Zeitschrift: L'Enseignement Mathématique

Herausgeber: Commission Internationale de l'Enseignement Mathématique

Band: 30 (1984)

Heft: 1-2: L'ENSEIGNEMENT MATHÉMATIQUE

Artikel: THE ARITHMETIC-GEOMETRIC MEAN OF GAUSS

Autor: Cox, David A.

Kapitel: 1. The arithmetic-geometric mean of real numbers

DOI: https://doi.org/10.5169/seals-53831

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

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

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

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. $M(\sqrt{2}, 1) = 1.198140234735592207 4...$

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:

$$
\begin{aligned}
&M(a, b) = M(a_1, b_1) = M(a_2, b_2) = \ldots \\
(1.3)\qquad & \\
&M(\lambda a, \lambda b) = \lambda M(a, b)\,.
\end{aligned}
$$

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

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

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

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

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

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

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

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

(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

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

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

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

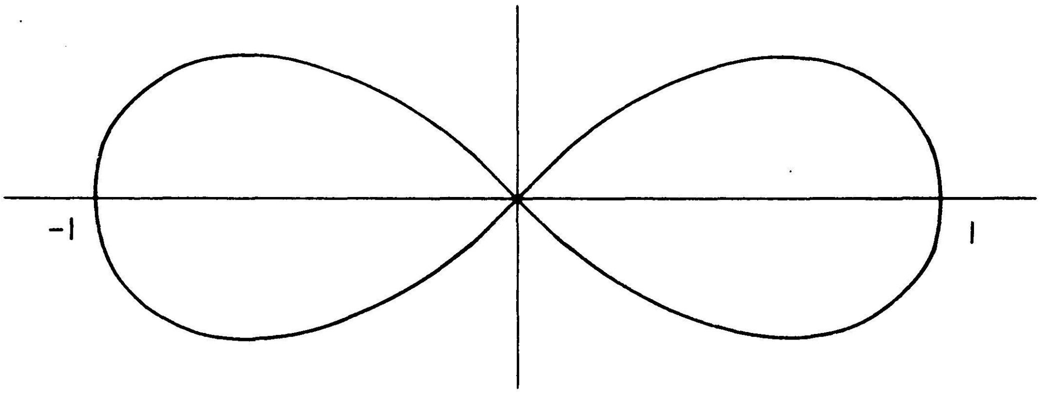
Finally, it is customary to set $k' = \sqrt{1-k^2}$. Then, using (1.3), we can rewrite (1.7) as

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

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

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

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

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

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

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

A comparison of the coefficients of powers of t gives an infinite system of equations in $A, B, C, \dots$. Gauss showed that this system is equivalent to the equations $0 = 1 - 4A = 9A - 16B = 25B - 36C = \dots$, 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

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

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