Zeitschrift: L'Enseignement Mathématique
Band: 30 (1984)
Heft: 1-2: L'ENSEIGNEMENT MATHÉMATIQUE

Artikel: THE ARITHMETIC-GEOMETRIC MEAN OF GAUSS
Kapitel: 2. The arithmetic-geometric mean of complex numbers
Autor: Cox, David A.
DOI: https://doi.org/10.5169/seals-53831

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 If we turn to Gauss' unpublished manuscript of 1800 (where we got the example $M(\sqrt{2}, 1)$), we find (1.7) and (1.8) as expected, but also the observation that a complete solution of the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

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

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

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 $0 \leqslant \operatorname{ang}(a, b) \leqslant \pi$ denote the unoriented angle between a and b.

Then we have:

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

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

To prove (2.2), note that

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

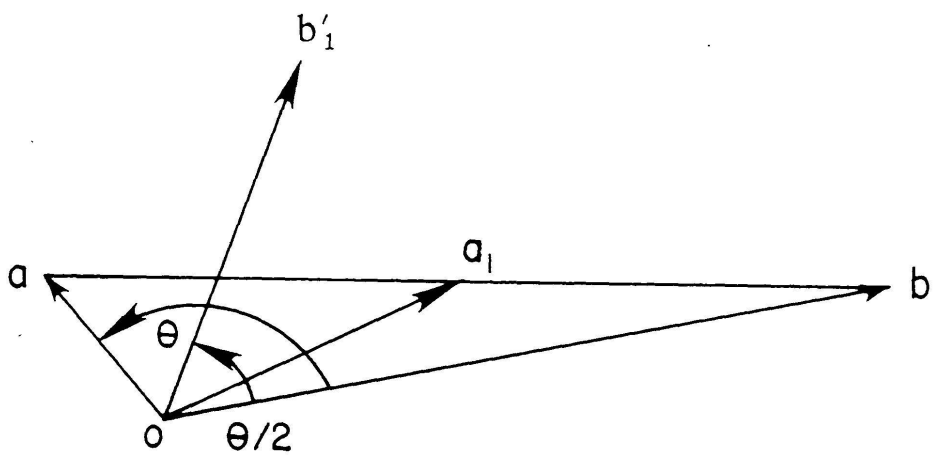
Since $|a_1 - b_1| \leqslant |a_1 + b_1|$, (2.2) follows immediately. To prove (2.3), let $\theta_1 = \operatorname{ang}(a_1, b_1)$ and $\theta = \operatorname{ang}(a, b)$. From the law of cosines

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

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

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

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

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

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 $\theta_0 = \text{ang}(a, b) < \pi$). 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 $\mathfrak{H} = \{\tau \in \mathbf{C} : \mathrm{Im}\,\tau > 0\}$ 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 $e^{-\pi t}$, $\mathrm{Re}\,t > 0$ (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 $\mathrm{Re}(-i\tau)^{1/2} > 0$), 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),*

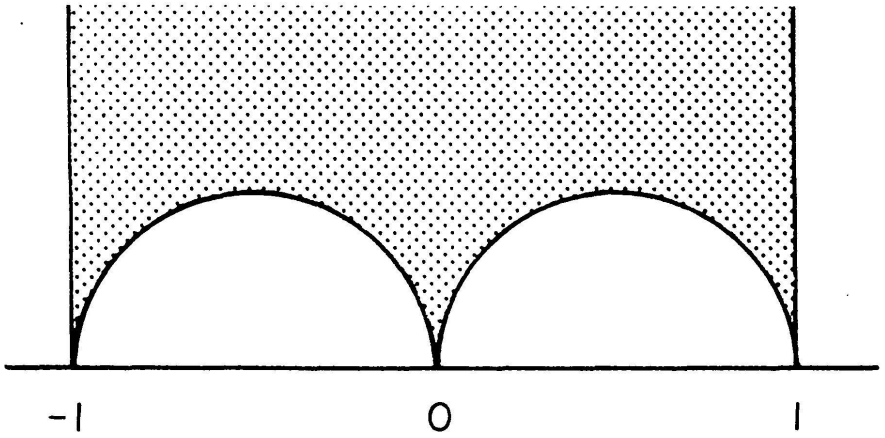
(ii) $\lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n = \mu$.

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 $F_1 \subseteq \mathfrak{H}$:

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

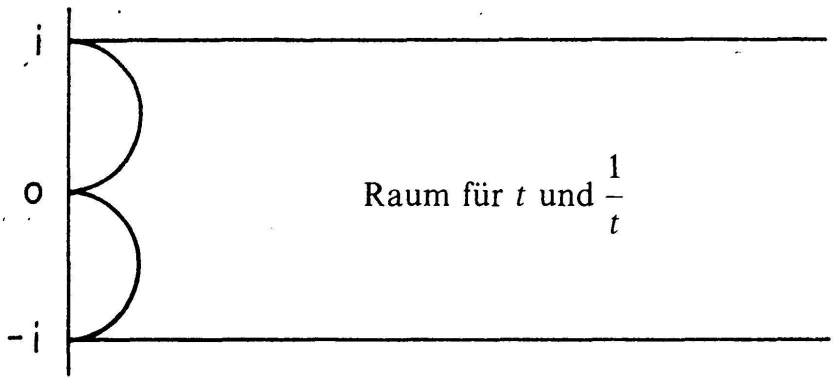


F_1 shaded

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

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

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

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

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

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

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

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

Thus $|\, 2\tau \pm 1 \,| \geqslant 1$, or $|\, \tau \pm (1/2) \,| \geqslant 1/2$. This is equivalent to $|\operatorname{Re} 1/\tau'| \leqslant 1$, 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

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

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

$$\begin{aligned} p(\gamma\tau)^2 &= (c\tau + d)\, p(\tau)^2 , \quad \gamma \in \Gamma(2)_0 , \\ q(\gamma\tau)^2 &= (c\tau + d)\, q(\tau)^2 , \quad \gamma \in \Gamma_2(4) . \end{aligned} \tag{2.12}$$

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.

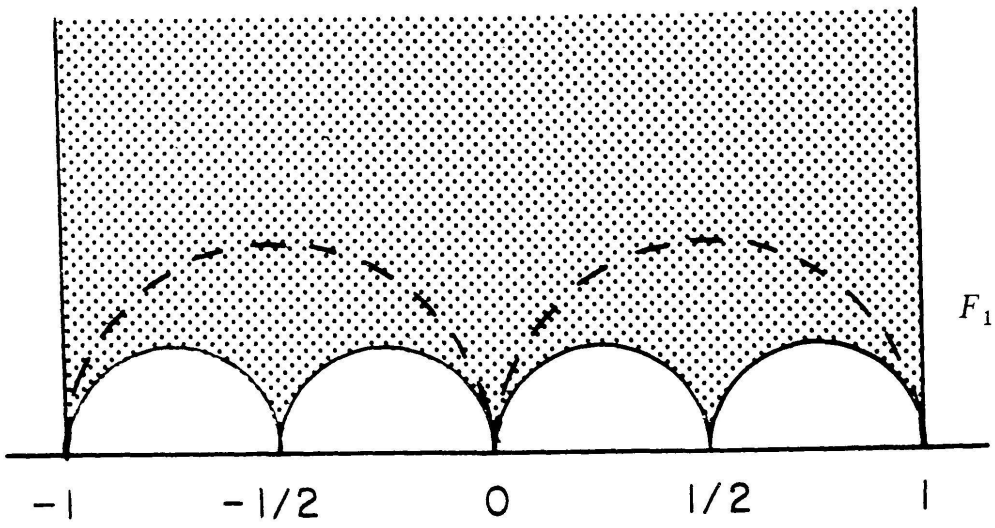
In general, it is well known that the square of a theta function is a modular form of weight one (see [27, Ch. I, § 9]), although the general theory only says that our functions are modular forms for the group

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

(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:

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



F shaded

F_1 above dashed lines

We also set

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

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

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*

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

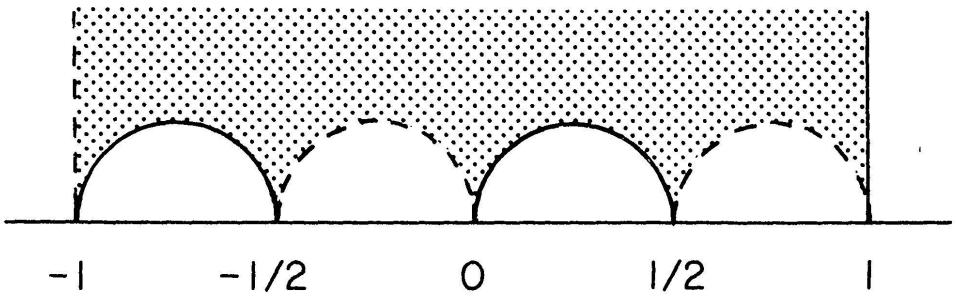
$$\overline{k'} : \mathfrak{H}/\Gamma_2(4) \tilde{\to} \mathbf{C} - \{0, \pm 1\}.$$

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

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



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:

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

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

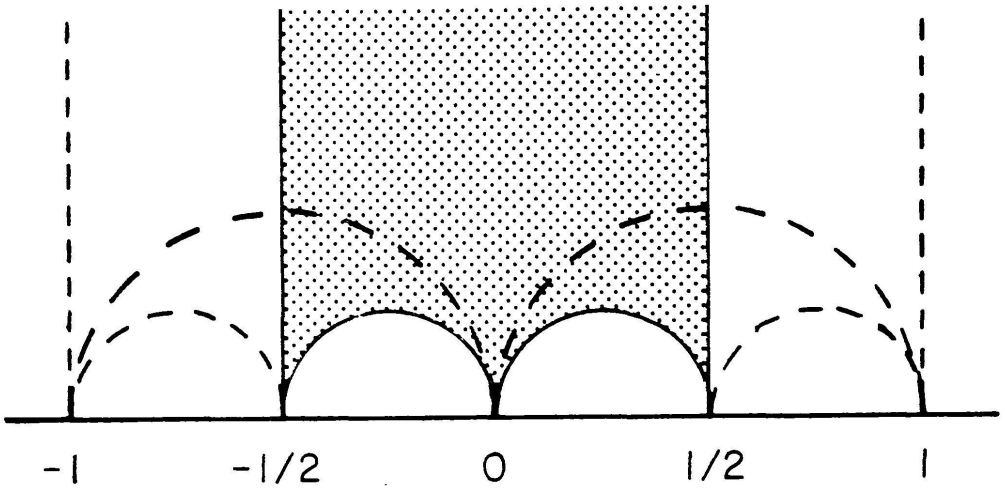
We should point out that $r(\tau)^2$ has properties similar to $p(\tau)^2$ and $q(\tau)^2$. Specifically, $r(\tau)^2$ is a modular form of weight one for the group

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

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 $\Gamma_2(4)^t$ and induces a biholomorphism $\overline{k} : \mathfrak{H}/\Gamma_2(4)^t$

$\to \mathbf{C} - \{0, \pm 1\}$. 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

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

(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

(2.15)

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

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

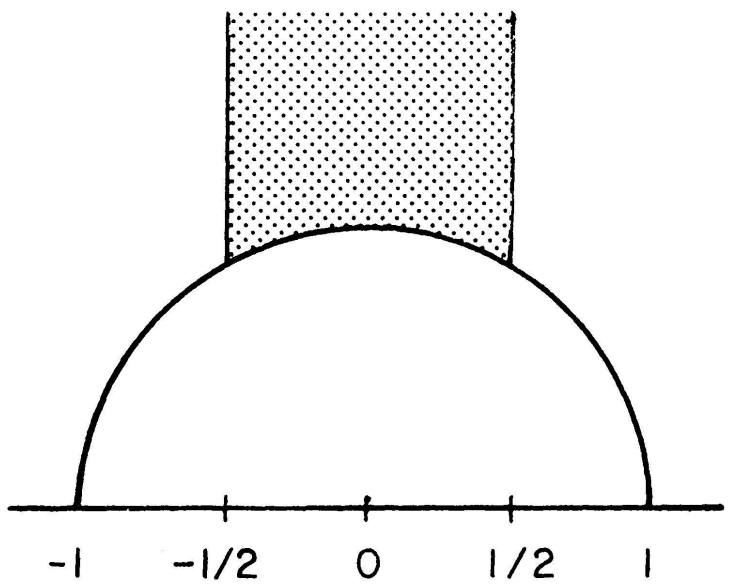
$$k'(\gamma\tau) = i^c k'(\tau)$$

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



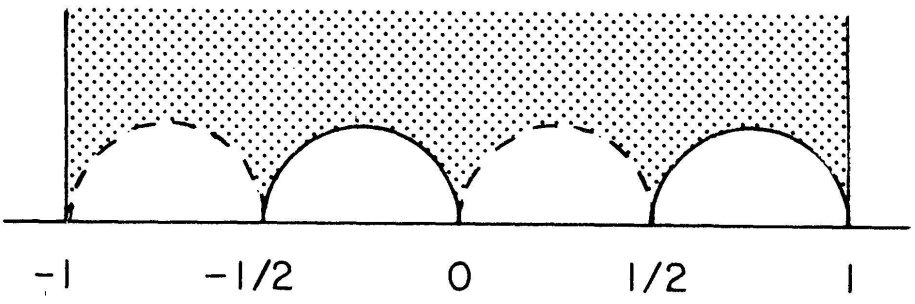
which is well known to be the fundamental domain of $SL(2, \mathbf{Z})$ acting on $\mathfrak{H}$ (see [29, Ch. VII, Thm. 1]).

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 $F^\wedge = \{\tau \in F : |\tau - 1/4| > 1/4, |\tau + 3/4| > 1/4\}$. 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 $n \geq 0$.

The following equivalences are very easy to prove:

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

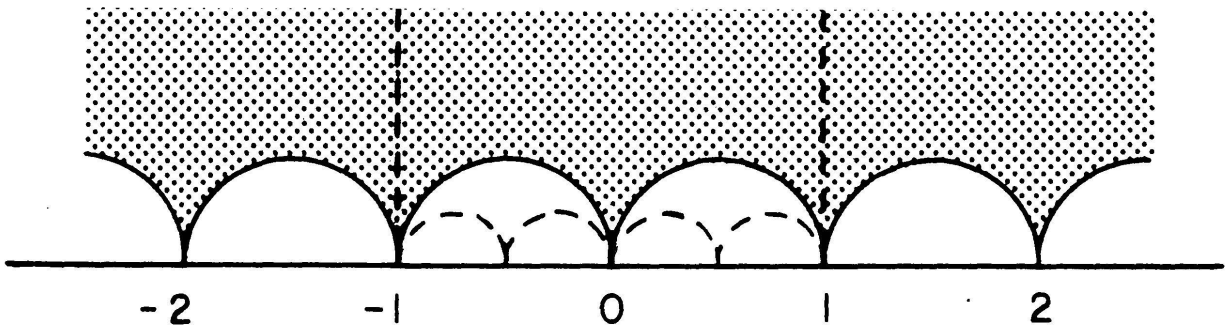
$$|a_{n+1} - b_{n+1}| = |a_{n+1} + b_{n+1}| \Leftrightarrow \operatorname{Re}\left(\frac{b_{n+1}}{a_{n+1}}\right) = 0.$$

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.

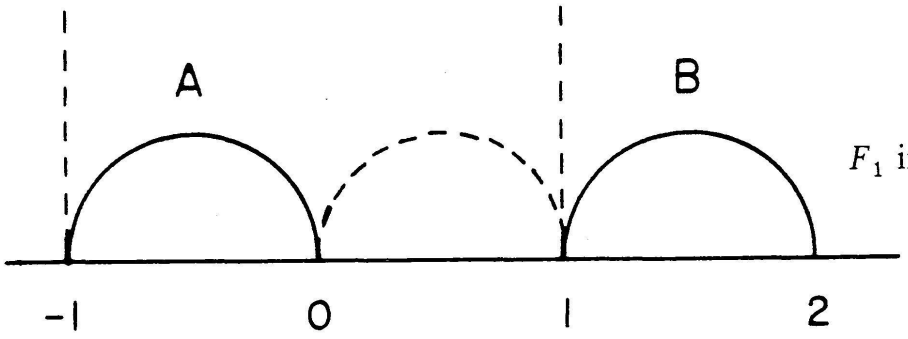


$\tilde{F}_1$ shaded
F indicated by dashed lines

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 $\tau \in F^{\wedge}$ 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 $\mathrm{Im}(k'(2\tau)) > 0$ for $2\tau \in A$. Since $S = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ maps the line $\mathrm{Re}\sigma = 1$ to A, we can write $2\tau = -1/\sigma$, where $\mathrm{Re}\sigma = 1$. 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 $\mathrm{Re}\sigma = 1$, the product expansions (2.6) easily show that

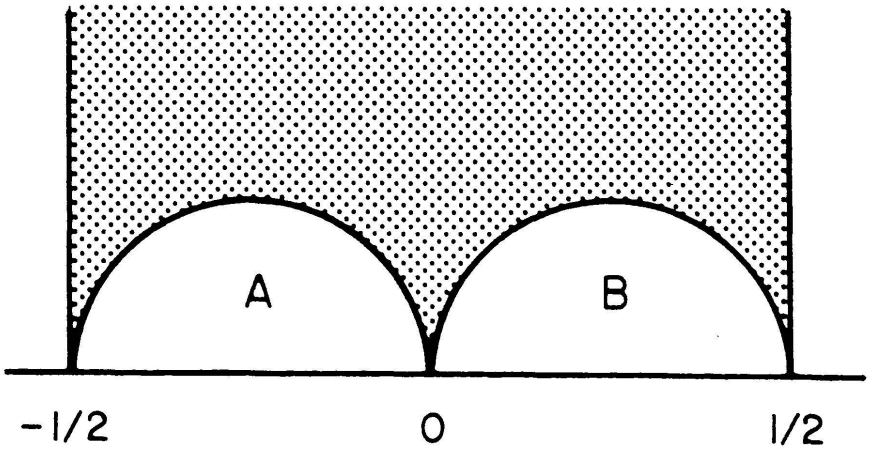
$$\mathrm{Im}(r(\sigma)^2/p(\sigma)^2) > 0,$$

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

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

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

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

(2.21)

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

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 $\int_0^{\pi/2} (1-k^2\sin^2\phi)^{-1/2}d\phi$. 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 $\text{ang}(a, b) \neq \pi$. 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

(2.22)

$$\mu' = \mu/(c\tau_0+d) ,$$

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 α, β, γ, δ signify all integers which satisfy the equation $\alpha\delta - 4\beta\gamma = 1$." Recall that 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:

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

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.