The Arithmetico-geometric Mean of Gauss

How to find the Perimeter of an Ellipse

B Sury

The author enjoys writing about such great masters as:

A person who does arouse
much admiration and a
million wows!
Such a mathematicians' prince
has not been born since.
I talk of Carl Friedrich Gauss!

Elementary Calculus

As all of us learn very early in school, the length of a circle is $\pi \times$ diameter. Somewhat later, one learns that elementary calculus gives the length of an arc of the unit circle to be the value of the integral $s = s(x) = \int_0^x \frac{dt}{\sqrt{1-t^2}} = \sin^{-1}(x)$.

Before going further, it is worthwhile to recall how length of (part of) a curve is defined. Suppose first that the curve is given on the plane and is described by the pair of equations $x = x(t), y = y(t)$ in terms of a parameter t. Intuitively, one would think of the length of any part of the curve to be that of parts of a polygon which increasingly approximates that part of the curve. Thus, one divides the interval $[a, b]$ where t varies into finitely many parts $a = t_0 < t_1 < \quad < t_n = b$. The points $P_i = (x(t_i), y(t_i))$ can be joined by line segments and the sum $P_0P_1 + P_1P_2 + \quad + P_{n-1}P_n$ of the lengths of these segments is an approximation to the length of the part of the curve which corresponds to the parameter t varying from a to b. This approximation might be a crude one but it is at once clear (see *Figure* 1) that one can take more points to obtain a better approximation. At this point, calculus comes to the rescue. It is heuristically clear that the best possible notion of length is obtained as the *limiting case* as one increases the number n of points P_i indefinitely while simultaneously letting the lengths of all the line segments P_iP_{i+1} diminish indefinitely. In other words, if the limit exists, it can be defined as the length of the curve. A curve for which this limit exists is called *rectifiable* for evident reasons. Clearly, circles, ellipses and hyperbolae are rectifiable curves.

Exercise *Show that the curve defined by*

$$y = x^2 \sin 1/x \quad 0 < x \leq 1 \quad y(0) = 0$$

is rectifiable while that defined by

$$y = x \sin 1/x \quad 0 < x \leq 1 \quad y(0) = 0$$

is not. What is the length of the first curve?

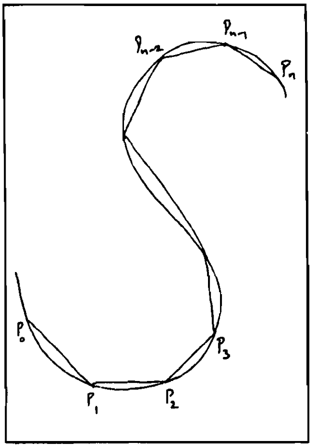


Figure 1.

Is there a nice condition which is sufficient to ensure that a given curve is rectifiable? Let us look at a curve with a parametrisation $x = x(t), y = y(t)$ for $a \leq t \leq b$ where these functions are assumed to have continuous first derivatives. This condition is indeed enough to enforce rectifiability as we show now. Let $a = t_0 < t_1 <$ $< t_n = b$ be a subdivision; then the approximating perimeter

$$S_n = \sum_{i=0}^{n-1} P_i P_{i+1} =$$

$$\sum_{i=0}^{n-1} \sqrt{[\{x(t_{i+1}) - x(t_i)\}^2 + \{y(t_{i+1}) - y(t_i)\}^2]}.$$

By the mean-value theorem, there are $a_i, b_i \in (t_i, t_{i+1})$ such that $x(t_{i+1}) - x(t_i) = \frac{dx}{dt}(a_i)(t_{i+1} - t_i)$ and $y(t_{i+1}) - y(t_i) = \frac{dy}{dt}(b_i)(t_{i+1} - t_i)$. When we let the number of intervals (t_i, t_{i+1}) tend to infinity while simultaneously letting their lengths tend to 0, the limit is, by definition, just the integral $\int_a^b \sqrt{(\frac{dx}{dt})^2 + (\frac{dy}{dt})^2 dt}$. We note that the above limit exists by uniform continuity of $\frac{dx}{dt}, \frac{dy}{dt}$ – consequence of the assumption that the parametrising functions $x = x(t), y = y(t)$ for $a \leq t \leq b$ are functions which have continuous first derivatives. Actually, one can even allow the functions $\frac{dx}{dt}$ and $\frac{dy}{dt}$ to have finitely many discontinuities and still the integral makes sense.

It is clear from the change of variable formula for integrals that the above notion of length is independent of which parametrisation we use. For this reason, if one is given the curve in nonparametric form as $y = f(x)$, then one can write its length from $x = a$ to $x = b$ to be the integral $\int_a^b \sqrt{1 + (\frac{dy}{dx})^2}dx$. If the curve is given in polar coordinates as $x = r(\theta)\cos\theta$, $y = r(\theta)\sin\theta$, the length can, therefore, be written as the integral $\int_{\theta_0}^{\theta_1} \sqrt{r^2 + (\frac{dr}{d\theta})^2}d\theta$ where $r(\theta_0)\cos\theta_0 = a, r(\theta_1)\cos\theta_1 = b$.

Let us look at the circle of radius r_0; its polar equation is $r = r_0$ (constant); the integral gives the length $(\theta_1 - \theta_0)r_0$.

Already, it might be clear that even if a sequence of functions f_n defined on some interval (a, b) converges in the interval to a function f, the corresponding integrals $\int_a^b f_n(x)dx$ need not converge to $\int_a^b f(x)dx$. The same problem persists with the integrals of their derivatives etc.

This, for instance, provides the correct explanation of one of the 'Think-it-over' problems: 'Where is the missing string?' posed in *Resonance*, Vol. 3, No. 1, 1998. The area under the string in its various positions can get arbitrarily close to that under the diagonal string without the arc lengths tending to the diagonal length.

Elliptic Integrals

After the brief digression to recall the fundamental concept of length of a curve, let us come back a whole circle viz., look at $x^2 + y^2 = 1$. The length $s = s(x_0)$ from $(x, y) = (0, 1)$ to (x_0, y_0) is given as $\int_0^{x_0} \frac{dx}{\sqrt{1-x^2}} = \sin^{-1}x_0$. We express this as $x = \sin(s)$ and, viewed this way, the addition formula for the sine function amounts to the formula

$$\int_0^x \frac{dt}{\sqrt{1-t^2}} + \int_0^y \frac{dt}{\sqrt{1-t^2}} = \int_0^z \frac{dt}{\sqrt{1-t^2}}$$

where $z = x\sqrt{1-y^2} + y\sqrt{1-x^2}$.

What happens when we try to calculate the length of an arc in an ellipse

$\frac{x^2}{a^2}+\frac{y^2}{b^2}=1$?

Such a length is given by an integral of the form

$$\begin{aligned} s(x_0) &= \int_0^{x_0} \sqrt{1+(\frac{dy}{dx})^2}dx \\ &= \int_0^{x_0} \frac{a^2-e^2x^2}{\sqrt{(a^2-e^2x^2)(a^2-x^2)}}dx \\ &= \int_0^{x_0} R(x,\sqrt{p(x)})dx \end{aligned}$$

where R is a rational function (i.e. a quotient of two polynomials) and p is a polynomial of degree 4 in x. Here $e=\frac{\sqrt{a^2-b^2}}{a}$ is the *eccentricity* of the ellipse.

Lemniscate can be described as the locus of points, the product of whose distances from two fixed points is constant.

Similarly, let us look at a *lemniscate* which can be described as *the locus of points, the product of whose distances from two fixed points is constant.* It is given in polar co-ordinates by the equation $r^2=\cos(2\theta)$, the arc length is given by the integral

$$\begin{aligned} s(\theta_0) &= \int_0^{\theta_0} \sqrt{(r^2+(\frac{dr}{d\theta})^2)}d\theta \\ &= \int_0^{\theta_0} \cos(2\theta)^{-1/2}d\theta \\ &= \int_0^{x_0} (1-x^4)^{-1/2}dx. \end{aligned}$$

The analogy to the circular length $\int_0^{x_0}(1-x^2)^{-1/2}dx$ is evident. For the circle, one evaluates the integral by *rationalising* the integrand i.e., by making a change of variables which makes the integrand into a rational function i.e., into a quotient of two polynomials. One can rationalise the integrand by the substitution $x=\frac{2t}{1+t^2}$ (How?). Guided by this analogy in 1718, Fagnano tried the substitution $x=\frac{2t^2}{1+t^4}$ for the lemniscate. Although the integrand does not rationalise, he noticed that $\int_0^{x_0}(1-x^4)^{-1/2}dx=\sqrt{2}\int_0^{t_0}(1+t^4)^{-1/2}dt$

if $0 \leq x \leq 1$. Following this up with the substitution $t = \frac{2u^2}{1-u^4}$, one obtains $\int_0^{x_0}(1-x^4)^{-1/2}dx = 2\int_0^{u_0}(1-u^4)^{-1/2}du$. Let us note the interesting consequence of this that using a ruler and compass, one can double the arc length of a lemniscate – see [1] for a more general discussion of this aspect.

Using a ruler and compass, one can double the arc length of a lemniscate.

A fundamental idea in the general problem of evaluating integrals of the form $\int_0^{x_0} R(x, \sqrt{p(x)})dx$ is due to Abel and Jacobi. This is simply to *invert the problem*! So, one considers x_0 as a function of s in the formula $s(x_0) = \int_0^{x_0} R(x, \sqrt{p(x)})dx$. Note that the integrand is a multi-valued function while however their inverse functions arise as single-valued functions as Abel and Jacobi realised in the 1820's. This yields addition formulae for these functions akin to those for the trigonometric functions. These integrals and, in general, the integrals of the form $s(x_0) = \int_0^{x_0} R(x, \sqrt{p})dx$ are called *elliptic integrals* where R is a rational function and p is a polynomial of degree 3 or 4 in x (why did we leave out degrees 1 and 2?). The inverse functions $x = x(s)$ are called *elliptic functions*. They have properties analogous to those for trigonometric functions; in place of the usual periodicity, they are doubly[1] periodic!

Although we can't prove it here, the fundamental fact involved here is that *any cubic curve over the complex numbers which is smooth i.e., has no singularities, is biholomorphic to a complex torus*. We digress now to discuss a notion first introduced by Gauss and studied in relation with elliptic integrals to get some beautiful results.

The agM

Let $a \geq b > 0$ be two real numbers. Construct the following sequence of arithmetic and geometric means:

$a_1 = \frac{a+b}{2}, b_1 = \sqrt{ab}, a_2 = \frac{a_1+b_1}{2}, b_2 = \sqrt{a_1 b_1}, \quad , a_{n+1} = \frac{a_n+b_n}{2}, b_{n+1} = \sqrt{a_n b_n},$ etc.

For example, look at $a = 2, b = 1$. We have the following sequence of values correct to eight decimal places:

n	a_n	b_n
0	2	1
1	1.5	1.41421356
2	1.45710678	1.45647531
3	1.45679105	1.45679101
4	1.45679103	1.45679103

The fundamental fact involved here is that any cubic curve over the complex numbers which is smooth is biholomorphic to a complex torus.

It is natural to guess that the two sequences approach a common number. This is indeed true in general and is seen quite easily as follows. From the familiar inequality which asserts that the arithmetic mean is at least as big as the geometric mean, one has $a_n \geq b_n$. In fact, $a \geq a_1 \geq a_2 \geq \quad \geq a_n \geq b_n \geq b_{n-1} \geq \quad \geq b_1 \geq b$. Can one say more?

Indeed, it is easy to see that any of the a's is at least as big as any of the b's (why?) i.e.,

$$a \geq a_1 \geq a_2 \geq \quad \geq a_n \geq \quad \geq b_m \geq b_{m-1} \geq$$

$$\geq b_1 \geq b.$$

Moreover, by mathematical induction, it follows that $a_n - b_n < \frac{a-b}{2^n}$ for all $n \geq 1$. What do these two observations give us? The first implies that the limits $\lim_{n\to\infty} a_n$ and $\lim_{n\to\infty} b_n$ exist. The second implies that these two limits are actually equal! This common limit, denoted by $M(a, b)$ is called the *arithmetico-geometric mean* and denoted by agM, a notation introduced by Gauss himself. Some of the evident properties of the agM are:

$$M(a, a) = a$$

$$M(a, b) = M(a_n, b_n) \ \ \forall \ n$$

$$M(ta, tb) = tM(a, b)$$

agM and the Perimeter

Gauss proved the following beautiful result on the perimeter of an ellipse:

Theorem

If $a \geq b > 0$, then

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

Gauss's proof is ingenious and goes as follows. If $I(a,b)$ denotes the integral above, the key step is to show that $I(a,b) = I(a_1,b_1)$ for then, repeating this, and going to the limit, one will get $I(a,b) = I(M(a,b), M(a,b)) = M(a,b)^{-1}\frac{\pi}{2}$, which is the assertion. The proof of the key step is achieved in the following manner by making the substitution

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

This is now called Gauss's transformation.

The change of variables gives easily that

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

Let us differentiate both sides of

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

to compute the integral in terms of the new variable ϕ. Derivative of the left hand side of (Δ) is

$$(a^2\cos^2\theta + b^2\sin^2\theta)^{-1/2}(b^2-a^2)\sin\theta\cos\theta d\theta.$$

Differentiating the right hand side of (Δ), one gets the expression

$$\frac{4a(b^2-a^2)\sin\phi\cos\phi}{(a+b+(a-b)\sin^2\phi)^2}d\phi.$$

Substituting for the denominator, this further simplifies to $4a(b^2 - a^2)\sin\phi\cos\phi\frac{\sin^2\theta}{4a^2\sin^2\phi}d\phi$ i.e., to the expression

$$\frac{(b^2 - a^2)\cos\phi\sin^2\theta\sin\phi}{a\sin\phi}d\phi$$

On the other hand, using $a + b = 2a_1$ and $ab = b_1^2$, one has

$$(a_1^2\cos^2\phi + b_1^2\sin^2\phi)\cos^2\phi = a_1^2 - (2b_1^2 - a_1^2)\sin^2\phi +$$

$$(a_1^2 - b_1^2)\sin^4\phi = \frac{(a+b)^2 - 2(a^2 + b^2)\sin^2\phi + (a-b)^2\sin^4\phi}{4}$$

$$= \frac{(a + b + (a-b)\sin^2\phi)^2 - 4a^2\sin^2\phi}{4} = \frac{a^2\cos^2\theta\sin^2\phi}{\sin^2\theta}.$$

Thus,

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

Mutiplying this with $(b^2 - a^2)\sin\theta\cos\theta d\phi$ yields just the derivative of the right hand side of (Δ). In other words, we have verified that

$$(a^2\cos^2\theta + b^2\sin^2\theta)^{-1/2}d\theta =$$

$$(a_1^2\cos^2\phi + b_1^2\sin^2\phi)^{-1/2}d\phi.$$

From this, the key step follows immediately and so does Gauss's theorem.

The agM $M(a, b)$ can be expressed in terms of one parameter, viz., the eccentricity $e = \sqrt{a^2 - b^2}/a$ of the ellipse. Indeed, $M(1 + e, 1 - e) = M(1, \sqrt{1 - e^2}) = M(1, \frac{b}{a}) = M(a, b)/a$. Therefore, the theorem can be restated as

$$M(1 + e, 1 - e) = \frac{\pi}{2K(e)}$$

where $K(e) = \int_0^{\pi/2}(1 - e^2\sin^2\theta)^{-1/2}d\theta$.

There is a transformation similar to the one used by Gauss in the proof above. It is due to a little-known English mathematician J Landen and has turned out to be very fruitful in evaluating elliptic integrals. Landen's transformation is the substitution $e\sin\theta = \sin(2\phi - \theta)$.

Exercise. *Verify that this transformation can also be used to prove the above theorem of Gauss.*

Yet another result that can be proved by a repeated usage of Landen's transformation is the following. For $a > b > 0$, denote by $J(a, b)$, the integral $\int_0^{\pi/2}(a^2\cos^2\theta + b^2\sin^2\theta)^{1/2}d\theta$. Note that $J(a, b) = aE(e)$ where $E(e) = \int_0^{\pi/2}(1 - e^2\sin^2\theta)^{1/2}d\theta$. Then, we have:

Proposition

$$J(a,b) = (a^2 - \sum_{n=0}^{\infty} 2^{n-1}(a_n^2 - b_n^2))I(a,b).$$

We shall return to this proposition in the next section. For some other applications of Landen's transformation, the interested reader may see [2].

In the case of a lemniscate, the total perimeter is

$$\int_0^{\pi} \cos(2\theta)^{-1/2}d\theta = 4\int_0^{\pi/2}(2\cos^2\alpha + \sin^2\alpha)^{-1/2}d\alpha$$

from the change of variables

$$\cos(2\theta) = \cos^2\alpha.$$

So, in this case, the constants a and b in Gauss's theorem are $\sqrt{2}$ and 1, respectively. Therefore, one has

$$M(\sqrt{2}, 1) = \frac{\pi}{2\varpi} \text{ where } \varpi = \int_0^1 (1 - z^4)^{-1/2}dz. \qquad (\diamond)$$

Gauss's 92nd entry made in 1798 in his famous mathematical diary reads "*I have obtained most elegant results concerning the lemniscate, which surpasses all expectation – indeed, by methods which open an entirely* *new field to us*" The 98th entry made in May 1799 says that "*demonstration of this fact (the identity $\diamondsuit$) will open an entirely new field of analysis*" What is this 'new field of analysis' that is being talked about? When a and b are complex numbers, the square-root causes trouble i.e., at each stage b_n has two possible choices. Thus, $M(a,b)$ is a multi-valued function and the various different values are related. Determining the relations between them requires the so-called theory of modular forms of weight 1 for congruence subgroups of $SL(2,\mathbf{Z})$. This is what Gauss refers to as 'the new field of analysis.'

The fascination of calculating approximations to π is said to have existed among mathematicians for as long as 4000 years – even laymen have had their fingers in the π !

Ways to Approximate π

The fascination of calculating approximations to π is said to have existed among mathematicians for as long as 4000 years – even laymen have had their fingers in the π! The study of π has led to deep questions – an example is the ancient Greek problem of *squaring the circle* – see the discussion of this problem and other related ones in [3]. It was only at the end of the 19th century that Lindemann (1852-1939) finally proved that π is transcendental i.e., that there is no nonconstant polynomial with rational coefficients that has π as a root.

Several great mathematicians have given'formulae' or approximations. Perhaps, the earliest was Archimedes (287-212 B.C.) who gave us the approximation $\frac{223}{71} < \pi < \frac{22}{7}$. The notation π is also due to him. Aryabhatta (476-550 A.D.) gave the approximation $\pi \sim \frac{62832}{20000} = 3.1416$. Leibniz (1646-1716) and Euler (1707-1783) gave the respective formulae:

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} +$$

$$\frac{\pi^2}{8} = 1 + \frac{1}{3^2} + \frac{1}{5^2} + \frac{1}{7^2} +$$

Legendre gave the following formula for π in terms of the elliptic integrals ([4] has a new and rather easy proof):

Legendre's Formula

Let $0 < e < 1$ *and* $e' = \sqrt{1-e^2}$. *Then*

$$K(e)E(e') + K(e')E(e) - K(e)K(e') = \frac{\pi}{2}.$$

Let us use this formula with the special value $e = e' = 1/\sqrt{2}$ to give an approximation for π which involves the agM. This will be seen to be reminiscent of Archimedes's method. His method was to approximate the unit circle by inscribed and circumscribed regular polygons. Then, one has $P_n < 2\pi < Q_n$ where P_n, Q_n are, respectively, the perimeters of the inscribed and the circumscribed regular polygons of n sides. In fact, the numbers $a_n := 1/Q_{2^n}$ and $b_n = 1/P_{2^n}$ satisfy the recursion $a_{n+1} = (a_n + b_n)/2, b_{n+1} = \sqrt{a_{n+1}b_n}$. This is very similar to the defining recursion for $M(1/2, 1/4)$. However, Archimedes's sequences converge much more slowly.

Now, Legendre's formula gives

$$2K(1/\sqrt{2})E(1/\sqrt{2}) - K(1/\sqrt{2})^2 = \frac{\pi}{2}.$$

On the other hand, the proposition above implies (with $a = 1, b = 1/\sqrt{2}$) that

$$E(1/\sqrt{2}) = (1 - \sum_{n=0}^{\infty} 2^{n-1}(a_n^2 - b_n^2))K(1/\sqrt{2}).$$

Combining these two with Gauss's theorem, we have

$$\pi(1 - \sum_{i=1}^{\infty} 2^{i+1}(a_i^2 - b_i^2)) = 4M(1, 1/\sqrt{2})^2$$

Thus, one has the approximations

$\pi_n = 4a_{n+1}^2 / \sum_{i=1}^{n} 2^{i+1}(a_i^2 - b_i^2)$ which converge quadratically to π. In other words, the number of correct digits doubles at each step.

Finally, we end with a formula due to Ramanujan (see [5]). Ramanujan used the theory of modular functions (this was alluded to at the end of the last section) to give some amazing formulae for π as well as for the perimeters of ellipses. One of these is:

$$\frac{1}{2\pi\sqrt{2}} = \sum_{n=0}^{\infty} \frac{1103+26390n}{99^{4n+2}} \frac{(4n)!}{4^{4n}(n!)^4}.$$

Using just the first term, one gets the approximation 0.112539536... to $\frac{1}{2\pi\sqrt{2}} = 0.112539539...$

Suggested Reading

[1] **B Sury, *Resonance*, Vol.4, No.12, p.48, 1999.**

[2] **B Berndt, *Ramanujan's Notebooks*, Part III, Springer-Verlag, 1991.**

[3] **Shashidhar Jagadeesan, *Resonance*, Vol. 4, No. 3, 1999.**

[4] **G Almkvist and B Berndt, *Amer. Math. Monthly*, Vol.94 585-608, 1987.**

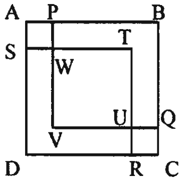
[5] **J H Borwein and P B Borwein, *Pi and the AGM*, Wiley, Canada, 1987.**

[6] **D Cox, *L'Enseignement Mathematique*, Vol.30, 275-330, 1984.**

Address for Correspondence
B Sury
Statistics & Mathematics Unit
Indian Statistical Institute
Bangalore 560050, India.
Email: sury@isibang.ac.in

The Square Root by Infinite Descent

One of the first instances of a proof by contradiction that one learns in school is that of the irrationality of $\sqrt{2}$. Here is a proof due to Tannenbaum by 'infinite descent' – the method employed by Fermat and by Brahmagupta even earlier in other contexts. Suppose $2 = \frac{a^2}{b^2}$ where a, b are natural numbers. In the figure, ABCD is a square of side a, PBQV and STRD are both squares of side b. As $a^2 = 2b^2$, one has Area(TUVW) = Area(PWSA) + Area(QCRU). This means $(2b-a)^2 = 2(a-b)^2$ i.e., $2 = \frac{(2b-a)^2}{(a-b)^2}$. On the other hand, it is clear from the figure that $0 < 2b - a < a$ which produces a smaller numerator than the original one setting in a process of infinite descent. The resultant contradiction forced on us proves the irrationality of $\sqrt{2}$.



Kanakku Puly