ c & d \end{pmatrix}$, $V\gamma = \begin{pmatrix} * & * \\ 2a + c & 2b + d \end{pmatrix}$, so that (i) now holds for $U\gamma$ and $V\gamma$.

(ii) Using (2.11) and arguing as above, we see that if $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} = U^{a_1} V^{b_1} \ldots U^{a_n} V^{b_n}$, then

$$q(\gamma\tau)^2 = (-1)^{\Sigma b_i}(c\tau + d)\,q(\tau)^2.$$

However, U and V commute modulo 4, so that

$$\gamma \equiv \begin{pmatrix} 1 & 2\Sigma a_i \\ 2\Sigma b_i & 1 \end{pmatrix} \bmod 4 .$$

Thus $c \equiv 2\Sigma b_i \bmod 4$, and (ii) follows. QED

Note that (2.10) is an immediate consequence of Lemma 2.6.

In order to fully exploit this lemma, we introduce the following subgroups of $\Gamma(2)$:

$$\Gamma(2)_0 = \{\gamma \in \Gamma(2) : a \equiv d \equiv 1 \bmod 4\},$$

$$\Gamma_2(4) = \{\gamma \in \Gamma(2)_0 : c \equiv 0 \bmod 4\}$$

Note that $\Gamma(2) = \{\pm 1\} \cdot \Gamma(2)_0$ and that $\Gamma_2(4)$ has index 2 in $\Gamma(2)_0$. From Lemma 2.6 we obtain

$$p(\gamma\tau)^2 = (c\tau + d)\, p(\tau)^2 , \quad \gamma \in \Gamma(2)_0 ,$$

$$q(\gamma\tau)^2 = (c\tau + d)\, q(\tau)^2 , \quad \gamma \in \Gamma_2(4) . \tag{2.12}$$

Since these functions are holomorphic on $\mathfrak{H}$, one says that $p(\tau)^2$ and $q(\tau)^2$ are weak modular forms of weight one for $\Gamma(2)_0$ and $\Gamma_2(4)$ respectively. The term more commonly used is modular form, which requires that the functions be holomorphic at the cusps (see [30, pp. 28-29] for a precise definition). Because $\Gamma(2)_0$ and $\Gamma_2(4)$ are congruence subgroups of level $N = 4$, this condition reduces to proving that

$$(c\tau + d)^{-1}\, p(\gamma\tau)^2 , \quad (c\tau + d)^{-1}\, q(\gamma\tau)^2 , \tag{2.13}$$

are holomorphic functions of $q^{1/2} = \exp(2\pi i\tau/4)$ for all $\gamma \in SL(2, \mathbf{Z})$. This will be shown later.

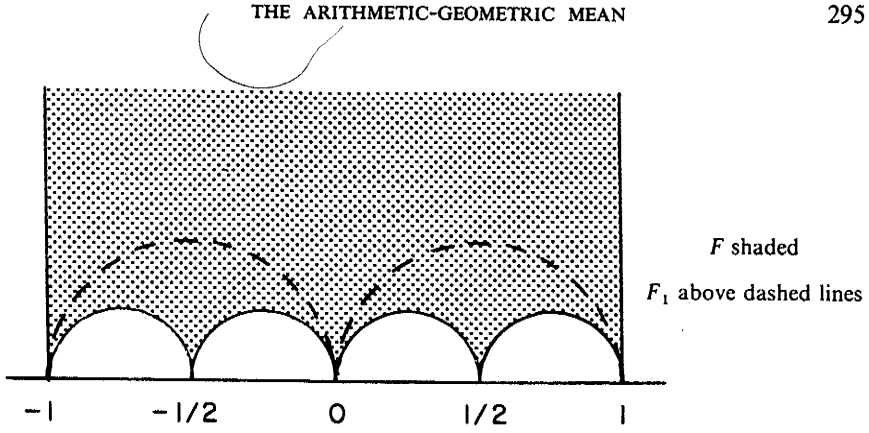
In general, it is well known that the square of a theta function is a modular form of weight one (see [27, Ch. I, § 9]), although the general theory only says that our functions are modular forms for the group

$$\Gamma(4) = \{\gamma \in SL(2, \mathbf{Z}) : \gamma \equiv \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \bmod 4\}$$

(see [27, Ch. I, Prop. 9.2]). We will need the more precise information given by (2.12).

We next study the quotients of $\mathfrak{H}$ by $\Gamma(2)$ and $\Gamma_2(4)$. From Step 1, recall the region $F_1 \subseteq \mathfrak{H}$. We now define a larger region F:

$$F = \{\tau \in \mathfrak{H} : |\operatorname{Re}\tau| \leqslant 1,\ |\tau \pm 1/4| \geqslant 1/4,\ |\tau \pm 3/4| \geqslant 1/4\} .$$



We also set

$$F_1' = F_1 - (\partial F_1 \cap \{\tau \in \mathfrak{H} : \operatorname{Re}\tau < 0\})$$

$$F' = F - (\partial F \cap \{\tau \in \mathfrak{H} : \operatorname{Re}\tau < 0\}) .$$

LEMMA 2.7. *F_1' and F' are fundamental domains for $\Gamma(2)$ and $\Gamma_2(4)$ respectively, and the functions k'^2 and k' induce biholomorphic maps*

$$\overline{k'^2} : \mathfrak{H}/\Gamma(2) \xrightarrow{\sim} \mathbf{C} - \{0, 1\}$$

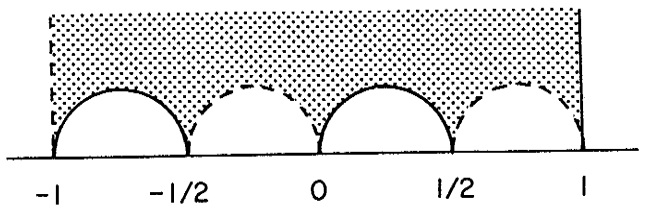
$$\overline{k'} : \mathfrak{H}/\Gamma_2(4) \xrightarrow{\sim} \mathbf{C} - \{0, \pm 1\} .$$

Proof. A simple modification of the proof of Lemma 2.6 shows that if $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma(2)$, then $p(\gamma\tau)^4 = (c\tau + d)^2\, p(\tau)^4$, $q(\gamma\tau)^4 = (c\tau + d)^2\, q(\tau)^4$. Thus k'^2 is invariant under $\Gamma(2)$.

Given $\tau \in \mathfrak{H}$, Lemma 2.5 shows that $\gamma\tau \in F_1$ for some $\gamma \in \Gamma(2)$. Since $U = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}$ maps the left vertical line in ∂F_1 to the right one and $V = \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix}$ maps the left semicircle in ∂F_1 to the right one, we may assume that $\gamma\tau \in F_1'$. If we also had $\sigma\tau \in F_1'$ for $\sigma \in \Gamma(2)$, then $k'(\sigma\tau)^2 = k'(\tau)^2 = k'(\gamma\tau)^2$, so that $\sigma\tau = \gamma\tau$ by Lemma 2.4. This shows that F_1' is a fundamental domain for $\Gamma(2)$.

Since $\Gamma(2)_0 \simeq \Gamma(2)/\{\pm 1\}$, F_1' is also a fundamental domain for $\Gamma(2)_0$. Since $\Gamma_2(4)$ has index 2 in $\Gamma(2)_0$ with 1 and V as coset representatives, it follows that

$$F^* = F_1' \cup V(F_1' \cap \{\tau \in \mathfrak{H} : \operatorname{Re}\tau \leqslant 0\}) \cup V^{-1}(F_1' \cap \{\tau \in \mathfrak{H} : \operatorname{Re}\tau > 0\})$$



F^* shaded

is a fundamental domain for $\Gamma_2(4)$. Since $\begin{pmatrix} -3 & -2 \\ -4 & -3 \end{pmatrix} \in \Gamma_2(4)$ takes the far left semicircle in ∂F to the far right one, it follows that F' is a fundamental domain for $\Gamma_2(4)$.

It now follows easily from Lemma 2.4 that k'^2 induces a bijection $\overline{k'^2}: \mathfrak{H}/\Gamma(2) \to \mathbf{C} - \{0, 1\}$. Since $\Gamma(2)/\{\pm 1\}$ acts freely on $\mathfrak{H}$ by Lemma 2.5, $\mathfrak{H}/\Gamma(2)$ is a complex manifold and $\overline{k'^2}$ is holomorphic. A straightforward argument then shows that $\overline{k'^2}$ is biholomorphic.

Next note that k' is invariant under $\Gamma_2(4)$ by (2.12), and thus induces a map $\overline{k'}: \mathfrak{H}/\Gamma_2(4) \to \mathbf{C} - \{0, \pm 1\}$. Since $\mathfrak{H}/\Gamma(2) = \mathfrak{H}/\Gamma(2)_0$, we obtain a commutative diagram:

$$\begin{array}{ccc} \mathfrak{H}/\Gamma_2(4) & \xrightarrow{\overline{k'}} & \mathbf{C} - \{0, 1\} \\ f \downarrow & & \downarrow g \\ \mathfrak{H}/\Gamma(2)_0 & \xrightarrow{\overline{k'^2}} & \mathbf{C} - \{0, 1\} \end{array}$$

where f is induced by $\Gamma_2(4) \subseteq \Gamma(2)_0$ and g is just $g(z) = z^2$. Note that g is a covering space of degree 2, and the same holds for f since $[\Gamma(2)_0 : \Gamma_2(4)] = 2$ and $\Gamma(2)_0$ acts freely on $\mathfrak{H}$. We know that $\overline{k'^2}$ is a biholomorphism, and it now follows easily that $\overline{k'}$ is also. QED

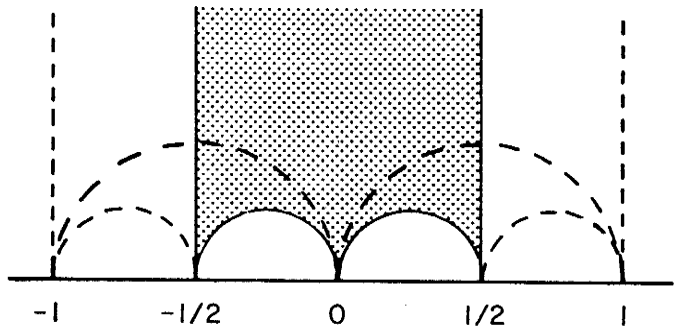
We should point out that $r(\tau)^2$ has properties similar to $p(\tau)^2$ and $q(\tau)^2$. Specifically, $r(\tau)^2$ is a modular form of weight one for the group

$$\Gamma_2(4)' = \{\gamma \in \Gamma(2) : \gamma \equiv \begin{pmatrix} 1 & 0 \\ * & 1 \end{pmatrix} \bmod 4\},$$

which is a conjugate of $\Gamma_2(4)$. Furthermore, if we set $k(\tau) = r(\tau)^2/p(\tau)^2$, then k is invariant under $\Gamma_2(4)'$ and induces a biholomorphism $\overline{k}: \mathfrak{H}/\Gamma_2(4)'$

$\to \mathbf{C} - \{0, \pm 1\}$. We leave the proofs to the reader. Note also that $k(\tau)^2 + k'(\tau)^2 = 1$ by (2.9).

Our final lemma will be useful in studying the agM. Let F_2 be the region $(1/2)F_1$, pictured below. Note that $F_2 \subseteq F$.



F_2 shaded

F, F_1 indicated by dashed lines

LEMMA 2.8.

$$k'(F_1) = \{z \in \mathbf{C} - \{0, \pm 1\} : \mathrm{Re}\, z \geqslant 0\},$$

$$k'(F_2) = \{z \in \mathbf{C} - \{0, \pm 1\} : |z| \leqslant 1\}.$$

Proof. We will only treat $k'(F_2)$, the proof for $k'(F_1)$ being quite similar. We first claim that $\{k'(\tau) : \mathrm{Re}\,\tau = \pm 1/2\} = S^1 - \{\pm 1\}$. To see this, note that $\mathrm{Re}\,\tau = \pm 1/2$ and the product expansions (2.6) easily imply that $\overline{k'(\tau)} = k'(\tau)^{-1}$, i.e., $|k'(\tau)| = 1$. How much of the circle is covered? It is easy to see that $k'(\pm 1/2 + it) \to 1$ as $t \to +\infty$. To study the limit as $t \to 0$, note that by (2.10) we have

$$k'(\pm 1/2 + it) = -k'\left(\pm 1/2 + \frac{i}{4t}\right).$$

As $t \to 0$, the right-hand side clearly approaches -1. Then connectivity arguments easily show that all of $S^1 - \{\pm 1\}$ is covered.

Since k' is injective on F' by Lemma 2.7, it follows that $k'(F_2) - S^1$ is connected. Since $|k'(it)| < 1$ for $t > 0$ by (2.6), we conclude that

$$k'(F_2) \subseteq \{z \in \mathbf{C} - \{0, \pm 1\} : |z| \leqslant 1\}.$$

Similar arguments show that

$$k'(F - F_2) \subseteq \{z \in \mathbf{C} : |z| > 1\}.$$

Since $k'(F) = \mathbf{C} - \{0, \pm 1\}$ by Lemma 2.7, both inclusions must be equalities. QED

Gauss' collected works show that he was familiar with most of this material, though it's hard to tell precisely what he knew. For example, he basically has two things to say about $k'(\tau)$:

(i) $k'(\tau)$ has positive real part for $\tau \in F_1$,

(ii) the equation $k'(\tau) = A$ has one and only one solution $\tau \in F_2$.

(See [12, III, pp. 477-478].) Neither statement is correct as written. Modifications have to be made regarding boundary behavior, and Lemma 2.8 shows that we must require $|A| \leqslant 1$ in (ii). Nevertheless, these statements show that Gauss essentially knew Lemma 2.8, and it becomes clear that he would not have been greatly surprised by Lemmas 2.4 and 2.7.

Let us see what Gauss had to say about other matters we've discussed. He was quite aware of linear fractional transformations. Since he used the right half plane, he wrote

$$t' = \frac{at - bi}{cti + d}, \quad ad - bc = 1, \quad a, b, c, d \in \mathbf{Z}, \quad \mathrm{Ret} > 0$$

(see [12, III, p. 386]). To prevent confusion, we will always translate formulas into ones involving $\tau \in \mathfrak{H}$.

Gauss decomposed an element $\gamma \in SL(2, \mathbf{Z})$ into simpler ones by means of continued fractions. For example, Gauss considers those transformations $\tau^* = \gamma\tau$ which can be written as

$$\begin{aligned} \tau' &= \frac{-1}{\tau} + 2a_1 \\ \tau'' &= \frac{-1}{\tau'} + 2a_2 \\ &\vdots \\ \tau^* = \tau^{(n)} &= \frac{-1}{\tau^{(n-1)}} + 2a_n \end{aligned} \tag{2.14}$$

(see [12, X.1, p. 223]). If $U = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}$ and $V = \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix}$, then $\tau'' = U^{a_2}V^{-a_1}\tau$, so that for n even we see a similarity to the proof of Lemma 2.5 (ii). The similarity becomes deeper once we realize that the algorithm used in the proof gives a continued fraction expansion for a/c, where $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$.

However, since n can be odd in (2.14), we are dealing with more than just elements of $\Gamma(2)$.

Gauss' real concern becomes apparent when we see him using (2.14) together with the transformation properties of $p(\tau)$. From (2.7) he obtains

$$p(\tau^*) = \sqrt{(-i\tau)(-i\tau')\cdots(-i\tau^{(n-1)})}\, p(\tau)$$

(see [12, X.1, p. 223]). The crucial thing to note is that if $\tau^* = \gamma\tau$, $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, then $(-i\tau)\cdots(-i\tau^{(n-1)})$ is just $c\tau + d$ up to a power of i. This tells us how $p(\tau)$ transforms under those γ's described by (2.14). In general, Gauss used similar methods to determine how $p(\tau)$, $q(\tau)$ and $r(\tau)$ transform under arbitrary elements γ of $SL(2, \mathbf{Z})$. The answer depends in part on how $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ reduces modulo 2. Gauss labeled the possible reductions as follows:

a	1	1	1	0	1	0
b	0	1	0	1	1	1
c	0	0	1	1	1	1
d	1	1	1	1	0	0
	1	2	3	4	5	6

(see [12, X.1, p. 224]). We recognize this as the isomorphism $SL(2, \mathbf{Z})/\Gamma(2) \simeq SL(2, \mathbf{F}_2)$, and note that (2.14) corresponds to cases 1 and 6. Then the transformations of $p(\tau)$, $q(\tau)$ and $r(\tau)$ under $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL(2, \mathbf{Z})$ are given by

		1	2	3	4	5	6
	$h^{-1}\, p(\gamma\tau) =$	$p(\tau)$	$q(\tau)$	$r(\tau)$	$q(\tau)$	$r(\tau)$	$p(\tau)$
(2.15)	$h^{-1}\, q(\gamma\tau) =$	$q(\tau)$	$p(\tau)$	$p(\tau)$	$r(\tau)$	$p(\tau)$	$r(\tau)$
	$h^{-1}\, r(\gamma\tau) =$	$r(\tau)$	$r(\tau)$	$q(\tau)$	$p(\tau)$	$q(\tau)$	$q(\tau)$

where $h = (i^\lambda(c\tau + d))^{1/2}$ and λ is an integer depending on both γ and which one of $p(\tau)$, $q(\tau)$ or $r(\tau)$ is being transformed (see [12, X.1, p. 224]). Note that Lemma 2.6 can be regarded as giving a careful analysis of λ in case 1. An analysis of the other cases may be found in [13, pp. 117-123]. One consequence of this table is that the functions (2.13) are holomorphic functions

of $q^{1/2}$, which proves that $p(\tau)^2$, $q(\tau)^2$ and $r(\tau)^2$ are modular forms, as claimed earlier.

Gauss did not make explicit use of congruence subgroups, although they appear implicitly in several places. For example, the table (2.15) shows Gauss using $\Gamma(2)$. As for $\Gamma(2)_0$, we find Gauss writing

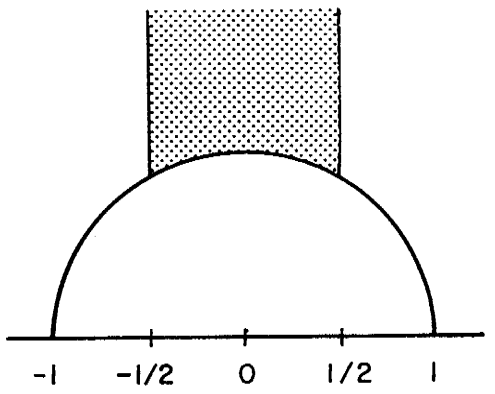
$$k'(\gamma\tau) = i^c k'(\tau)$$

where $\gamma = \begin{pmatrix} a & -b \\ -c & d \end{pmatrix}$ and, as he carefully stipulates, "$ad - bc = 1$, $a \equiv d \equiv 1 \bmod 4$, b, c even" (see [12, III, p. 478]). Also, if we ask which of these γ's leave k' unchanged, then the above equation immediately gives us $\Gamma_2(4)$, though we should be careful not to read too much into what Gauss wrote.

More interesting is Gauss' use of the reduction theory of positive definite quadratic forms as developed in Disquisitiones Arithmeticae (see [11, § 171]). This can be used to determine fundamental domains as follows. A positive definite quadratic form $ax^2 + 2bxy + cy^2$ may be written $a \mid x - \tau y \mid^2$ where $\tau \in \mathfrak{H}$. An easy computation shows that this form is equivalent via an element γ of $SL(2, \mathbf{Z})$ to another form $a' \mid x - \tau' y \mid^2$ if and only if $\tau' = \gamma^{-1}\tau$. Then, given $\tau \in \mathfrak{H}$, Gauss applies the reduction theory mentioned above to $\mid x - \tau y \mid^2$ and obtains a $SL(2, \mathbf{Z})-$ equivalent form $A \mid x - \tau' y \mid^2 = Ax^2 + 2Bxy + Cy^2$ which is reduced, i.e.

$$2 \mid B \mid \leqslant A \leqslant C$$

(see [11, § 171] and [12, X.1, p. 225]). These inequalities easily imply that $\mid \mathrm{Re}\tau' \mid \leqslant 1/2$, $\mid \mathrm{Re}\, 1/\tau' \mid \leqslant 1/2$, so that τ' lies in the shaded region



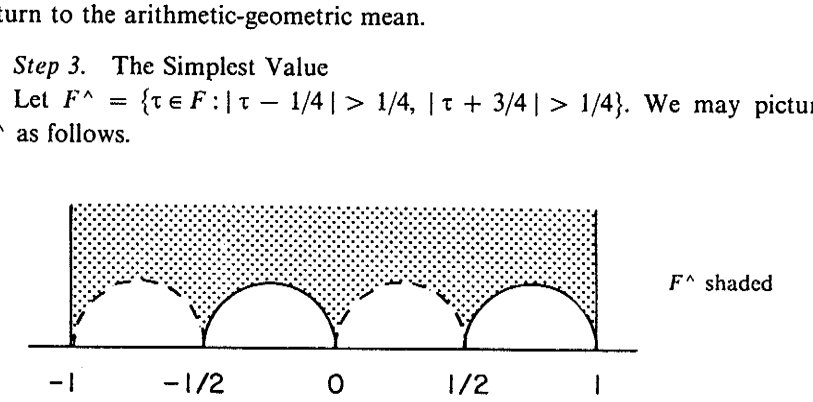
which is well known to be the fundamental domain of $SL(2, \mathbf{Z})$ acting on $\mathfrak{H}$ (see [29, Ch. VII, Thm. 1]).

This seems quite compelling, but Gauss never gave a direct connection between reduction theory and fundamental domains. Instead, he used reduction as follows: given $\tau \in \mathfrak{H}$, the reduction algorithm gives $\tau' = \gamma\tau$ as above and *at the same time* decomposes γ into a continued fraction similar to (2.14). Gauss then applies this to relate $p(\tau')$ and $p(\tau)$, etc., bringing us back to (2.15) (see [12, X.1, p. 225]). But in another place we find such continued fraction decompositions in close conjunction with geometric pictures similar to F_1 and the above (see [12, VIII, pp. 103-105]). Based on this kind of evidence, Gauss' editors decided that he did see the connection (see [12, X.2, pp. 105-106]). Much of this is still a matter of conjecture, but the fact remains that reduction theory is a powerful tool for finding fundamental domains (see [6, Ch. 12]) and that Gauss was aware of some of this power.

Having led the reader on a rather long digression, it is time for us to return to the arithmetic-geometric mean.

Step 3. The Simplest Value

Let $F^\wedge = \{\tau \in F : \mid \tau - 1/4 \mid > 1/4, \mid \tau + 3/4 \mid > 1/4\}$. We may picture $F^\wedge$ as follows.



$F^\wedge$ shaded

Let $a, b \in \mathbf{C}^*$ be as usual, and let $\tau \in \mathfrak{H}$ satisfy $k'(\tau) = b/a$. From Lemma 2.3 we know that $\mu = a/p(\tau)^2$ is a value of $M(a, b)$. The goal of Step 3 is to prove the following lemma.

LEMMA 2.9. *If* $\tau \in F^\wedge$, *then* μ *is the simplest value of* $M(a, b)$.

Proof. From Lemma 2.3 we know that

$$a_n = \mu\, p(2^n\tau)^2, \quad b_n = \mu\, q(2^n\tau)^2, \quad n = 0, 1, 2, \ldots \tag{2.16}$$

gives us good sequences converging to μ. We need to show that b_{n+1} is the right choice for $(a_n b_n)^{1/2}$ for all $n \geqslant 0$.

The following equivalences are very easy to prove:

$$|a_{n+1} - b_{n+1}| \leqslant |a_{n+1} + b_{n+1}| \Leftrightarrow \operatorname{Re}\left(\frac{b_{n+1}}{a_{n+1}}\right) \geqslant 0$$

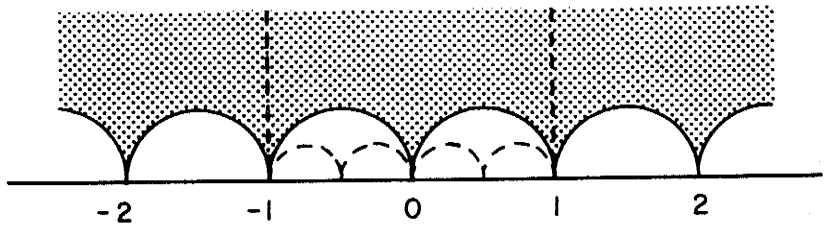
$$|a_{n+1} - b_{n+1}| = |a_{n+1} + b_{n+1}| \Leftrightarrow \operatorname{Re}\left(\frac{b_{n+1}}{a_{n+1}}\right) = 0\,.$$

Recalling the definition of the right choice, we see that we have to prove, for all $n \geqslant 0$, that $\operatorname{Re}\left(\frac{b_{n+1}}{a_{n+1}}\right) \geqslant 0$, and if $\operatorname{Re}\left(\frac{b_{n+1}}{a_{n+1}}\right) = 0$, then $\operatorname{Im}\left(\frac{b_{n+1}}{a_{n+1}}\right) > 0$. From (2.16) we see that

$$\frac{b_{n+1}}{a_{n+1}} = \frac{q(2^{n+1}\tau)^2}{p(2^{n+1}\tau)^2} = k'(2^{n+1}\tau)\,,$$

so that we are reduced to proving that if $\tau \in F^\wedge$, then for all $n \geqslant 0$, $\operatorname{Re}(k'(2^{n+1}\tau)) \geqslant 0$, and if $\operatorname{Re}(k'(2^{n+1}\tau)) = 0$, then $\operatorname{Im}(k'(2^{n+1}\tau)) > 0$.

Let $\tilde{F}_1$ denote the region obtained by translating F_1 by ± 2, ± 4, etc. The drawing below pictures both $\tilde{F}_1$ and F.

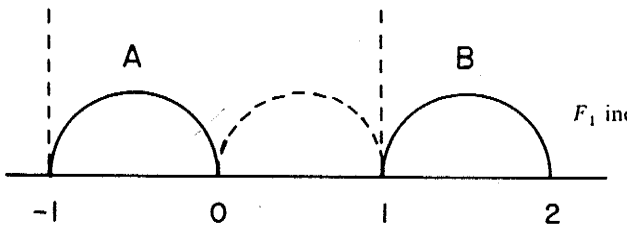


$\tilde{F}_1$ shaded
F indicated by dashed lines

Since $k'(\tau)$ has period 2 and its real part is nonnegative on F_1 by Lemma 2.8, it follows that the real part of $k'(\tau)$ is nonnegative on all of $\tilde{F}_1$. Furthermore, it is clear that on F_1, $\operatorname{Re}(k'(\tau)) = 0$ can occur only on ∂F_1. The product expansions (2.6) show that $k'(\tau)$ is real when $\operatorname{Re}\tau = \pm 1$, so that on F_1, $\operatorname{Re}(k'(\tau)) = 0$ can occur only on the boundary semicircles. From the periodicity of $k'(\tau)$ we conclude that $k'(\tau)$ has positive real part on the interior $\tilde{F}_1^0$ of $\tilde{F}_1$.

If $\tau \in F^\wedge$, then the above drawing makes it clear that $2^{n+1}\tau \in \tilde{F}_1$ for $n \geqslant 0$ and that $2^{n+1}\tau \in \tilde{F}_1^0$ for $n \geqslant 1$. We thus see that $\operatorname{Re}(k'(2^{n+1}\tau)) > 0$ for $n \geqslant 0$ unless $n = 0$ and $2\tau \in \partial\tilde{F}_1$. Thus the lemma will be proved once we show that $\operatorname{Im}(k'(2\tau)) > 0$ when $\tau \in F_1^\wedge$ and $2\tau \in \partial\tilde{F}_1$.

These last two conditions imply that 2τ lies on one of the semicircles A and B pictured below.



F_1 indicated by dashed lines

By periodicity, k' takes the same values on A and B. Thus it suffices to show that $\operatorname{Im}(k'(2\tau)) > 0$ for $2\tau \in A$. Since $S = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ maps the line $\operatorname{Re}\sigma = 1$ to A, we can write $2\tau = -1/\sigma$, where $\operatorname{Re}\sigma = 1$. Then, using (2.7), we obtain

$$k'(2\tau) = k'(-1/\sigma) = \frac{q(-1/\sigma)^2}{p(-1/\sigma)^2} = \frac{r(\sigma)^2}{p(\sigma)^2}\,.$$

Since $\operatorname{Re}\sigma = 1$, the product expansions (2.6) easily show that

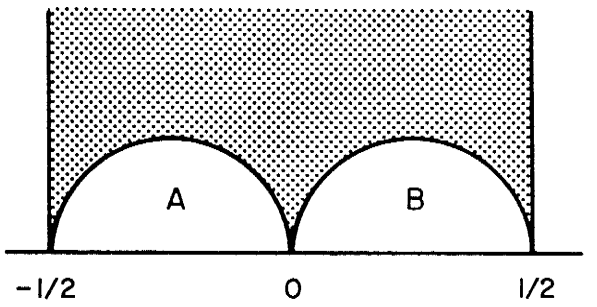
$$\operatorname{Im}(r(\sigma)^2/p(\sigma)^2) > 0\,,$$

which completes the proof of Lemma 2.9. QED

Step 4. Conclusion of the Proof.

We can now prove Theorem 2.2. Recall that at the end of Step 1 we were left with three tasks: to find all solutions τ of $k'(\tau) = b/a$, to relate the values of $a/p(\tau)^2$ thus obtained, and to show that all values of $M(a, b)$ arise in this way.

We are given $a, b \in \mathbf{C}^*$ with $a \neq \pm b$ and $|a| \geqslant |b|$. We will first find $\tau_0 \in F_2 \cap F^\wedge$ such that $k'(\tau_0) = b/a$. Since $|b/a| \leqslant 1$, Lemma 2.8 gives us $\tau_0 \in F_2$ with $k'(\tau_0) = b/a$. Could τ_0 fail to lie in $F^\wedge$? From the definition of $F^\wedge$, this only happens when τ_0 lies in the semicircle B pictured below.



F_2 shaded

However, $\gamma = \begin{pmatrix} 1 & 0 \\ -4 & 1 \end{pmatrix} \in \Gamma_2(4)$ takes B to the semicircle A. Since k' is invariant under $\Gamma_2(4)$, we have $k'(\gamma\tau_0) = k'(\tau_0) = b/a$. Thus, replacing τ_0 by $\gamma\tau_0$, we may assume that $\tau_0 \in F_2 \cap F^\wedge$.

It is now easy to solve the first two of our tasks. Since k' induces a bijection $\mathfrak{H}/\Gamma_2(4) \cong \mathbf{C} - \{0, \pm 1\}$, it follows that all solutions of $k'(\tau) = b/a$ are given by $\tau = \gamma\tau_0$, $\gamma \in \Gamma_2(4)$. This gives us the following set of values of $M(a, b)$:

$$\{a/p(\gamma\tau_0)^2 : \gamma \in \Gamma_2(4)\}.$$

Recalling the statement of Theorem 2.2, it makes sense to look at the reciprocals of these values:

$$R = \{p(\gamma\tau_0)^2/a : \gamma \in \Gamma_2(4)\}$$

By (2.12), $p(\gamma\tau_0)^2 = (c\tau_0 + d)\, p(\tau_0)^2$ for $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma_2(4) \subseteq \Gamma(2)_0$. Setting $\mu = a/p(\tau_0)^2$, we have

$$R = \{(c\tau_0 + d)\, p(\tau_0)^2/a : \gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma_2(4)\}$$

$$= \{(c\tau_0 + d)/\mu : \gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma_2(4)\}.$$

An easy exercise in number theory shows that the bottom rows (c, d) of elements of $\Gamma_2(4)$ are precisely those pairs (c, d) satisfying $GCD(c, d) = 1$, $c \equiv 0 \bmod 4$ and $d \equiv 1 \bmod 4$. We can therefore write

$$R = \{(c\tau_0 + d)/\mu : GCD(c, d) = 1, \quad c \equiv 0 \bmod 4, \quad d \equiv 1 \bmod 4\}.$$

Then setting $\lambda = i\mu/\tau_0$ gives us

$$R = \left\{\frac{d}{\mu} + \frac{ic}{\lambda} : GCD(c, d) = 1, \quad d \equiv 1 \bmod 4, \quad c \equiv 0 \bmod 4\right\}. \tag{2.17}$$

Finally, we will show that μ and λ are the simplest values of $M(a, b)$ and $M(a+b, a-b)$ respectively. This is easy to see for μ: since $\tau_0 \in F^\wedge$, Lemma 2.9 implies that $\mu = a/p(\tau_0)^2$ is the simplest value of $M(a, b)$. Turning to λ, recall from Lemma 2.3 that $a = \mu p(\tau_0)^2$ and $b = \mu q(\tau_0)^2$. Thus by (2.8) and (2.7),

$$a + b = \mu(p(\tau_0)^2 + q(\tau_0)^2) = 2\mu\, p(2\tau_0)^2 = 2\mu \left(\frac{i}{2\tau_0}\right) p\left(\frac{-1}{2\tau_0}\right)^2,$$

$$a - b = \mu(p(\tau_0)^2 - q(\tau_0)^2) = 2\mu\, r(2\tau_0)^2 = 2\mu \left(\frac{i}{2\tau_0}\right) q\left(\frac{-1}{2\tau_0}\right)^2,$$

which implies that

$$a + b = \lambda\, p(-1/2\tau_0)^2, \quad a - b = \lambda\, q(-1/2\tau_0)^2.$$

Hence λ is a value of $M(a+b, a-b)$. To see that it is the simplest value, we must show that $-1/2\tau_0 \in F^\wedge$ (by Lemma 2.9). Since $\tau_0 \in F_2$, we have $2\tau_0 \in F_1$. But F_1 is stable under $S = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$, so that $-1/2\tau_0 \in F_1$. The inclusion $F_1 \subseteq F^\wedge$ is obvious, and $-1/2\tau_0 \in F^\wedge$ follows. This completes our first two tasks.

Our third and final task is to show that (2.17) gives the reciprocals of *all* values of $M(a, b)$. This will finish the proof of Theorem 2.2. So let μ' be a value of $M(a, b)$, and let $\{a_n\}_{n=0}^\infty$ and $\{b_n\}_{n=0}^\infty$ be the good sequences such that $\mu' = \lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n$. Then there is some m such that b_{n+1} is the right choice for $(a_n b_n)^{1/2}$ for all $n \geqslant m$, and thus μ' is the simplest value of $M(a_m, b_m)$. Since $k' : F' \to \mathbf{C} - \{0, \pm 1\}$ is surjective by Lemma 2.7, we can find $\tau \in F'$ such that $k'(\tau) = b_m/a_m$. Arguing as above, we may assume that $\tau \in F^\wedge$. Then Lemma 2.9 shows that $\mu' = a_m/p(\tau)^2$ and also that for $n \geqslant m$,

$$a_n = \mu'\, p(2^{n-m}\tau)^2, \quad b_n = \mu'\, q(2^{n-m}\tau)^2. \tag{2.18}$$

Let us study a_{m-1} and b_{m-1}. Their sum and product are $2\,a_m$ and b_m^2 respectively. From the quadratic formula we see that

$$\{a_{m-1}, b_{m-1}\} = \{a_m \pm (a_m^2 - b_m^2)^{1/2}\}.$$

Using (2.9), we obtain

$$a_m^2 - b_m^2 = \mu'^2(p(\tau)^4 - q(\tau)^4) = \mu'^2 r(\tau)^4,$$

so that, again using (2.9), we have

$$a_m \pm (a_m^2 - b_m^2)^{1/2} = \mu'(p(\tau)^2 \pm r(\tau)^2) = \begin{cases} \mu' p(\tau/2)^2 \\ \mu' q(\tau/2)^2. \end{cases}$$

Thus, either

$$a_{m-1} = \mu'\, p(\tau/2)^2,\ b_{m-1} = \mu'\, q(\tau/2)^2 \text{ or } a_{m-1} = \mu'\, q(\tau/2)^2,\ b_m = \mu'\, p(\tau/2)^2.$$

In the former case, set $\tau_1 = \tau/2$. Then from (2.18) we easily see that for $n \geqslant m-1$,

$$a_n = \mu' p(2^{n-m+1}\tau_1)^2, \quad b_n = \mu' q(2^{n-m+1}\tau_1)^2. \tag{2.19}$$

If the latter case holds, let $\tau_1 = \tau/2 + 1$. From (2.7) we see that $a_{m-1} = \mu' p(\tau_1)^2$, $b_{m-1} = \mu' q(\tau_1)^2$, and it also follows easily that $p(2^{n-m+1}\tau_1) = p(2^{n-m}\tau)$ and $q(2^{n-m+1}\tau_1) = q(2^{n-m}\tau)$ for all $n \geqslant m$. Thus (2.19) holds for this choice of τ_1 and $n \geqslant m-1$.

By induction, this argument shows that there is $\tau_m \in \mathfrak{H}$ such that for all $n \geqslant 0$,

$$a_n = \mu' p(2^n\tau_m)^2, \quad b_n = \mu' q(2^n\tau_m)^2.$$

In particular, $\mu' = a/p(\tau_m)^2$ and $k'(\tau_m) = b/a$. Thus $(\mu')^{-1} = p(\tau_m)^2/a$ is in the set R of (2.17). This shows that R consists of the reciprocals of all values of $M(a, b)$, and the proof of Theorem 2.2 is now complete. QED

We should point out that the proof just given, though arrived at independently, is by no means original. The first proofs of Theorem 2.2 appeared in 1928 in [15] and [35]. Geppert's proof [15] is similar to ours in the way it uses the theory of theta functions and modular functions. The other proof [35], due to von David, is much shorter; it is a model of elegance and conciseness.

Let us discuss some consequences of the proof of Theorem 2.2. First, the formula $\lambda = i\mu/\tau_0$ obtained above is quite interesting. We say that τ_0 "uniformizes" the simplest value μ of $M(a, b)$, where

$$a = \mu\, p(\tau_0)^2, \quad b = \mu\, q(\tau_0)^2.$$

Writing the above formula as $\tau_0 = i\dfrac{\mu}{\lambda}$, we see how to *explicitly compute* τ_0 in terms of the simplest values of $M(a, b)$ and $M(a+b, a-b)$. This is especially useful when $a > b > 0$. Here, if we set $c = \sqrt{a^2 - b^2}$, then, using the notation of § 1, the simplest values are $M(a, b)$ and $M(a, c)$, so that

$$\tau_0 = i\frac{M(a, b)}{M(a, c)}. \tag{2.20}$$

A nice example is when $a = \sqrt{2}$ and $b = 1$. Then $c = 1$, which implies $\tau_0 = i$! Thus $M(\sqrt{2}, 1) = \sqrt{2}/p(i)^2 = 1/q(i)^2$. From § 1 we know $M(\sqrt{2}, 1) = \pi/\varpi$, which gives us the formulas

$$\varpi/\pi = 2^{-1/2}p(i)^2 = 2^{-1/2}(1+2e^{-\pi}+2e^{-4\pi}+2e^{-9\pi}+\ldots)^2,$$

$$\varpi/\pi = q(i)^2 = (1-2e^{-\pi}+2e^{-4\pi}-2e^{-9\pi}+\ldots)^2. \tag{2.21}$$

We will discuss the importance of this in § 3.

Turning to another topic, note that $M(a, b)$ is clearly homogeneous of degree 1, i.e., if μ is a value of $M(a, b)$, then $c\mu$ is a value of $M(ca, cb)$ for $c \in \mathbf{C}^*$. Thus, it suffices to study $M(1, b)$ for $b \in \mathbf{C} - \{0, \pm 1\}$. Its values are given by $\mu = 1/p(\tau)^2$ where $k'(\tau) = b$. Since $k' : \mathfrak{H} \to \mathbf{C} - \{0, \pm 1\}$ is a local biholomorphism, it follows that $M(1, b)$ is a multiple valued holomorphic function. To make it single valued, we pull back to the universal cover via k', giving us $M(1, k'(\tau))$. We thus obtain

$$M(1, k'(\tau)) = 1/p(\tau)^2.$$

This shows that the agM may be regarded as a meromorphic modular form of weight -1.

Another interesting multiple valued holomorphic function is the elliptic integral $\displaystyle\int_0^{\pi/2} (1-k^2\sin^2\phi)^{-1/2}d\phi$. This is a function of $k \in \mathbf{C} - \{0, \pm 1\}$. If we pull back to the universal cover via $k : \mathfrak{H} \to \mathbf{C} - \{0, \pm 1\}$ (recall from Step 2 that $k(\tau) = r(\tau)^2/p(\tau)^2$), then it is well known that

$$\frac{2}{\pi}\int_0^{\pi/2} (1-k(\tau)^2\sin^2\phi)^{-1/2}d\phi = p(\tau)^2$$

(see [36, p. 500]). Combining the above two equations, we obtain

$$\frac{1}{M(1, k'(\tau))} = p(\tau)^2 = \frac{2}{\pi}\int_0^{\pi/2} (1-k(\tau)^2\sin^2\phi)^{-1/2}d\phi,$$

which may be viewed as a rather amazing generalization of (1.9).

Finally, let us make some remarks about the set $\mathcal{M}$ of values of $M(a, b)$, where a and b are fixed. If μ denotes the simplest value of $M(a, b)$, then it can be shown that $|\mu| \geqslant |\mu'|$ for $\mu' \in \mathcal{M}$, and $|\mu|$ is a strict maximum if $\operatorname{ang}(a, b) \neq \pi$. This may be proved directly from the definitions (see [35]). Another proof proceeds as follows. We know that any $\mu' \in \mathcal{M}$ can be written

$$\mu' = \mu/(c\tau_0 + d), \tag{2.22}$$

where $\tau_0 \in F_2$ and $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma_2(4)$. Thus it suffices to prove that $|c\tau_0 + d| \geqslant 1$ whenever $\tau_0 \in F_2$ and $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma_2(4)$. This is left as an exercise for the reader.

We can also study the accumulation points of $\mathscr{M}$. Since $|c\tau_0 + d|$ is a positive definite quadratic form in c and d, it follows from (2.22) that $0 \in \mathbf{C}$ is the only accumulation point of $\mathscr{M}$. This is very satisfying once we recall from Proposition 2.1 that $0 \in \mathbf{C}$ is the common limit of all non-good sequences $\{a_n\}_{n=0}^{\infty}$ and $\{b_n\}_{n=0}^{\infty}$ coming from (2.1).

The proof of Theorem 2.2 makes one thing very clear: we have now seen "an entirely new field of analysis." However, before we can say that Gauss' prediction of May 30, 1799 has been fulfilled, we need to show that the proof given above reflects what Gauss actually did. Since we know from Step 2 about his work with the theta functions $p(\tau)$, $q(\tau)$ and $r(\tau)$ and the modular function $k'(\tau)$, it remains to see how he applied all of this to the arithmetic-geometric mean.

The connections we seek are found in several places in Gauss' notes. For example, he states very clearly that if

$$a = \mu\, p(\tau)^2, \quad b = \mu\, q(\tau)^2, \tag{2.23}$$

then the sequences $a_n = \mu\, p(2^n\tau)^2$, $b_n = \mu\, q(2^n\tau)^2$ satisfy the agM algorithm (2.1) with μ as their common limit (see [12, III, p. 385 and pp. 467-468]). This is precisely our Lemma 2.3. In another passage, Gauss defines the "einfachste Mittel" (simplest mean) to be the limit of those sequences where $\mathrm{Re}(b_{n+1}/a_n) > 0$ for all $n \geqslant 0$ (see [12, III, p. 477]). This is easily seen to be equivalent to our definition of simplest value when $\mathrm{ang}(a, b) \neq \pi$. On the same page, Gauss then asserts that for $\tau \in F_2$, μ is the simplest value of $M(a, b)$ for a, b as in (2.23). This is a weak form of Lemma 2.9. Finally, consider the following quote from [12, VIII, p. 101]: "In order to solve the equation $\dfrac{q(t)}{p(t)} = A$, one sets $A^2 = n/m$ and takes the agM of m and n; let this be μ. One further takes the agM of m and $\sqrt{m^2 - n^2}$, or, what is the same, of $m + n$ and $m - n$; let this be λ. One then has $t = \mu/\lambda$. This gives only one value of t; all others are contained in the formula

$$t' = \frac{\alpha t - 2\beta i}{\delta - 2\gamma t i},$$

where $\alpha, \beta, \gamma, \delta$ signify all integers which satisfy the equation $\alpha\delta - 4\beta\gamma = 1$." Recall that $\mathrm{Re}\, t > 0$, so that our τ is just ti. Note also that the last assertion is not quite correct.

Unfortunately, in spite of these compelling fragments, Gauss never actually stated Theorem 2.2. The closest he ever came is the following quote from [12, X.1, p. 219]: "The agM changes, when one chooses the negative value for one of n', n'', n''' etc.: however all resulting values are of the following form:

$$\frac{1}{(\mu)} = \frac{1}{\mu} + \frac{4ik}{\lambda}." \tag{2.24}$$

Here, Gauss is clearly dealing with $M(m, n)$ where $m > n > 0$. The fraction $1/\mu$ in (2.24) is correct: in fact, it can be shown that if the negative value of $n^{(r)}$ is chosen, and all other choices are the right choice, then the corresponding value μ' of $M(m, n)$ satisfies

$$\frac{1}{\mu'} = \frac{1}{\mu} + \frac{2^{r+1}i}{\lambda}$$

(see [13, p. 140]). So (2.24) is only a very special case of Theorem 2.2.

There is one final piece of evidence to consider: the 109th entry in Gauss' mathematical diary. It reads as follows:

> Between two given numbers there are always infinitely many means both arithmetic-geometric and harmonic-geometric, the observation of whose mutual connection has been a source of happiness for us.

(See [12, X.1, p. 550]. The harmonic-geometric mean of a and b is $M(a^{-1}, b^{-1})^{-1}$.) What is amazing is the date of this entry: June 3, 1800, a little more than a year after May 30, 1799. We know from § 1 that Gauss' first proofs of Theorem 1.1 date from December 1799. So less than six months later Gauss was aware of the multiple valued nature of $M(a, b)$ and of the relations among these values! One tantalizing question remains: does the phrase "mutual connection" refer only to (2.24), or did Gauss have something more like Theorem 2.2 in mind? Just how much did he know about modular functions as of June 3, 1800? In order to answer these questions, we need to examine the history of the whole situation more closely.

3. Historical remarks

The main difficulty in writing about the history of mathematics is that so much has to be left out. The mathematics we are studying has a richness which can never be conveyed in one article. For instance, our discussion of Gauss' proofs of Theorem 1.1 in no way does justice to the complexity of his mathematical thought; several important ideas were simplified or omitted altogether. This is not entirely satisfactory, yet to rectify such gaps is beyond the scope of this paper. As a compromise, we will explore the three following topics in more detail:

A. The history of the lemniscate,

B. Gauss' work on inverting lemniscatic integrals, and

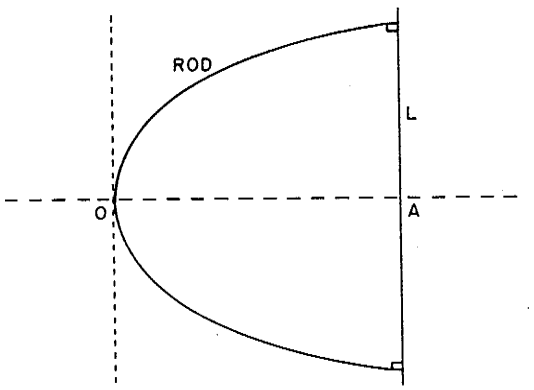
C. The chronology of Gauss' work on the agM and theta functions.

A. The lemniscate was discovered by Jacob Bernoulli in 1694. He gives the equation in the form

$$xx + yy = a\sqrt{xx - yy}$$

(in § 1 we assumed that $a = 1$), and he explains that the curve has "the form of a figure 8 on its side, as of a band folded into a knot, or of a lemniscus, or of a knot of a French ribbon" (see [2, p. 609]). "Lemniscus" is a Latin word (taken from the Greek) meaning a pendant ribbon fastened to a victor's garland.

More interesting is that the integral $\int_0^1 (1-z^4)^{-1/2}dz$, which gives one-quarter of the arc length of the lemniscate, had been discovered three years earlier in 1691! This was when Bernoulli worked out the equation of the so-called *elastic curve*. The situation is as follows: a thin elastic rod is bent until the two ends are perpendicular to a given line L.



After introducing cartesian coordinates as indicated and letting a denote 0A, Bernoulli was able to show that the upper half of the curve is given by the equation

$$y = \int_0^x \frac{z^2dz}{\sqrt{a^4 - z^4}}, \tag{3.1}$$

where $0 \leqslant x \leqslant a$ (see [2, pp. 567-600]).

It is convenient to assume that $a = 1$. But as soon as this is done, we no longer know how long the rod is. In fact, (3.1) implies that the arc length from the origin to a point (x, y) on the rescaled elastic curve is $\int_0^x (1-z^4)^{-1/2}dz$. Thus the length of the whole rod is $2\int_0^1 (1-z^4)^{-1/2}dz$, which is precisely Gauss' ϖ!

How did Bernoulli get from here to the lemniscate? He was well aware of the transcendental nature of the elastic curve, and so he used a standard seventeenth century trick to make things more manageable: he sought "an algebraic curve... whose rectification should agree with the rectification of the elastic curve" (this quote is from Euler [9, XXI, p. 276]).

Jacob actually had a very concrete reason to be interested in arc length: in 1694, just after his long paper on the elastic curve was published, he solved a problem of Leibniz concerning the "isochrona paracentrica" (see [2, pp. 601-607]). This called for a curve along which a falling weight recedes from or approaches a given point equally in equal times. Since Bernoulli's solution involved the arc length of the elastic curve, it was natural for him to seek an algebraic curve with the same arc length. Very shortly thereafter, he found the equation of the lemniscate (see [2, pp. 608-612]). So we really can say that the arc length of the lemniscate was known well before the curve itself.

But this is not the full story. In 1694 Jacob's younger brother Johann independently discovered the lemniscate! Jacob's paper on the isochrona paracentrica starts with the differential equation

$$(xdx+ydy)\sqrt{y} = (xdy-ydx)\sqrt{a},$$

which had been derived earlier by Johann, who, as Jacob rather bluntly points out, hadn't been able to solve it. Johann saw this comment for the first time when it appeared in June 1694 in Acta Eruditorum. He took up the challenge and quickly produced a paper on the isochrona paracentrica which gave the equation of the lemniscate and its relation to the elastic curve. This appeared in Acta Eruditorum in October 1694 (see [3, pp. 119-

122]), but unfortunately for Johann, Jacob's article on the lemniscate appeared in the September issue of the same journal. There followed a bitter priority dispute. Up to now relations between the brothers had been variable, sometimes good, sometimes bad, with always a strong undercurrent of competition between them. After this incident, amicable relations were never restored. (For details of this controversy, as well as a fuller discussion of Jacob's mathematical work, see [18].)

We need to mention one more thing before going on. Near the end of Jacob's paper on the lemniscate, he points out that the y-value $\int_0^x z^2(a^4-z^4)^{-1/2}dz$ of the elastic curve can be expressed as the difference of an arc of the ellipse with semiaxes $a\sqrt{2}$ and a, and an arc of the lemniscate (see [2, pp. 611-612]). This observation is an easy consequence of the equation

$$\int_0^x \frac{a^2dz}{(a^4-z^4)^{1/2}} + \int_0^x \frac{z^2dz}{(a^4-z^4)^{1/2}} = \int_0^x \left(\frac{a^2+z^2}{a^2-z^2}\right)^{1/2} dz\,. \tag{3.2}$$

What is especially intriguing is that the ratio $\sqrt{2}:1$, so important in Gauss' observation of May 30, 1799, was present at the very birth of the lemniscate.

Throughout the eighteenth century the elastic curve and the lemniscate appeared in many papers. A lot of work was done on the integrals $\int_0^1 (1-z^4)^{-1/2}dz$ and $\int_0^1 z^2(1-z^4)^{-1/2}dz$. For example, Stirling, in a work written in 1730, gives the approximations

$$\int_0^1 \frac{dz}{\sqrt{1-z^4}} = 1.31102877714605987$$

$$\int_0^1 \frac{z^2dz}{\sqrt{1-z^4}} = .59907011736779611$$

(see [31, pp. 57-58]). Note that the second number doubled is 1.19814023473559222, which agrees with $M(\sqrt{2}, 1)$ to sixteen decimal places. Stirling also comments that these two numbers add up to one half the circumference of an ellipse with $\sqrt{2}$ and 1 as axes, a special case of Bernoulli's observation (3.2).

Another notable work on the elastic curve was Euler's paper "De miris proprietatibus curvae elasticae sub equatione $y = \int \frac{xxdx}{\sqrt{1-x^4}}$ contentae"

which appeared posthumously in 1786. In this paper Euler gives approximations to the above integrals (not as good as Stirling's) and, more importantly, proves the amazing result that

$$\int_0^1 \frac{dz}{\sqrt{1-z^2}} \cdot \int_0^1 \frac{z^2dz}{\sqrt{1-z^4}} = \frac{\pi}{4} \tag{3.3}$$

(see [9, XXI, pp. 91-118]). Combining this with Theorem 1.1 we see that

$$M(\sqrt{2}, 1) = 2\int_0^1 \frac{z^2dz}{\sqrt{1-z^4}},$$

so that the coincidence noted above has a sound basis in fact.

We have quoted these two papers on the elastic curve because, as we will see shortly, Gauss is known to have read them. Note that each paper has something to contribute to the equality $M(\sqrt{2}, 1) = \pi/\varpi$: from Stirling, we get the ratio $\sqrt{2}$: 1, and from Euler we get the idea of using an equation like (3.3).

Unlike the elastic curve, the story of the lemniscate in the eighteenth century is well known, primarily because of the key role it played in the development of the theory of elliptic integrals. Since this material is thoroughly covered elsewhere (see, for example, [1, Ch. 1-3], [8, pp. 470-496], [19, § 1-§ 4] and [21, § 19.4]), we will mention only a few highlights. One early worker was C. G. Fagnano who, following some ideas of Johann Bernoulli, studied the ways in which arcs of ellipses and hyperbolas can be related. One result, known as Fagnano's Theorem, states that the sum of two appropriately chosen arcs of an ellipse can be computed algebraically in terms of the coordinates of the points involved. He also worked on the lemniscate, starting with the problem of halving that portion of the arc length of the lemniscate which lies in one quadrant. Subsequently he found methods for dividing this arc length into n equal pieces, where $n = 2^m$, $3\cdot 2^m$ or $5\cdot 2^m$. These researches of Fagnano's were published in the period 1714-1720 in an obscure Venetian journal and were not widely known. In 1750 he had his work republished, and he sent a copy to the Berlin Academy. It was given to Euler for review on December 23, 1751. Less than five weeks later, on January 27, 1752, Euler read a paper giving new derivations for Fagnano's results on elliptic and hyperbolic arcs as well as significantly new results on lemniscatic arcs. By 1753 he had a general addition theorem for lemniscatic integrals, and by 1758 he had the addition theorem for elliptic integrals (see [9, XX, pp. VII-VIII]). This material was finally published in 1761,

and for the first time there was a genuine theory of elliptic integrals. For the next twenty years Euler and Lagrange made significant contributions, paving the way for Legendre to cast the field in its classical form which we glimpsed at the end of § 1. Legendre published his definitive treatise on elliptic integrals in two volumes in 1825 and 1826. The irony is that in 1828 he had to publish a third volume describing the groundbreaking papers of Abel and Jacobi which rendered obsolete much of his own work (see [23]).

An important problem not mentioned so far is that of computing tables of elliptic integrals. Such tables were needed primarily because of the many applications of elliptic integrals to mechanics. Legendre devoted the entire second volume of his treatise to this problem. Earlier Euler had computed these integrals using power series similar to (1.8) (see also [9, XX, pp. 21-55]), but these series often converged very slowly. The real breakthrough came in Lagrange's 1785 paper "Sur une nouvelle méthode de calcul intégral" (see [22, pp. 253-312]). Among other things, Lagrange is concerned with integrals of the form

$$\int \frac{M\,dy}{\sqrt{(1+p^2y^2)\,(1+q^2y^2)}}, \tag{3.4}$$

where M is a rational function of y^2 and $p \geqslant q > 0$. He defines sequences $p, p', p'', \ldots, q, q', q'', \ldots$ as follows:

$$\begin{aligned} p' &= p + (p^2-q^2)^{1/2}, & q' &= p - (p^2-q^2)^{1/2}, \\ p'' &= p' + (p'^2-q'^2)^{1/2}, & q'' &= p' - (p'^2-q'^2)^{1/2}, \\ &\vdots & &\vdots \end{aligned} \tag{3.5}$$

and then, using the substitution

$$y' = \frac{y\left((1+p^2y^2)\,(1+q^2y^2)\right)^{1/2}}{1+q^2y^2} \tag{3.6}$$

he shows that

$$\left((1+p^2y^2)\,(1+q^2y^2)\right)^{-1/2}dy = \left((1+p'^2y'^2)\,(1+q'^2y'^2)\right)^{-1/2}dy'. \tag{3.7}$$

Two methods of approximation are now given. The first starts by observing that the sequence $p, p', p'', \ldots$ approaches $+\infty$ while $q, q', q'', \ldots$ approaches 0. Thus by iterating the substitution (3.6) in the integral of (3.4),

one can eventually assume that $q = 0$, which gives an easily computable integral. The second method consists of doing the first backwards: from (3.5) one easily obtains

$$p = (p'+q')/2, \qquad q = (p'q')^{1/2}.$$

Lagrange then observes that continuing this process leads to sequences $p', p, 'p, ''p, \ldots, q', q, 'q, ''q, \ldots$ which converge to a common limit (see [22, p. 271]). Hence iterating (3.6) allows one to eventually assume $p = q$, again giving an easily computable integral.

So here we are in 1785, staring at the definition of the arithmetic-geometric mean, six years before Gauss' earliest work on the subject. By setting $py = \tan\phi$, one obtains

$$\left((1+p^2y^2)\,(1+q^2y^2)\right)^{-1/2}dy = (p^2\cos^2\phi + q^2\sin^2\phi)^{-1/2}d\phi,$$

so that (3.6) and (3.7) are precisely (1.5) and (1.6) from the proof of Theorem 1.1. Thus Lagrange not only could have defined the agM, he could have also proved Theorem 1.1 effortlessly. Unfortunately, none of this happened; Lagrange never realized the power of what he had discovered.

One question emerges from all of this: did Gauss ever see Lagrange's article? The library of the Collegium Carolinum in Brunswick had some of Lagrange's works (see [4, p. 9]) and the library at Gottingen had an extensive collection (see [12, X.2, p. 22]). On the other hand, Gauss, in the research announcement of his 1818 article containing the proof of Theorem 1.1, claims that his work is independent of that of Lagrange and Legendre (see [12, III, p. 360]). A fuller discussion of these matters is in [12, X.2, pp. 12-22]. Assuming that Gauss did discover the agM independently, we have the amusing situation of Gauss, who anticipated so much in Abel, Jacobi and others, himself anticipated by Lagrange.

The elastic curve and the lemniscate were equally well known in the eighteenth century. As we will soon see, Gauss at first associated the integral $\int (1-z^4)^{-1/2}dz$ with the elastic curve, only later to drop it in favor of the lemniscate. Subsequent mathematicians have followed his example. Today, the elastic curve has been largely forgotten, and the lemniscate has suffered the worse fate of being relegated to the polar coordinates section of calculus books. There it sits next to the formula for arc length in polar coordinates, which can never be applied to the lemniscate since such texts know nothing of elliptic integrals.

B. Our goal in describing Gauss' work on the lemniscate is to learn more of the background to his observation of May 30, 1799. We will see that the lemniscatic functions played a key role in Gauss' development of the arithmetic-geometric mean.

Gauss began innocently enough in September 1796, using methods of Euler to find the formal power series expansion of the inverse function of first $\int (1-x^3)^{-1/2}dx$, and then more generally $\int (1-x^n)^{-1/2}dx$ (see [12, X.1, p. 502]). Things became more serious on January 8, 1797. The 51st entry in his mathematical diary, bearing this date, states that "I have begun to investigate the elastic curve depending on $\int (1-x^4)^{-1/2}dx$." Notes written at the same time show that Gauss was reading the works of Euler and Stirling on the elastic curve, as discussed earlier. Significantly, Gauss later struck out the word "elastic" and replaced it with "lemniscatic" (see [12, X.1, pp. 147 and 510]).

Gauss was strongly motivated by the analogy to the circular functions. For example, notice the similarity between $\varpi/2 = \int_0^1 (1-z^4)^{-1/2}dz$ and $\pi/2 = \int_0^1 (1-z^2)^{-1/2}dz$. (This similarity is reinforced by the fact that many eighteenth century texts used ϖ to denote π — see [12, X.2, p. 33].) Gauss then defined the lemniscatic functions as follows:

$$\operatorname{sinlemn}\left(\int_0^x (1-z^4)^{-1/2}dz\right) = x$$

$$\operatorname{coslemn}\left(\varpi/2 - \int_0^x (1-z^4)^{-1/2}dz\right) = x$$

(see [12, III, p. 404]). Gauss often used the abbreviations $\operatorname{sl}\phi$ and $\operatorname{cl}\phi$ for $\operatorname{sinlemn}\phi$ and $\operatorname{coslemn}\phi$ respectively, a practice we will adopt. From Euler's addition theorem one easily obtains

$$\operatorname{sl}^2\phi + \operatorname{cl}^2\phi + \operatorname{sl}^2\phi \operatorname{cl}^2\phi = 1 \tag{3.8}$$

$$\operatorname{sl}(\phi+\phi') = \frac{\operatorname{sl}\phi \operatorname{cl}\phi' + \operatorname{sl}\phi' \operatorname{cl}\phi}{1 - \operatorname{sl}\phi \operatorname{sl}\phi' \operatorname{cl}\phi \operatorname{cl}\phi'} \tag{3.9}$$

(see [12, X.1, p. 147]).

Other formulas can now be derived in analogy with the trigonometric functions (see [25, pp. 155-156] for a nice treatment), but Gauss went much, much farther. A series of four diary entries made in March 1797 reveal the amazing discoveries that he made in the first three months of 1797. We will need to describe these results in some detail.

Gauss started with Fagnano's problem of dividing the lemniscate into n equal parts. Since this involved an equation of degree n^2, Gauss realized that most of the roots were complex (see [12, X.1, p. 515]). This led him to define $\operatorname{sl}\phi$ and $\operatorname{cl}\phi$ for complex numbers ϕ. The first step is to show that

$$\operatorname{sl}(iy) = i \operatorname{sl} y, \qquad \operatorname{cl}(iy) = 1/\operatorname{cl}(y),$$

(the first follows from the change of variable $z = iz'$ in $\int (1-z^4)^{-1/2}dz$, and the second follows from (3.8)). Then (3.9) implies that

$$\operatorname{sl}(x+iy) = \frac{\operatorname{sl} x + i \operatorname{sl} y \operatorname{cl} x \operatorname{cl} y}{\operatorname{cl} y - i \operatorname{sl} x \operatorname{sl} y \operatorname{cl} x}$$

(see [12, X.1, p. 154]).

It follows easily that $\operatorname{sl}\phi$ is doubly periodic, with periods 2ϖ and $2i\varpi$. The zeros and poles of $\operatorname{sl}\phi$ are also easy to determine; they are given by $\phi = (m+in)\varpi$ and $\phi = ((2m-1)+i(2n-1))(\varpi/2)$, $m, n \in \mathbf{Z}$, respectively. Then Gauss shows that $\operatorname{sl}\phi$ can be written as

$$\operatorname{sl}\phi = \frac{M(\phi)}{N(\phi)}$$

where $M(\phi)$ and $N(\phi)$ are entire functions which are doubly indexed infinite products whose factors correspond to the zeros and poles respectively (see [12, X.1, pp. 153-155]). In expanding these products, Gauss writes down the first examples of Eisenstein series (see [12, X.1, pp. 515-516]). He also obtains many identities involving $M(\phi)$ and $N(\phi)$, such as

$$N(2\phi) = M(\phi)^4 + N(\phi)^4 \tag{3.10}$$

(see [12, X.1, p. 157]). Finally, Gauss notices that the numbers $N(\varpi)$ and $e^{\pi/2}$ agree to four decimal places (see [12, X.1, p. 158]). He comments that a proof of their equality would be "a most important advancement of analysis" (see 12, X.1, p. 517]).

Besides being powerful mathematics, what we have here is almost a rehearsal for what Gauss did with the arithmetic-geometric mean: the

observation that two numbers are equal, the importance to analysis of proving this, and the passage from real to complex numbers in order to get at the real depth of the subject. Notice also that identities such as (3.10) are an important warm-up to the theta function identities needed in § 2.

Two other discoveries made at this time require comment. First, only a year after constructing the regular 17-gon by ruler and compass, Gauss found a ruler and compass construction for dividing the lemniscate into five equal pieces (see [12, X.1, p. 517]). This is the basis for the remarks concerning $\int (1-x^4)^{-1/2}dx$ made in Disquisitiones Arithmeticae (see [11, § 335]). Second, Gauss discovered the complex multiplication of elliptic functions when he gave formulas for $\mathrm{sl}(1+i)\phi$, $N(1+i)\phi$, etc. (see [12, III, pp. 407 and 411]). These discoveries are linked: complex multiplication on the elliptic curve associated to the lemniscate enabled Abel to determine all n for which the lemniscate can be divided into n pieces by ruler and compass. (The answer is the same as for the circle! See [28] for an excellent modern account of Abel's theorem.)

After this burst of progress, Gauss left the lemniscatic functions to work on other things. He returned to the subject over a year later, in July 1798, and soon discovered that there was a better way to write $\mathrm{sl}\,\phi$ as a quotient of entire functions. The key was to introduce the new variable $s = \sin\left(\frac{\pi}{\varpi}\phi\right)$. Since $\mathrm{sl}\,\phi$ has period 2ϖ, it can certainly be written as a function of s. By expressing the zeros and poles of $\mathrm{sl}\,\phi$ in terms of s, Gauss was able to prove that

$$\mathrm{sl}\,\phi = \frac{P(\phi)}{Q(\phi)},$$

where

$$P(\phi) = \frac{\varpi}{\pi} s\left(1 + \frac{4s^2}{(e^{\pi}-e^{-\pi})^2}\right)\left(1 + \frac{4s^2}{(e^{2\pi}-e^{-2\pi})^2}\right)\cdots$$

$$Q(\phi) = \left(1 - \frac{4s^2}{(e^{\pi/2}+e^{-\pi/2})^2}\right)\left(1 - \frac{4s^2}{(e^{3\pi/2}+e^{-3\pi/2})^2}\right)\cdots$$

(see [12, III, pp. 415-416]). Relating these to the earlier functions $M(\phi)$ and $N(\phi)$, Gauss obtains (letting $\phi = \psi\varpi$)

$$M(\psi\varpi) = e^{\pi\psi^2/2} P(\psi\varpi),$$

$$N(\psi\varpi) = e^{\pi\psi^2/2} Q(\psi\varpi),$$

(see [12, III, p. 416]). Notice that $N(\varpi) = e^{\pi/2}$ is an immediate consequence of the second formula.

Many other things were going on at this time. The appearance of π/ϖ sparked Gauss' interest in this ratio. He found several ways of expressing ϖ/π, for example

$$\frac{\varpi}{\pi} = \frac{\sqrt{2}}{2}\left(1 + \left(\frac{1}{2}\right)^2\frac{1}{2} + \left(\frac{3}{8}\right)^2\frac{1}{4} + \left(\frac{5}{16}\right)^2\frac{1}{8} + \cdots\right), \tag{3.11}$$

and he computed ϖ/π to fifteen decimal places (see [12, X.1, p. 169]). He also returned to some of his earlier notes and, where the approximation $2\int_0^1 z^2(1-z^4)^{-1/2}dz \approx 1.198$ appears, he added that this is π/ϖ (see [12, X.1, pp. 146 and 150]). Thus in July 1798 Gauss was intimately familiar with the right-hand side of the equation $M(\sqrt{2}, 1) = \pi/\varpi$. Another problem he studied was the Fourier expansion of $\mathrm{sl}\,\phi$. Here, he first found the numerical value of the coefficients, i.e.

$$\mathrm{sl}\,\psi\varpi = .95500599 \sin\psi\pi - .04304950 \sin 3\psi\pi + \cdots,$$

and then he found a formula for the coefficients, obtaining

$$\mathrm{sl}\,\psi\varpi = \frac{4\pi}{\varpi(e^{\pi/2}+e^{-\pi/2})}\sin\psi\pi - \frac{4\pi}{\varpi(e^{3\pi/2}+e^{-3\pi/2})}\sin 3\psi\pi + \cdots$$

see [12, X.1, p. 168 and III, p. 417]).

The next breakthrough came in October 1798 when Gauss computed the Fourier expansions of $P(\phi)$ and $Q(\phi)$. As above, he first computed the coefficients numerically and then tried to find a general formula for them. Since he suspected that numbers like $e^{-\pi}$, $e^{-\pi/2}$, etc., would be involved, he computed several of these numbers (see [12, III, pp. 426-432]). The final formulas he found were

$$P(\psi\varpi) = 2^{3/4}(\pi/\varpi)^{1/2}\left(e^{-\pi/4}\sin\psi\pi - e^{-9\pi/4}\sin 3\psi\pi + e^{-25\pi/4}\sin 5\psi\pi - \cdots\right)$$

$$Q(\psi\varpi) = 2^{-1/4}(\pi/\varpi)^{1/2}\left(1 + 2e^{-\pi}\cos 2\psi\pi + 2e^{-4\pi}\cos 4\psi\pi + 2e^{-9\pi}\cos 6\psi\pi + \cdots\right) \tag{3.12}$$

(see [12, X.1, pp. 536-537]). A very brief sketch of how Gauss proved these formulas may be found in [12, X.2, pp. 38-39].

These formulas are remarkable for several reasons. First, recall the theta functions Θ_1 and Θ_3:

$$\Theta_1(z, q) = 2q^{1/4} \sin z - 2q^{9/4} \sin 3z + 2q^{25/4} \sin 5z - \dots$$

(3.13)

$$\Theta_3(z, q) = 1 + 2q \cos 2z + 2q^4 \cos 4z + 2q^9 \cos 6z + \dots$$

(see [36, p. 464]). Up to the constant factor $2^{-1/4}(\pi/\varpi)^{1/2}$, we see that $P(\psi\varpi)$ and $Q(\psi\varpi)$ are precisely $\Theta_1(\psi\pi, e^{-\pi})$ and $\Theta_3(\psi\pi, e^{-\pi})$ respectively. Even though this is just a special case, one can easily discern the general form of the theta functions from (3.12). (For more on the relation between theta functions and sl ϕ, see [36, pp. 524-525]).

Several interesting formulas can be derived from (3.12) by making specific choice for ψ. For example, if we set $\psi = 1$, we obtain

$$\sqrt{\varpi/\pi} = 2^{-1/4}(1+2e^{-\pi}+2e^{-4\pi}+2e^{-9\pi}+\dots).$$

Also, if we set $\psi = 1/2$ and use the nontrivial fact that $P(\varpi/2) = Q(\varpi/2) = 2^{-1/4}$ (this is a consequence of the formula $Q(2\phi) = P(\phi)^4 + Q(\phi)^4$ – see (3.10)), we obtain

$$\sqrt{\varpi/\pi} = 2(e^{-\pi/4}+e^{-9\pi/4}+e^{-25\pi/4}+\dots)$$

(3.14)

$$\sqrt{\varpi/\pi} = 1 - 2e^{-\pi} + 2e^{-4\pi} - 2e^{-9\pi} + \dots .$$

Gauss wrote down these last two formulas in October 1798 (see [12, III, p. 418]). We, on the other hand, derived the first and third formulas as (2.21) in § 2, only after a very long development. Thus Gauss had some strong signposts to guide his development of modular functions.

These results, all dating from 1798, were recorded in Gauss' mathematical diary as the 91st and 92nd entries (in July) and the 95th entry (in October). The statement of the 92nd entry is especially relevant: "I have obtained most elegant results concerning the lemniscate, which surpasses all expectation—indeed, by methods which open an entirely new field to us" (see [12, X.1, p. 535]). There is a real sense of excitement here; instead of the earlier "advancement of analysis" of the 63rd entry, we have the much stronger phrase "entirely new field." Gauss knew that he had found something of importance. This feeling of excitement is confirmed by the

95th entry: "A new field of analysis is open before us, that is, the investigation of functions, etc." (see [12, X.1, p. 536]). It's as if Gauss were so enraptured he didn't even bother to finish the sentence.

More importantly, this "new field of analysis" is clearly the same "entirely new field of analysis" which we first saw in § 1 in the 98th entry. Rather than being an isolated phenomenon, it was the culmination of years of work. Imagine Gauss' excitement on May 30, 1799: this new field which he had seen grow up around the lemniscate and reveal such riches, all of a sudden expands yet again to encompass the arithmetic-geometric mean, a subject he had known since age 14. All of the powerful analytic tools he had developed for the lemniscatic functions were now ready to be applied to the agM.

C. In studying Gauss' work on the agM, it makes sense to start by asking where the observation $M(\sqrt{2}, 1) = \pi/\varpi$ came from. Using what we have learned so far, part of this question can now be answered: Gauss was very familiar with π/ϖ, and from reading Stirling he had probably seen the ratio $\sqrt{2}:1$ associated with the lemniscate. (In fact, this ratio appears in most known methods for constructing the lemniscate—see [24, pp. 111-117].) We have also seen, in the equation $N(\varpi) = e^{\pi/2}$, that Gauss often used numerical calculations to help him discover theorems. But while these facts are enlightening, they still leave out one key ingredient, the idea of taking the agM of $\sqrt{2}$ and 1. Where did this come from? The answer is that every great mathematical discovery is kindled by some intuitive spark, and in our case, the spark came on May 30, 1799 when Gauss decided to compute $M(\sqrt{2}, 1)$.

We are still missing one piece of our picture of Gauss at this time: how much did he know about the agM? Unfortunately, this is a very difficult question to answer. Only a few scattered fragments dealing with the agM can be dated before May 30, 1799 (see [12, X.1, pp. 172-173 and 260]). As for the date 1791 of his discovery of the agM, it comes from a letter he wrote in 1816 (see [12, X.1, p. 247]), and Gauss is known to have been sometimes wrong in his recollections of dates. The only other knowledge we have about the agM in this period is an oral tradition which holds that Gauss knew the relation between theta functions and the agM in 1794 (see [12, III, 493]). We will soon see that this claim is not as outrageous as one might suspect.

It is not our intention to give a complete account of Gauss' work on the agM. This material is well covered in other places (see [10], [12, X.2,

pp. 62-114], [13], [14] and [25]—the middle three references are especially complete), and furthermore it is impossible to give the full story of what happened. To explain this last statement, consider the following formulas:

$$B + (1/4)B^3 + (9/64)B^5 + \ldots = (2z^{1/2} + 2z^{9/2} + \ldots)^2 = r^2,$$

(3.15)

$$\frac{a}{M(a, b)} = 1 + (1/4)B^2 + (9/64)B^4 + \ldots,$$

where $B = (1-(b/a)^2)^{1/2}$. These come from the first surviving notes on the agM that Gauss wrote after May 30, 1799 (see [12, X.1, pp. 177-178]). If we set $a = 1$ and $b = k' = \sqrt{1-k^2}$, then $B = k$, and we obtain

$$\frac{1}{M(1, k')} = 1 + (1/4)k^2 + (9/64)k^4 + \ldots$$

(3.16)

$$\frac{k}{M(1, k')} = (2z^{1/2} + 2z^{9/2} + \ldots)^2 = r^2.$$

The first formula is (1.8), and the second, with $z = e^{\pi i\tau/2}$, follows easily from what we learned in § 2 about theta functions and the agM. Yet the formulas (3.15) appear with neither proofs nor any hint of where they came from. The discussion at the end of § 1 sheds some light on the bottom formula of (3.15), but there is nothing to prepare us for the top one.

It is true that Gauss had a long-standing interest in theta functions, going back to when he first encountered Euler's wonderful formula

$$\sum_{n=-\infty}^{\infty} (-1)^n x^{(3n^2+n)/2} = \prod_{n=1}^{\infty} (1-x^n).$$

The right-hand side appears in a fragment dating from 1796 (see [12, X.1, p. 142]), and the 7th entry of his mathematical diary, also dated 1796, gives a continued fraction expansion for

$$1 - 2 + 8 - 64 + \ldots.$$

Then the 58th entry, dated February 1797, generalizes this to give a continued fraction expansion for

$$1 - a + a^3 - a^6 + a^{10} - \ldots$$

(see [12, X.1, pp. 490 and 513]). The connection between these series and lemniscatic functions came in October 1798 with formulas such as (3.14).

This seems to have piqued his interest in the subject, for at this time he also set himself the problem of expressing

(3.17)
$$1 + x + x^3 + x^6 + x^{10} + \ldots$$

as an infinite product (see [12, X.1, p. 538]). Note also that the first formula of (3.14) gives r with $z = e^{-\pi/2}$.

Given these examples, we can conjecture where (3.15) came from. Gauss could easily have defined p, q and r in general and then derived identities (2.8)-(2.9) (recall the many identities obtained in 1798 for $P(\phi)$ and $Q(\phi)$—see (3.10) and [12, III, p. 410]). Then (3.15) would result from noticing that these identities formally satisfy the agM algorithm, which is the basic content of Lemma 2.3. This conjecture is consistent with the way Gauss initially treated z as a purely formal variable (the interpretation $z = e^{-\pi i\tau/2}$ was only to come later—see [12, X.1, pp. 262-263 and X.2, pp. 65-66]).

The lack of evidence makes it impossible to verify this or any other reasonable conjecture. But one thing is now clear: in Gauss' observation of May 30, 1799, we have not two but three distinct streams of his thought coming together. Soon after (or simultaneous with) observing that $M(\sqrt{2}, 1) = \pi/\varpi$, Gauss knew that there were inimate connections between lemniscatic functions, the agM, and theta functions. The richness of the mathematics we have seen is in large part due to the many-sided nature of this confluence.

There remain two items of unfinished business. From § 1, we want to determine more precisely when Gauss first proved Theorem 1.1. And recall from § 2 that on June 3, 1800, Gauss discovered the "mutual connection" among the infinitely many values of $M(a, b)$. We want to see if he really knew the bulk of § 2 by this date. To answer these questions, we will briefly examine the main notebook Gauss kept between November 1799 and July 1800 (the notebook is "Scheda Ac" and appears as pp. 184-206 in [12, X.1]).

The starting date of this notebook coincides with the 100th entry of Gauss' mathematical diary, which reads "We have uncovered many new things about arithmetic-geometric means" (see [12, X.1, p. 544]). After several pages dealing with geometry, one all of a sudden finds the formula (3.11) for ϖ/π. Since Gauss knew (3.15) at this time, we get an immediate proof of $M(\sqrt{2}, 1) = \pi/\varpi$. Gauss must have had this in mind, for otherwise why would he so carefully recopy a formula proved in July 1798? Yet one could also ask why such a step is necessary: isn't Theorem 1.1 an immediate consequence of (3.15)? Amazingly enough, it appears that Gauss wasn't yet

aware of this connection (see [12, X.1, p. 262]). Part of the problem is that he had been distracted by the power series, closely related to (3.15), which gives the arc length of the ellipse (see [12, X.1, p. 177]). This distraction was actually a bonus, for an asymptotic formula of Euler's for the arc length of the ellipse led Gauss to write

$$M(x, 1) = \frac{(\pi/2)\,(x-\alpha x^{-1}-\beta x^{-3}-\ldots)}{\log(1/z)} \tag{3.18}$$

where $x = k^{-1}$, and z and k are as in (3.16) (see [12, X.1, pp. 186 and 268-270]). He was then able to show that the power series on top was $(k\,M(1, k'))^{-1}$, which implies that

$$z = \exp\left(-\frac{\pi}{2}\cdot\frac{M(1, k')}{M(1, k)}\right)$$

(see [12, X.1, pp. 187 and 190]). Letting $z = e^{\pi i\tau/2}$, we obtain formulas similar to (2.20). More importantly, we see that Gauss is now in a position to uniformize the agM; z is no longer a purely formal variable.

In the process of studying (3.18), Gauss also saw the relation between the agM and complete elliptic integrals of the first kind. The formula

$$\frac{1}{M(1, k')} = \frac{2}{\pi}\int_0^1 \left((1-x^2)\,(1-k^2x^2)\right)^{-1/2}dx$$

follows easily from [12, X.1, p. 187], and this is trivially equivalent to (1.7). Furthermore, we know that this page was written on December 14, 1799 since on this date Gauss wrote in his mathematical diary that the agM was the quotient of two transcendental functions (see (3.18)), one of which was itself an integral quantity (see the 101st entry, [12, X.1, 544]). Thus Theorem 1.1 was proved on December 14, 1799, nine days earlier than our previous estimate.

Having proved this theorem, Gauss immediately notes one of its corollaries, that the "constant term" of the expression $(1+\mu\cos^2\phi)^{-1/2}$ is $M(\sqrt{1+\mu}, 1)^{-1}$ (see [12, X.1, p. 188]). By "constant term" Gauss means the coefficient A in the Fourier expansion

$$(1+\mu\cos^2\phi)^{-1/2} = A + A'\cos\phi + A''\cos 2\phi + \ldots .$$

Since A is the integral $\frac{2}{\pi}\int_0^{\pi/2}(1+\mu\cos^2\phi)^{-1/2}d\phi$, the desired result follows from Theorem 1.1. This interpretation is important because these coefficients

are useful in studying secular perturbations in astronomy (see [12, X.1, pp. 237-242]). It was in this connection that Gauss published his 1818 paper [12, III, pp. 331-355] from which we got our proof of Theorem 1.1.

What Gauss did next is unexpected: he used the agM to generalize the lemniscate functions to arbitrary elliptic functions, which for him meant inverse functions of elliptic integrals of the form

$$\int(1+\mu^2\sin^2\phi)^{-1/2}d\phi = \int\left((1-x^2)\,(1+\mu^2x^2)\right)^{-1/2}dx\,.$$

Note that $\mu = 1$ corresponds to the lemniscate. To start, he first set $\mu = \tan\nu$,

$$\varpi = \frac{\pi\cos\nu}{M(1, \cos\nu)}, \qquad \varpi' = \frac{\pi\cos\nu}{M(1, \sin\nu)}$$

and finally

$$z = \exp\left(-\frac{\pi}{2}\cdot\frac{\varpi'}{\varpi}\right) = \exp\left(-\frac{\pi}{2}\cdot\frac{M(1, \cos\nu)}{M(1, \sin\nu)}\right). \tag{3.19}$$

Then he defined the elliptic function $S(\phi)$ by $S(\phi) = \dfrac{T(\phi)}{W(\phi)}$ where

$$T(\psi\varpi) = 2\mu^{-1/2}\sqrt{M(1, \cos\nu)}\,(z^{1/2}\sin\psi\pi - z^{9/2}\sin 3\psi\pi + \ldots)$$

$$W(\psi\varpi) = \sqrt{M(1, \cos\nu)}\,(1 + 2z^2\cos 2\psi\pi + 2z^8\cos 4\psi\pi + \ldots) \tag{3.20}$$

(see [12, X.1, pp. 194-195 and 198]). In the pages that follow, we find the periods 2ϖ and $2i\varpi'$, the addition formula, and the differential equation connecting $S(\phi)$ to the above elliptic integral. Thus Gauss had a complete theory of elliptic functions.

In general, there are two basic approaches to this subject. One involves direct inversion of the elliptic integral and requires a detailed knowledge of the associated Riemann surface (see [17, Ch. VII]). The other more common approach defines elliptic functions as certain series ($\mathfrak{P}$-functions) or quotients of series (theta functions). The difficulty is proving that such functions invert all elliptic integrals. Classically, this uniformization problem is solved by studying a function such as $k(\tau)^2$ (see [36, § 20.6 and § 21.73]) or $j(\tau)$ (as in most modern texts—see [30, § 4.2]). Gauss uses the agM to solve this problem: (3.19) gives the desired uniformizing parameter! (In this connection,

the reader should reconsider the from [12, VIII, p. 101] given near the end of § 2.)

For us, the most interesting aspect of what Gauss did concerns the functions T and W. Notice that (3.20) is a direct generalization of (3.12); in fact, in terms of (3.13), we have

$$T(\psi\varpi) = \mu^{-1/2}\sqrt{M(1,\cos\nu)}\,\Theta_1(\psi\pi, z^2),$$

$$W(\psi\varpi) = \sqrt{M(1,\cos\nu)}\,\Theta_3(\psi\pi, z^2).$$

Gauss also introduces $T(\varpi/2-\phi)$ and $W(\varpi/2-\phi)$, which are related to the theta functions Θ_2 and Θ_4 by similar formulas (see [12, X.1, pp. 196 and 275]). He then studies the squares of these functions and he obtains identities such as

$$2\Theta_3(0, z^4)\,\Theta_3(2\phi, z^4) = \Theta_3(\phi, z^2)^2 + \Theta_4(\phi, z^2)^2$$

(this, of course, is the modern formulation—see [12, X.1, pp. 196 (Eq. 14) and 275]). When $\phi = 0$, this reduces to the first formula

$$p(\tau)^2 + q(\tau)^2 = 2p(2\tau)^2$$

of (2.8). The other formulas of (2.8) appear similarly. Gauss also obtained product expansions for the theta functions (see [12, X.1, pp. 201-205]). In particular, one finds all the formulas of (2.6). These manipulations yielded the further result that

$$1 + z + z^3 + z^6 z^{10} + \ldots = \prod_{n=1}^{\infty}(1-z)^{-1}(1-z^2),$$

solving the problem he had posed a year earlier in (3.17).

From Gauss' mathematical diary, we see that the bulk of this work was done in May 1800 (see entries 105, 106 and 108 in [12, X.1, pp. 546-549]). The last two weeks were especially intense as Gauss realized the special role played by the agM. The 108th entry, dated June 3, 1800, announces completion of a general theory of elliptic functions ("sinus lemniscatici universalissime accepti"). On the same day he recorded his discovery of the "mutual connection" among the values of the agM!

This is rather surprising. We've seen that Gauss knew the basic identities (2.6), (2.8) and (2.9), but the formulas (2.7), which tell us how theta functions behave under linear fractional transformations, are nowhere to be seen, nor do we find any hint of the fundamental domains used in § 2. Reading this notebook makes it clear that Gauss now knew the basic observation of

Lemma 2.3 that theta functions satisfy the agM algorithm, but there is no way to get from here to Theorem 2.2 without knowing (2.7). It is not until 1805 that this material appears in Gauss' notes (see [12, X.2, pp. 101-103]). Thus some authors, notably Markushevitch [25], have concluded that on June 3, 1800, Gauss had nothing approaching a proof of Theorem 2.2.

Schlesinger, the last editor of Gauss' collected works, feels otherwise. He thinks that Gauss knew (2.7) at this time, though knowledge of the fundamental domains may have not come until 1805 (see [12, X.2, p. 106]). Schlesinger often overestimates what Gauss knew about modular functions, but in this case I agree with him. As evidence, consider pp. 287-307 in [12, X.1]. These reproduce twelve consecutive pages from a notebook written in 1808 (see [12, X.1, p. 322]), and they contain the formulas (2.7), a clear statement of the basic observation of Lemma 2.3, the infinite product manipulations described above, and the equations giving the division of the agM into 3, 5 and 7 parts (in analogy with the division of the lemniscate). The last item is especially interesting because it relates to the second half of the 108th entry: "Moreover, in these same days, we have discovered the principles according to which the agM series should be interpolated, so as to enable us to exhibit by algebraic equations the terms in a given progression pertaining to any rational index" (see [12, X.1, p. 548]). There is no other record of this in 1800, yet here it is in 1808 resurfacing with other material (the infinite products) dating from 1800. Thus it is reasonable to assume that the rest of this material, including (2.7), also dates from 1800. Of course, to really check this conjecture, one would have to study the original documents in detail.

Given all of (2.6)-(2.9), it is still not clear where Gauss got the basic insight that $M(a, b)$ is a multiple valued function. One possible source of inspiration is the differential equation (1.12) whose solution (1.13) suggests linear combinations similar to those of Theorem 2.2. We get even closer to this theorem when we consiser the periods of $S(\phi)$:

$$m\varpi + in\varpi' = \pi\cos\nu\left(\frac{m}{M(1,\cos\nu)} + \frac{in}{M(1,\sin\nu)}\right)$$

where m, n are even integers. Gauss' struggles during May 1800 to understand the imaginary nature of these periods (see [12, X.2, pp. 70-71]) may have influenced his work on the agM. (We should point out that the above comments are related: Theorem 2.2 can be proved by analyzing the monodromy group—$\Gamma_2(4)$ in this case—of the differential equation (1.12).) On the other hand, Geppert suggests that Gauss may have taken a completely different

route, involving the asymptotic formula (3.18), of arriving at Theorem 2.2 (see [14, pp. 173-175]). We will of course never really know how Gauss arrived at this theorem.

For many years, Gauss hoped to write up these results for publication. He mentions this in Disquisitiones Arithmeticae (see [11, § 335]) and in the research announcement to his 1818 article (see [12, III, p. 358]). Two manuscripts written in 1800 (one on the agM, the other on lemniscatic functions) show that Gauss made a good start on this project (see [12, III, pp. 360-371 and 413-415]). He also periodically returned to earlier work and rewrote it in more complete form (the 1808 notebook is an example of this). Aside from the many other projects Gauss had to distract him, it is clear why he never finished this one: it was simply too big. Given his predilection for completeness, the resulting work would have been enormous. Gauss finally gave up trying in 1827 when the first works of Abel and Jacobi appeared. As he wrote in 1828, "I shall most likely not soon prepare my investigations on transcendental functions which I have had for many years—since 1798—because I have many other matters which must be cleared up. Herr Abel has now, I see, anticipated me and relieved me of the burden in regard to one third of these matters, particularly since he carried out all developments with great concision and elegance" (see [12, X.1, p. 248]).

The other two thirds "of these matters" encompass Gauss' work on the agM and modular functions. The latter were studied vigorously in the nineteenth century and are still an active area of research today. The agM, on the other hand, has been relegated to the history books. This is not entirely wrong, for the history of this subject is wonderful. But at the same time the agM is also wonderful as mathematics, and this mathematics deserves to be better known.

REFERENCES

[1] ALLING, N. L. *Real Elliptic Curves*. North-Holland Mathematics Studies, Vol. 54, North-Holland, Amsterdam, 1981.

[2] BERNOULLI, Jacob. *Opera*, Vol. I. Geneva, 1744.

[3] BERNOULLI, Johann. *Opera omnia*, Vol. I. Lausanne, 1742.

[4] BÜHLER, W. K. *Gauss: A Biographical Study*. Springer-Verlag, Berlin-Heidelberg-New York, 1981.

[5] CARLSON, B. C. Algorithms involving Arithmetic and Geometric Means. *Amer. Math. Monthly 78* (1971), 496-505.

[6] CASSELS, J. W. S. *Rational Quadratic Forms*. Academic Press, New York, 1978.

[7] COPSON, E. T. *An Introduction to the Theory of Functions of a Complex Variable*. Oxford U. Press, London, 1935.

[8] ENNEPER, A. *Elliptische Functionen: Theorie und Geschichte*. Halle, 1876.

[9] EULER, L. *Opera Omnia*, Series Prima, Vol. XX and XXI. Teubner, Leipzig and Berlin, 1912-1913.

[10] FUCHS, W. Das arithmetisch-geometrische Mittel in den Untersuchungen von Carl Friedrich Gauss. *Gauss-Gesellschaft Göttingen, Mittelungen No. 9* (1972), 14-38.

[11] GAUSS, C. F. *Disquisitiones Arithmeticae*. Translated by A. Clark, Yale U. Press, New Haven, 1965 (see also [12, I]).

[12] —— *Werke*. Göttingen-Leipzig, 1868-1927.

[13] GEPPERT, H. *Bestimmung der Anziehung eines elliptischen Ringes*. Ostwald's Klassiker, Vol. 225, Akademische Verlag, Leipzig, 1927.

[14] —— Wie Gauss zur elliptischen Modulfunktion kam. *Deutsche Mathematik 5* (1940), 158-175.

[15] —— Zur Theorie des arithmetisch-geometrischen Mittels. *Math. Annalen 99* (1928), 162-180.

[16] GRADSHTEYN, I. S. and I. M. RYZHIK. *Table of Integrals, Series and Products*. Academic Press, New York, 1965.

[17] HANCOCK, H. *Lectures on the Theory of Elliptic Functions*. Vol. I. Wiley, New York, 1910.

[18] HOFFMAN, J. E. Über Jakob Bernoullis Beiträge zur Infinitesimalmathematik. *L'Enseignement Math. 2* (1956), 61-171.

[19] HOUZEL, C. Fonctions Elliptiques et Intégrals Abéliennes. In *Abrégé d'histoire des mathématiques 1700-1900*, Vol. II. Ed. by J. Dieudonné, Hermann, Paris, 1978, 1-112.

[20] JACOBI, C. C. J. *Gesammelte Werke*. G. Reimer, Berlin, 1881.

[21] KLINE, M. *Mathematical Thought from Ancient to Modern Times*. Oxford U. Press, New York, 1972.

[22] LAGRANGE, J. L. *Œuvres*, Vol. II. Gauthier-Villars, Paris, 1868.

[23] LEGENDRE, A. M. *Traité des Fonctions Elliptiques*. Paris, 1825-1828.

[24] LOCKWOOD, E. H. *A Book of Curves*. Cambridge U. Press, Cambridge, 1971.

[25] MARKUSHEVITCH, A. I. Die Arbeiten von C. F. Gauss über Funktionentheorie. In *C. F. Gauss Gedenkband Anlässlich des 100. Todestages am 23. Februar 1955*. Ed. by H. Reichart, Teubner, Leipzig, 1957, 151-182.

[26] MIEL, G. Of Calculations Past and Present: The Archimedean Algorithm. *Amer. Math. Monthly 90* (1983), 17-35.

[27] MUMFORD, D. *Tata Lectures on Theta I*. Progress in Mathematics Vol. 28, Birkhäuser, Boston, 1983.

[28] ROSEN, M. Abel's Theorem on the Lemniscate. *Amer. Math. Monthly 88* (1981), 387-395.
[29] SERRE, J.-P. *Cours d'Arithmétique.* Presses U. de France, Paris, 1970.
[30] SHIMURA, G. *Introduction to the Arithmetic Theory of Automorphic Functions.* Princeton U. Press, Princeton, 1971.
[31] STIRLING, J. *Methodus Differentialis.* London, 1730.
[32] TANNERY, J. and J. MOLK. *Eléments de la Théorie des Fonctions Elliptiques,* Vol. 2. Gauthiers-Villars, Paris, 1893.
[33] TODD, J. The Lemniscate Constants. *Comm. of the ACM 18* (1975), 14-19.
[34] van der POL, B. Démonstration Elémentaire de la Relation $\Theta_3^4 = \Theta_0^4 + \Theta_2^4$ entre les Différentes Fonctions de Jacobi. *L'Enseignement Math. 1* (1955), 258-261.
[35] von DAVID, L. Arithmetisch-geometrisches Mittel und Modulfunktion. *J. für die Reine u. Ang. Math. 159* (1928), 154-170.
[36] WHITTAKER, E. T. and G. N. WATSON. *A Course of Modern Analysis,* 4th ed. Cambridge U. Press, Cambridge, 1963.

(Reçu le 21 novembre 1983)

David A. Cox
Department of Mathematics
Amherst College
Amherst, MA 01002 (USA)

SIAM REVIEW
Vol. 26, No. 3, July 1984

THE ARITHMETIC-GEOMETRIC MEAN AND FAST COMPUTATION OF ELEMENTARY FUNCTIONS*

J. M. BORWEIN† AND P. B. BORWEIN†

Abstract. We produce a self contained account of the relationship between the Gaussian arithmetic-geometric mean iteration and the fast computation of elementary functions. A particularly pleasant algorithm for π is one of the by-products.

Introduction. It is possible to calculate 2^n decimal places of π using only n iterations of a (fairly) simple three-term recursion. This remarkable fact seems to have first been explicitly noted by Salamin in 1976 [16]. Recently the Japanese workers Y. Tamura and Y. Kanada have used Salamin's algorithm to calculate π to 2^{23} decimal places in 6.8 hours. Subsequently 2^{24} places were obtained ([18] and private communication). Even more remarkable is the fact that all the elementary functions can be calculated with similar dispatch. This was proved (and implemented) by Brent in 1976 [5]. These extraordinarily rapid algorithms rely on a body of material from the theory of elliptic functions, all of which was known to Gauss. It is an interesting synthesis of classical mathematics with contemporary computational concerns that has provided us with these methods. Brent's analysis requires a number of results on elliptic functions that are no longer particularly familiar to most mathematicians. Newman in 1981 stripped this analysis to its bare essentials and derived related, though somewhat less computationally satisfactory, methods for computing π and log. This concise and attractive treatment may be found in [15].

Our intention is to provide a mathematically intermediate perspective and some bits of the history. We shall derive implementable (essentially) quadratic methods for computing π and all the elementary functions. The treatment is entirely self-contained and uses only a minimum of elliptic function theory.

1. 3.14159265358979323846264338327950288419 7. The calculation of π to great accuracy has had a mathematical import that goes far beyond the dictates of utility. It requires a mere 39 digits of π in order to compute the circumference of a circle of radius 2×10^{25} meters (an upper bound on the distance travelled by a particle moving at the speed of light for 20 billion years, and as such an upper bound on the radius of the universe) with an error of less than 10^{-12} meters (a lower bound for the radius of a hydrogen atom).

Such a calculation was in principle possible for Archimedes, who was the first person to develop methods capable of generating arbitrarily many digits of π. He considered circumscribed and inscribed regular n-gons in a circle of radius 1. Using $n = 96$ he obtained

$$3.1405\cdots = \frac{6336}{2017.25} < \pi < \frac{14688}{4673.5} = 3.1428.$$

If $1/A_n$ denotes the area of an inscribed regular 2^n-gon and $1/B_n$ denotes the area of a circumscribed regular 2^n-gon about a circle of radius 1 then

$$A_{n+1} = \sqrt{A_n B_n}, \qquad B_{n+1} = \frac{A_{n+1} + B_n}{2}. \tag{1.1}$$

*Received by the editors February 8, 1983, and in revised form November 21, 1983. This research was partially sponsored by the Natural Sciences and Engineering Research Council of Canada.
†Department of Mathematics, Dalhousie University, Halifax, Nova Scotia, Canada B3H 4H8.