How to show that various numbers either can or cannot be constructed using only a straightedge and compass

Nick Janetos

June 3, 2010

1 Introduction

> *It has been found that a circular area is to the square on a line equal to the quadrant of the circumference, as the area of an equilateral rectangle is to the square on one side...*
> -Indiana House Bill No. 246, 1897

Three problems of classical Greek geometry are to do the following using only a compass and a straightedge:

1. To "square the circle": Given a circle, to construct a square of the same area,
2. To "trisect an angle": Given an angle, to construct another angle 1/3 of the original angle,
3. To "double the cube": Given a cube, to construct a cube with twice the area.

Unfortunately, it is not possible to complete any of these tasks until additional tools (such as a marked ruler) are provided. In section 2 we will examine the process of constructing numbers using a compass and straightedge. We will then express constructions in algebraic terms. In section 3 we will derive several results about transcendental numbers. There are two goals: One, to show that the numbers e and π are transcendental, and two, to show that the three classical geometry problems are unsolvable. The two goals, of course, will turn out to be related.

2 Constructions in the plane

The discussion in this section comes from [8], with some parts expanded and others removed.

The classical Greeks were clear on what constitutes a construction. Given some set of points, new points can be defined at the intersection of lines with other lines, or lines with circles, or circles with circles. A line is defined by two points, a circle is defined by a point at the center of the circle and a point on the circle. If P and Q are points in the plane, then we call the line segment defined by them PQ. The distance between them is $|PQ|$. Denote the line defined by the two points by $\mathcal{L}(P, Q)$. Denote a circle drawn at point P with radius $|PQ|$ by $\mathcal{C}(P, Q)$.

2.1 Using the compass and straightedge

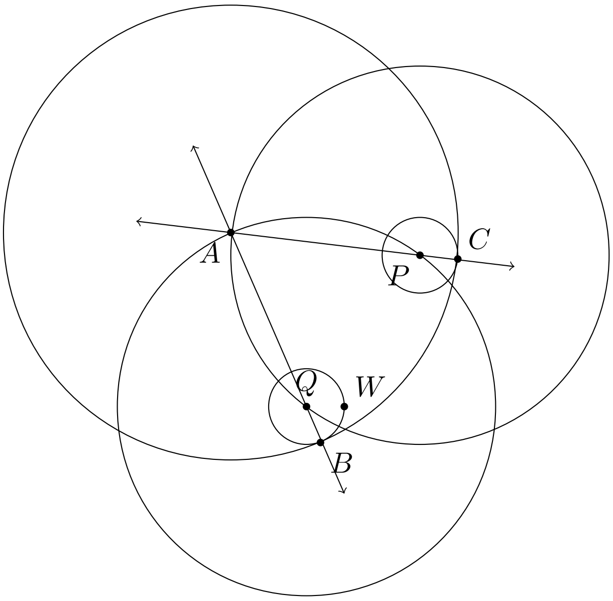
The following constructions and theorems will be taken without proof, refer to [5] for complete proofs. That it is possible to construct the midpoint of a line segment, to construct a line perpendicular to a point on a line, that the ratio of the corresponding sides of similar triangles are equal, and that it is possible to bisect an angle.

One difficult restriction on constructions we might want to make is the fact that we cannot directly construct circles with radius equal to the distance between any two points. We must use the distance between the point we have chosen as the center of the circle and some other point we have chosen as a point on its radius. However, this is an unnecessary restriction, as the following theorem makes clear.

Theorem 1. (Compass Equivalence)

Given points P, Q, and W, it is possible to construct a circle centered at P with radius $|QW|$.

Proof. The points W and B both lay on the same circle, $\mathcal{C}(Q, W)$. Hence $|QW| = |QB|$. The points B and C both lay on the same circle $\mathcal{C}(A, B)$. Hence $|AB| = |AC|$. And the point A lays on the intersection of two circles centered at P and Q. Hence $|AQ| = |AP|$. Then $|PC| = |AC| - |AP| = |AB| - |AQ| = |QB| = |QW|$. So the circle $\mathcal{C}(P, C)$ has radius $|QW|$.



□

From now on, then, it is possible to define a circle in a third way: Given points P, Q, and W, we write $\mathcal{C}(P, |QW|)$ to denote the circle centered at P of radius $|QW|$.

2.2 Representing constructions algebraically

Now that we have covered some basic results of classical geometry, it is time to represent compass and straightedge constructions algebraically.

2.2.1 Some definitions

First, we identify Euclidean space with the complex plane. Choose two points, O and R. Define $|OR|$ to be the unit distance, 1. Construct a line L perpendicular to $\mathcal{L}(O,R)$ through O and define $I = \mathcal{C}(O,R) \cup L$. Then by construction $|OI| = |OR| = 1$. The line $\mathcal{L}(O,R)$ is the *real axis*, or x-axis; $\mathcal{L}(O,I)$ is the *imaginary axis*, or y-axis; and O is the *origin.* Also, we have a coordinate system defined by $|OR| = 1$, e.g., $O = (0,0)$, $R = (1,0)$, and $I = (0,1)$. So we can identify a point in the complex plane, $x + iy$, with a point (x,y) in Euclidean space. From now on, the two notations will be used interchangeably. For example, we might take $P = (x,y)$ to be a point in Euclidean space and write $z = P + i$, or $\overline{P}$. This will be understood to mean $z = x + i(y+1)$ or $x - iy$.

Definition 1. *A point P is* 0-constructable *if $P \in \{O,R,I\}$.*

Using this definition, we define what it means to be constructable using induction:

Definition 2. *A point P is n-*constructable*, or simply* constructable*, if there exist points $P_1, \ldots, P_n$ (where P_i is j-constructable for $0 \leq j < n$) such that there are A, B, C, and D, not neccesarily all distinct, in $\{O,R,I\} \cup \{P_1, \ldots, P_n\}$ where*

1. *P is an intersection of $\mathcal{L}(A,B)$ and $\mathcal{L}(C,D)$,*
2. *P is an intersection of $\mathcal{L}(A,B)$ and $\mathcal{C}(C,D)$,*
3. *P is an intersection of $\mathcal{C}(A,B)$ and $\mathcal{C}(C,D)$.*

Note that there are only countably many constructable points, because the set of 0-constructable points has order 3, and given the set of n-constructable points there are only finitely many more elements in the set of $n+1$-constructable points. So we have already established that almost all points are not constructable – we're over 99% there. But to say more about the real number line, we'll need to establish a correspondence between constructable points and real numbers.

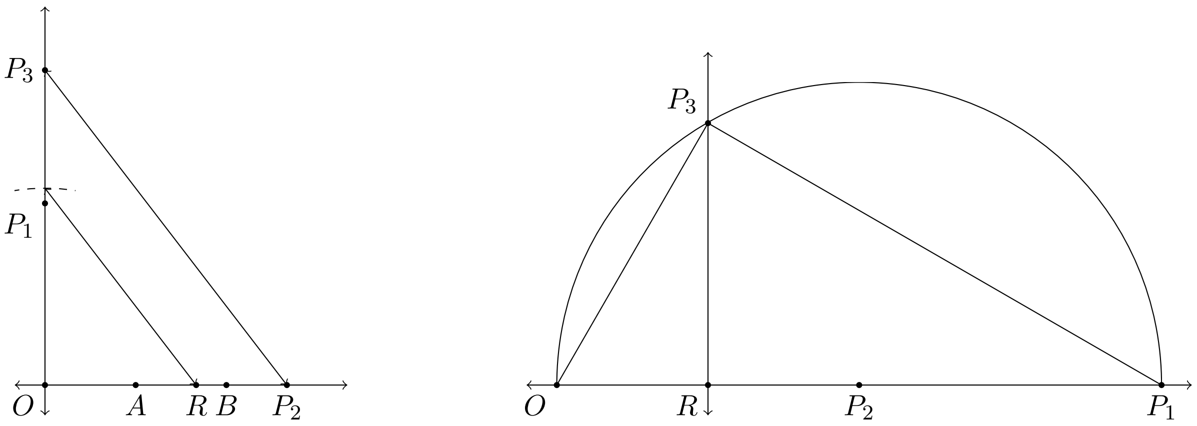
2.2.2 A correspondence between constructable points and real numbers

Lemma 1. *Let $\mathcal{K}$ be the set of complex numbers in $\mathbb{C}$ corresponding to constructable points in Euclidean space. Then $\mathcal{K} \cap \mathbb{R}$ is closed under addition, multiplication, square roots, every element has a multiplicative inverse (except O), and every element has a additive inverse.*

Proof. We must verify that given $a, b \in \mathcal{K} \cap \mathbb{R}$, that $a + b, ab, -a, a^{-1} \in \mathcal{K} \cup \mathbb{R}$. The general idea will be to present the appropriate geometric proof that all of these complex numbers are identified with constructable points. So let $A = (a,0)$ and $B = (b,0)$. Also, assume that $a \neq 0$ and $b \neq 0$, otherwise the following results are trivial.

$a + b$ This is easily shown by drawing $\mathcal{C}(A, |B|)$. It intersects the real axis at two points, by construction, these points are $(a+b,0)$ and $(a-b,0)$.

$-a$ Obvious: Draw $\mathcal{C}(O,A)$ and let P_1 be its other intersection with the real axis. Then $P_1 = (-a,0)$.



(a) ab

(b) $\sqrt{a}$

ab Assume $a > 0$ and $b > 0$. Draw $\mathcal{C}(O, B)$ and call the upper point where it intersects the imaginary axis $P_1 = (0, b)$. Because $R, A \in K$ we have already shown that $(1 + a, 0) \in K$. Let $P_2 = (1 + a, 0)$. Draw $\mathcal{L}(P_1, R)$, and construct a line through P_2 parallel to it, call the point where it intersects the imaginary axis P_3.

Then by construction the triangles ORP_1 and OP_2P_3 are similar triangles, hence $|OP_2|/|OR| = |OP_3|/|OP_1|$, which means that $a+1 = (b+|P_1P_3|)/b$, therefore, $|P_1P_3| = ab$. If either $a < 0$ or $b < 0$, then we have constructed $-ab$ and hence by (b) we have $ab \in \mathcal{K}$. If both $a < 0$ and $b < 0$ then we have constructed ab.

a^{-1} The proof for this is similar to the previous proof. Let $P_1 = (1 + a, 0)$. Let P_2 be the upper intersection of $\mathcal{C}(O, P_1)$ with the y-axis. Define P_3 to be the intersection of the x-axis and the line through P_2 parallel to RA. As in the previous proof, we have similar triangles, and $|RP_3| = a^{-1}$.

$\sqrt{a}$ Let $P_1 = (1 + a, 0)$. Let P_2 be the midpoint of OP_1. Let P_3 be the upper intersection of $\mathcal{C}(P_2, O)$ with the vertical line through R. Then by construction the triangles ROP_3 and RP_1P_3 are similar, so that $|OR|/|RP_3| = |RP_3|/|RP_1|$. So $|RP_3|^2 = |OR||RP_1| = |RP_1| = a$, i.e., $|RP_3| = \sqrt{a}$.

□

Lemma 2. *$P = (x, y)$ is constructable if and only if $P_1 = (x, 0)$ and $P_2 = (y, 0)$ are constructable.*

Proof. Assume that $P_1 = (x, 0)$ and $P_2 = (y, 0)$ are constructable. Let $P_3 = (0, y)$. Then P is the intersection of the line through P_1 perpendicular to the x-axis and the line through P_3 perpendicular to the y-axis.

Now assume that $P = (x, y)$ is constructable. The point P_1 is the intersection of the x-axis and the line through P perpendicular to the x-axis, and the same for P_2.

□

Theorem 2. *$\mathcal{K}$ is closed under addition, multiplication, square roots, every element has a multiplicative inverse (except O), and every element has an additive inverse.*

Proof. This is a consequence of the preceeding two lemmas. For let P_1 and P_2 be points in $\mathcal{K}$ and denote their corresponding complex numbers by $x_1 + iy_1$ and $x_2 + iy_2$. Then $P_1 + P_2 = x_1+x_2+i(y_1+y_2)$ is in $\mathcal{K}$ by lemmas 1 and 2. And $P_1P_2 = (x_1+iy_1)(x_2+iy_2) = x^2-y^2+i(x_1y_2+y_1x_2)$ is in $\mathcal{K}$ by the same lemmas. And if we denote P_1 as $re^{i\theta}$, then $e^{i\theta/2} \in \mathcal{K}$, because every angle can be bisected, and $r = \sqrt{x^2+y^2}$ is in $\mathcal{K} \cap \mathbb{R}$, hence so is $\sqrt{r}$, so $\sqrt{P_1} = \sqrt{r}e^{i\theta/2} \in \mathcal{K}$. □

Corollary 1. *$\mathcal{K}$ contains the set of rational numbers, or $\mathbb{Q} = \{p/q \mid p, q \in \mathbb{Z}\}$.*

Proof. This is easy to see: Since $1 \in \mathcal{K} \cap \mathbb{R}$ and $\mathcal{K} \cap \mathbb{R}$ is closed under addition and has additive inverses, we have $\mathbb{Z} \subseteq \mathcal{K}$. And since $\mathcal{K}$ is closed under multiplication and has multiplicative inverses, it also contains all elements of the form p/q where $p, q \in \mathbb{Z}$, which is $\mathbb{Q}$. □

Now we can write down some easy results. For example, if $a, b, c \in \mathcal{K}$, then the roots of the quadratic equation $az^2 + bz + c$ are given by the quadratic formula, which involves only addition, multiplication, and the square root. So the roots of the equation are also in $\mathcal{K}$.

2.2.3 Further results about constructable numbers

We're not concerned with which numbers are constructable, we're concerned with which numbers are not constructable. We've shown that $\mathcal{K}$ is closed under square roots, now we show that it is not closed under any root which is not a power of 2. For example, it is closed under 4th roots, and 8th roots, and 16th roots, but not 7th roots.

Theorem 3. *A number $z \in \mathbb{C}$ is n-constructable if and only if it is obtainable by adding, multiplying, and taking square roots of rational numbers.*

Proof. One direction is obvious: If z has been obtained by adding, multiplying, and taking square roots of rational numbers, then by theorem 2 it is constructable.

Now assume that z is n-constructable, and denote the point in the Euclidean plane identified with z by P_n. Note that this theorem is trivially true for numbers which are 0-constructable, because those numbers are all rational numbers anyways. We'll proceed with a proof by induction on n. Assume that this theorem holds for all points which are $n-1$-constructable.

Since z is n-constructable there exist points $P_1, \dots, P_{n-1}$ such that P_n is either

1. an intersection of $\mathcal{L}(P_i, P_j)$ and $\mathcal{L}(P_k, P_l)$,
2. an intersection of $\mathcal{L}(P_i, P_j)$ and $\mathcal{C}(P_k, P_l)$,
3. an intersection of $\mathcal{C}(P_i, P_j)$ and $\mathcal{C}(P_k, P_l)$,

where P_i, P_j, P_k, and $P_l \in \{P_1, \dots P_n\}$. Let $P_i = (i_x, i_y)$, $P_j = (j_x, j_y)$, $P_k = (k_x, k_y)$, and $P_l = (l_x, l_y)$. Note that all of these numbers are expressible as sums, products, and square roots of rational numbers.

In the first case, we have the intersection of two lines, which we could express using the formulas $y - i_y = \dfrac{j_y - i_j}{j_x - i_x}(x - i_x)$ and $y - k_y = \dfrac{l_y - k_j}{l_x - k_x}(x - k_x)$, or perhaps one of the lines is of the form $x = i_x$. (Not both, however, because then they'd be parallel.) The intersecting point would be a solution of these two equations, and clearly we could solve for x and y by simply using addition and multiplication.

In the second case, we have the intersection of a line and a circle. Again we might express the circle as something of the form $y = \pm\sqrt{(x-k_x)^2 - |P_kP_l|^2} + k_y$ and the line as something of the form $y - i_y = \frac{j_y - i_j}{j_x - i_x}(x - i_x)$. And again, note that to find a solution for x and y we would use addition, multiplication, and square roots.

In the third case, we have the intersection of two circles. Then if we have $y = \pm\sqrt{(x-i_x)^2 - |P_kP_l|^2} + k_y$ and $y = \pm\sqrt{(x-k_x)^2 - |P_kP_l|^2} + i_y$ we again will use addition, multiplication, and square roots to find a solution for x and y.

In all these cases we are adding, multiplying, and taking square roots of numbers which by assumption were formed by adding, multiplying, and taking square roots of rational numbers. Hence by induction z satisfies the conditions of the theorem. □

We're able to dispose of a few of the classical Greeks' problems now.

Theorem 4. *It is impossible to double the cube.*

Proof. To double the cube means to construct a cube of with sides of length double the unit cube (the cube whose sides have length 1). But this means to construct the number $\sqrt[3]{2}$, which by theorem 3 is impossible. □

Theorem 5. *It is impossible to trisect an arbitrary angle.*

Proof. An angle θ is given by two intersecting lines. Let $\pi/3$ be the angle formed by the lines. Assume without loss of generality that the lines intersect at the origin and that one line is the x-axis. If we could trisect an angle, then we could construct the point $(\cos\theta/3, \sin\theta/3)$, which is the intersection of the line not on the x-axis and $\mathcal{C}(O, R)$. Compute

$$\mathrm{Re}(e^{3i\theta}) = \mathrm{Re}\big((\cos\theta + i\sin\theta)^3\big) = 4\cos^3\theta - 3\cos\theta,$$

hence $\cos 3\theta = 4\cos^3\theta - 3\cos\theta$. We have $\cos\theta = \frac{1}{2}$, and $\cos\pi = -1$. Let $\theta = \pi/9$ and $u = 2\cos\theta$, and we get the polynomial $u^3 - 3u - 1 = 0$. This polynomial is irreducible, (i.e., it has no factorization into polynomials with rational coefficients of degree greater than zero) and it is an easily-believed fram from algebra that since the degree of the polynomial is 3, u cannot be expressed as a combination of addition, multiplication, and square roots of rational numbers. So by theorem 3 it is impossible to trisect an angle. □

This last proof leads into the next section. If we want to show that it is impossible to square the circle, we must use a less clumsy term for the conditions of theorem 3 then "can be expressed as a combination of sums, products, and square roots of rational numbers".

3 Transcendental numbers

> *...the fourth important fact, that the ratio of the diameter and circumference [of the circle] is as five-fourths to four; and because of these facts and the further fact that the rule in present use fails to work both ways mathematically, it should be discarded as wholly wanting and misleading in its practical applications.*
>
> -Indiana House Bill No. 246, 1897

Definition 3. *A number $x \in \mathbb{R}$ is* irrational *if it is not rational.*

Definition 4. *A number $z \in \mathbb{C}$ is* algebraic *if it is the zero of a polynomial with rational coefficients.*

Definition 5. *A number $z \in \mathbb{C}$ is* transcendental *if it is not algebraic.*

Note that if z is transcendental, then clearly it is not constructable. In this section, we will first look at a specific proof that the number e is transcendental. Then we will examine why the methods used in the proof were appropriate. Finally, we will prove that π is irrational, a consequence of which will be that it is impossible to square the circle.

3.1 e is transcendental

Before we begin, let $f(x) = \sum_{k=0}^{n} a_k x^n$ be a polynomial of degree n with rational coefficients, and define

$$I_f(t) = e^t \int_0^t e^{-x} f(x)\, dx, \qquad t > 0.$$

Setting $u = f(x)$ and $dv = e^{-x}\, dx$ we use integration by parts to obtain

$$\begin{aligned} I_f(t) &= \int_0^t e^{-x} f(x)\, dx \\ &= e^t \left(-f(x)e^{-x}\Big]_0^t + \int_0^t e^{-x} f'(x)\, dx \right) \\ &= e^t f(0) - f(t) + e^t \int_0^t e^{-x} f'(x)\, dx. \end{aligned} \tag{1}$$

Now, since $f'(x)$ is a polynomial with rational coefficients of degree $n-1$, we can apply the same process of integration by parts to the integral in (1). Applying the process $n+1$ times to $I_f(t)$ we get the following equation:

$$I_f(t) = \sum_{k=0}^{n} \left(e^t f^{(k)}(0) - f^{(k)}(t) \right). \tag{2}$$

Now define $F(x) = \sum_{k=0}^{n} |a_k| x^k$. Then clearly $F(t) > F(0)$ for $t > 0$ and n at least 1. We use this fact and the triangle inequality to obtain the estimate

$$\begin{aligned} |I_f(t)| &= \left| e^t \int_0^t e^{-x} f(x)\, dx \right| \\ &\leq |e^t| \left(\int_0^t |e^{-x}||f(x)|\, dx \right) \\ &\leq |e^t| \left(\int_0^t |e^{-x}| F(x)\, dx \right) \\ &\leq |e^t| \left(tF(t) \right). \end{aligned} \tag{3}$$

Later, in section 3.2 we'll examine the motivations for choosing the particular integral we did for $I_f(t)$.

Theorem 6. *e is transcendental.*

Proof. Although various details have been fleshed out, the following proof is found in [2].

Assume otherwise. Then there is some polynomial $a_0 + a_1x + \ldots + a_nx^n$ of degree n such that $\sum_{k=0}^{n} a_k e^k = 0$, and where $a_0, \ldots, a_n \in \mathbb{Q}$. Without loss of generality, multiply both sides by a scalar factor to get a_0 a positive integer. Define

$$f(x) = x^{p-1}(x-1)^p(x-2)^p \ldots (x-n)^p, \tag{4}$$

where p is some large prime number at least as big as n and at least as big as a_0. Define $J = a_0I_f(0) + a_1I_f(1) + \ldots + a_nI_f(n)$.

We will now proceed to show that $(p-1)! \leq |J| \leq c^p$, for some constant c.

Proof that $(p-1)! \leq |J|$**:** Note that the degree of f is $(p-1) + np = (n+1)p - 1$. By (2), we have

$$I_f(t) = \sum_{k=0}^{(n+1)p-1} \left(e^t f^{(k)}(0) - f^{(k)}(t)\right).$$

Hence

$$\begin{aligned}
J &= \sum_{k=0}^{n} a_k I(k) \\
&= \sum_{k=0}^{n} a_k \left(\sum_{j=0}^{(n+1)p-1} \left(e^k f^{(j)}(0) - f^{(j)}(k)\right) \right) \\
&= \sum_{k=0}^{n} \left(e^k \sum_{j=0}^{(n+1)p-1} f^{(j)}(0) - \sum_{j=0}^{(n+1)p-1} f^{(j)}(k) \right) \\
&= \left(\sum_{k=0}^{n} a_k e^k \right) \left(\sum_{j=0}^{(n+1)p-1} f^{(j)}(0) \right) - \sum_{k=0}^{n} \sum_{j=0}^{(n+1)p-1} a_k f^{(j)}(k) \ , \qquad (5)
\end{aligned}$$

where in (5) we used the distributive law to expand the sums. But by assumption, e is algebraic, so $\sum_{k=0}^{n} a_k e^k = 0$. Therefore, the first term in (5) is zero, and we get

$$J = -\sum_{k=0}^{n} \sum_{j=0}^{(n+1)p-1} a_k f^{(j)}(k). \tag{6}$$

Let's break up J further:

If $j < p-1$ in (6) then none of the terms of $f(x)$ are differentiated enough times for them to disappear, hence $f^{(j)}(k) = 0$.

If $j = p-1$ then the only term which disappears due to differentiation is x^{p-1}, hence $f^{(j)}(k) = 0$ for $k > 0$. If $k = 0$, then computing by the product rule we see that the only term which contributes a nonzero value is the one in which x^{p-1} is differentiated $p-1$ times, hence $f^{(j)}(0) = (p-1)!(-1)^{np}(n!)^p$. Finally, if $j > p-1$ we get by the product rule a bunch of ugly terms. Note however that all of the terms must be divisible by $p!$, because using the product rule we see that the only terms that contribute anything are the terms where $(x-k)^p$ has been differentiated p times.

In other words, we can write

$$J = a_0 f^{(j-1)}(0) + \sum_{k=0}^{n} \sum_{j=p}^{n(p+1)-1} a_k f^{(j)}(k) = a_0 M(p-1)! + Np! = (p-1)!(a_0 M + Np), \quad (7)$$

where $M = (-1)^{np}(n!)^p$ and N is just some large number representing the result of dividing

$$\sum_{k=0}^{n} \sum_{j=p}^{n(p+1)-1} a_k f^{(j)}(k)$$

by $p!$. We already showed that N is an integer.

Note that p does not divide M, because $p > M$. Also, by assumption, $p > a_0$, so p does also not divide a_0. If $a_0 M + Np$ were to equal zero, then $-Np = a_0 M$. So $N = -\frac{a_0 M}{p}$. But p does not divide a_0 or M, hence $a_0 M + Np \neq 0$, so $|a_0 M + Np| > 0$. But $a_0 M + Np$ is an integer, so in fact $|a_0 M + Np| \geq 1$. Therefore, putting absolute values on either side on (7), we get that $|J| = (p-1)!|a_0 M + Np| \geq (p-1)!$, which is what we wanted to show.

Proof that $|J| \leq c^p$: On the other hand, note that by the triangle inequality $F(k)$ (which we defined near the beginning of this section) satisfies

$$\begin{aligned} F(k) &\leq |k|^{p-1}|k-1|^p \dots |k-n|^p \\ &\leq (2n)^{p-1}((2n)\dots(2n))^p \\ &= (2n)^{p-1+pn} \\ &= (2n)^{(n+1)p-1}. \qquad (8) \end{aligned}$$

Armed with this inequality, and (3), (and the fact that $1+2+\ldots+n \leq n^2$), we are prepared to estimate $|J|$ again. Let $a = \max(|a_1|, \ldots, |a_n|)$. Then

$$\begin{aligned} |J| &\leq |a_0||I(0)| + \ldots + |a_n||I(n)| && \text{(by the triangle inequality)} \\ &\leq |a_0|F(0) + |a_1|eF(1) + \ldots + |a_n|ne^n F(n) && \text{(by (3))} \\ &\leq an^2e^n(2n)^{(n+1)p-1} && \text{(by (8))} \\ &= \frac{an^2e^n}{2n}\left((2n)^{n+1}\right)^p \\ &\leq c^p, \end{aligned}$$

for some large constant c, because a and n are fixed. So $|J| \leq c^p$.

We have shown that $(p-1)! \leq |J| \leq c^p$, so $(p-1)! \leq c^p$. But clearly for large p we have $(p-1)! \geq c^p$, which is a contradiction. So e is transcendental.

□

3.2 Motivation for the estimates

The material in this section is pulled from [1], [3], and [7], with some changes to clarify and unify the presentation.

The proof that e was transcendental involved taking a transformation of a function, I_f, and making several estimates on it. But it is an unsatisfying proof, because there is no indication of the reason why I_f should be a good candidate for making estimates. The reason it is unsatisfying is because it skips the full process through which Hermite proved that e is transcendental. In this section, we will examine the subject of Padé approximants, which are useful in proving things like irrationality and transcendence.

Definition 6. *Let $f(z)$ be a function analytic on a domain, and choose z_0 in that domain. Then if $P(z), Q(z)$ are polynomials of degree less than or equal to some integers m and n respectivally, such that the lowest order term of the series expansion of $Q(z)f(z) - P(z)$ around z_0 has order at least $n + m + 1$, then we call $\frac{P(z)}{Q(z)}$ the* Padé approximant to $f(z)$ of order n, m at z_0.

Note that this is equivalent to saying that as $z \to z_0$ we have

$$f(z) = \frac{P(z)}{Q(z)} + \mathcal{O}((z - z_0)^{n+m+1}),$$

assuming that such a $P(z)$ and $Q(z)$ exist, which they may not. If they do exist, then it happens that they are unique, although this is not particularly relevant.

The reason why we are considering Padé approximants is shown in the following lemma:

Lemma 3. *Let $x \in \mathbb{R}$. If there exists series of integers $\{p_n\}$, $\{q_n\}$, such that*

$$\lim_{n\to\infty} q_n x - p_n = 0, \tag{9}$$

but

$$q_n x - p_n \neq 0, \quad \textit{for all } n, \tag{10}$$

then x is irrational.

Proof. Suppose that x is rational. Then $x = \frac{p}{q}$ for some integers p, q. So

$$q_n x - p_n = q_n \left(\frac{p}{q}\right) - p_n = \frac{q_n p - p_n q}{q}.$$

By (10), this fraction is nonzero for all n, which implies that $q_n p - p_n q$ is nonzero for all n. All of these numbers are integers, hence $|q_n p - p_n q| \geq 1$ for all n. Then $|q_n x - p_n| \geq 1/q$, but by (9), this quantity becomes very small for n large. So it cannot be bounded from below. Hence x must be irrational. □

Padé approximants provide a method of approximating numbers with rational functions, so it's possible to use this lemma in conjuction with them to show that e and π are irrational. But in order to say something about whether a number is transcendental, we will have to introduce a modification of Padé approximants. As we shall see, by making appropriate modifications in the above lemma and Padé approximants, almost the same reasoning will suffice to show that a number is transcendental.

Definition 7. *Let $f_1(z), \ldots, f_r(z)$ be r functions analytic on some domain. Define $\vec{n} = (n_1, \ldots, n_r)$ and $\vec{m} = (m_1, \ldots, m_r)$ where $m_1, \ldots, m_r, n_1, \ldots, n_r \in \mathbb{N}$. Denote $n_1 + \ldots + n_r$ by $|\vec{n}|$, and $m_1 + \ldots m_r$ by $|\vec{m}|$. Say there exists a polynomial $Q(z)$ such that the degree of $Q(z)$ is less than $|\vec{n}|$, and say there exist r polynomials $P_1(z), \ldots, P_r(z)$ such that $\deg P_j(z) \leq m_j$ for $1 \leq j \leq r$. If the lowest order term of the series expansion of $Q(z)f_j(z) - P_j(z)$ around z_0 has order at least $n_j + 1$ for all j, then these polynomials are called the* Hermite-Padé approximants to $f_1(z), \ldots, f_r(z)$ at z_0.

Note that these are similar to Padé approximants, the different being that we are using r rational functions with a common denominator to approximate r functions.

The usefulness of these polynomials arises from the following lemma:

Lemma 4. *Let $z \in \mathbb{C}$. If, given an arbitrary set of $r+1$ integers $a_0, \dots, a_r$, we can find $r+1$ sequences of integers $\{q_n\}$, $\{p_{1,n}\}, \dots, \{p_{r,n}\}$ such that*

$$\lim_{n\to\infty} q_n z^j - p_j = 0 \tag{11}$$

for all $1 \le j \le r+1$ but

$$q_n + \sum_{j=1}^{r} a_j p_{j,n} \neq 0 \tag{12}$$

for all n, then z is transcendental.

Proof. Suppose that z is not transcendental. This means that there is some sequence of integers $a_0, \dots, a_r$ such that $\sum_{j=0}^{r} a_j x^j = 0$. So

$$\sum_{j=0}^{r} a_j(q_n x^k - p_{j,n}) = \sum_{j=0}^{r} a_j(q_n x^k) - \sum_{j=0}^{r} a_j p_{j,n} = -\sum_{j=0}^{r} a_j p_{j,n}.$$

By (12), this is nonzero. Therefore,

$$\left| \sum_{j=0}^{r} a_k(q_n z^j - p_{j,n}) \right| \ge 1.$$

But by (11), every term in the sum goes to zero as $n \to \infty$, so the sequence of sums cannot be bounded from below. Hence z must be transcendental. □

Note the suprising similarity between this lemma and the previous one, suprising because the definitions for algebraic and rational seem in a certain sense to be fundamentally different.

Now we will prove again that e is transcendental. This is in the vein of the original proof used by Hermite to prove e irrational, and it explicitly makes use of estimates and equations we derived in the previous proof that e was transcendental in order to derive Hermite-Padé approximants to the function e^z.

Theorem 7. *e is transcendental.*

Proof. Let $a_0, \dots, a_r$ be a sequence of integers, and define $\vec{n} = (n_1, \dots, n_r)$ and $\vec{m} = (m_1, \dots, m_r)$ as in the definition of Hermite-Padé approximants. Let $|\vec{n}| = n_1 + \dots + n_r$ and similarly let $|\vec{m}| = m_1 + \dots + m_r$.

Hermite noticed that the Hermite-Padé approximants can be derived explicitely if $\vec{n}$ and $\vec{m}$ have the following condition: That $m_j + n_j = N + |\vec{n}|$ for $1 \le j \le r$ and for some positive integer N. Now define

$$Q(z) = z^{|\vec{n}|+N+1} \int_0^\infty T(x) e^{-zx}\, dx,$$

where $T(x)$ is some polynomial. Notice that this is simply the Laplace transform of $T(x)$ multiplied by $z^{|\vec{n}|+N+1}$ and recall that this will yield a polynomial of degree $(|\vec{n}|+N)-k$, where k is the degree of the lowest order term of $T(x)$. Define

$$P_j(z)=z^{|\vec{n}|+N+1}\int_0^\infty T(x+j)e^{-zx}\,dx,$$

for $1\le j\le r$. This is the Laplace transform of $T(x+j)$ multiplied by $z^{|\vec{n}|+N+1}$, and it will yield a polynomial of degree $(|\vec{n}|+N)-k'$, where k' is the degree of the lowest order term of $T(x+j)$.

We are trying to find polynomials which are the Hermite-Padé approximants to e^z of a certain order to satisfy the conditions of the lemma. Specifically, we want the degree of $Q(z)$ to be of order $|\vec{n}|$, and we want the degree of $P_j(z)$ to be of order m_j. This suggests that we choose a polynomial $T(x)$ such that $T(x)$ has a zero of order N at $x=0$ and a zero of order m_j at $x=j$. The following obvious choice presents itself:

$$T(x)=x^N(x-1)^{m_1}\dots(x-r)^{m_r}.$$

Note that $T(x)$ has degree $N+|\vec{m}|$. Recall by the previous discussion that $Q(z)$ has degree $(|\vec{n}|+N)-k$, where k is the lowest degree term of $T(x)$. By construction, that is N, so $Q(z)$ has degree $|\vec{n}|$. And similarly, $P_j(z)$ has degree $|\vec{n}|+N-k'$, where here $k'=n_j$ by construction, so it has degree $|\vec{n}|+N-n_j=m_j$.

Now the reason for choosing (4) becomes more clear – it has zeros of appropriate orders at the appropriate places to ensure that the Hermite-Padé approximants are close enough to e^z.

So we have successfully defined polynomials $Q(z)$, $P_1(z)$, …, $P_r(z)$ which have the appropriate degrees. Now we compute

$$\begin{aligned}Q(z)e^{jz}-P_j(z)&=e^{jz}z^{|\vec{n}|+N+1}\int_0^\infty T(x)e^{-zx}\,dx-z^{|\vec{n}|+N+1}\int_0^\infty T(x+j)e^{-zx}\,dx\\&=e^{zj}z^{|\vec{n}|+N+1}\int_0^\infty T(x)e^{-zx}\,dx-\int_j^\infty T(x)e^{jz-zx}\,dx\\&=e^{zj}z^{|\vec{n}|+N+1}\int_0^j T(x)e^{-zx}\,dx, \qquad (13)\end{aligned}$$

where we used a change of variables substitution on the second integral there. Notice that we have seen something similar to (13) before – see (1). We established in (2) that

$$\int_0^j T(x)e^{-x}\,dx=\sum_{k=0}^{N+\vec{m}}\left(e^jT^{(k)}(0)-T^{(k)}(j)\right).$$

Using the exact same method (repeated integration by parts), we get

$$\int_0^j T(x)e^{-zx}\,dx=\sum_{k=0}^{N+\vec{m}}\left(\frac{e^jT^{(k)}(0)-T^{(k)}(j)}{z^k}\right).$$

Note that by construction $T^{(k)}(j)=0$ for $k\le N+m_j$. Then the order of the lowest nonzero term of (13) must be $N+|\vec{n}|+1=n_j+m_j+1$, which is the appropriate condition for the Hermite-Padé approximant.

Now take $z = 1$, and denote the ith prime by $p(i)$. Take $N = p(i) - 1$ and $n_j = p(i)$ for $1 \leq j \leq r$. Let $q_i = Q(1)/(p(i) - 1)!$ and $p_{j,i} = P_j(1)/(p - 1)!$. We already showed in the first proof that q_i is not divisible by $p(i)$ and $p_{j,i}$ is divisible by $p(i)$. So $q_i + \sum_{k=0}^{r} a_k p_{k,i}$ is not divisible by $p(i)$ for all i, which implies that

$$q_i + \sum_{k=0}^{r} a_k p_{k,i} \neq 0.$$

So we have satisfied the second condition of lemma 4. On the other hand, using the estimate we derived in (8) we have $|T(x)| \leq (2r)^{(r+1)p-1}$, so we have

$$q_i e^i - p_{j,i} = \frac{e^j}{(p-1)!} \int_0^j T(x) e^{-x}\, dx \leq \frac{e^j (2r)^{(r+1)p-1}}{(p-1)!} \int_0^j e^{-x}\, dx \to 0$$

as $i \to \infty$, because the factorial function grows much faster than the power function. So we have satisfied condition 1 of lemma 4, hence e is transcendental.

□

In the original proof, these sorts of estimates were used to provide a contradiction, in other words lemma 4 was built into the proof itself. Here the reason for choosing $I_f(x)$ in (1) becomes clear: It allowed us to precisely control the lowest order nonzero term in the polynomials we produced and thus ensured that the polynomials would approximate e^z in just the right way. Combined with lemma 4, this was enough to show that e is transcendental. Now we will show that π is transcendental, a proof which Lindemann was the first to write down after Hermite's proof of the transcendence of π. As we shall see, it uses the same underlying technique. However, a few results are neccesary from outside this paper, specifically the fact that the product of two algebraic numbers is algebraic and a result regarding certain types of polynomial.

And remember, a consequence of theorem is that it is impossible to square the circle.

Theorem 8. *π is transcendental.*

Proof. The following proof is copied almost verbatim from [2] and [6].

Assume otherwise. Note that since the imaginary number i is the root of $x^2 + 1$ it is algebraic. It is a non-trivial result that the product of two algebraic numbers is itself algebraic, refer to [4] for a proof. So $i\pi$ is algebraic, so it is the root of some polynomial. Let d be the degree of the polynomial of least degree of which $i\pi$ is a root. Since this polynomial has rational coefficients, assume without loss of generality that it has integer coefficients (by multiplying through by some constant). Let l be the coefficient of the leading term. By the fundamental theorem of algebra there are d roots of this polynomial. Denote the roots of the polynomial by $\theta_1, \ldots, \theta_d$. Using the well-known identity $e^{i\pi} = -1$, we must have that

$$(1 + e^{\theta_1})(1 + e^{\theta_2}) \ldots (1 + e^{\theta_d}) = 0. \tag{14}$$

Expanding the left side of this equation, we get a series of 2^d exponents raised to a power of the form e^{Θ}, $\Theta = \epsilon_1\theta_1 + \ldots + \epsilon_d\theta_d$, where $\epsilon_j = 0$ or 1. Some of these powers are nonzero, let n be the number of nonzero powers and denote the nonzero powers by $\Theta_1, \ldots \Theta_n$. We then have

$$e^{\Theta_1} + \ldots + e^{\Theta_n} + 2^d - n = 0. \tag{15}$$

Note that since at least one of the powers is zero, $2^d - n$ is a positive integer.

Define $I_f(t)$ as in the proof that e was transcendental. Again let p be some large prime number. Define

$$J = I_f(\alpha_1) + \ldots + I_f(\alpha_n),$$

where

$$f(x) = l^{np} x^{p-1} (x - \alpha_1)^p \ldots (x - \alpha_n)^p.$$

By (2) and (15) we get

$$J = -q \sum_{k=0}^{(n+1)p-1} f^{(j)}(0) - \sum_{k=0}^{(n+1)p-1} \sum_{j=1}^{n} f^{(k)}(\alpha_j),$$

where recall that l is the coefficient of the leading term of the polynomial which $i\pi$ is a root of.

The sum over j is a symmetric polynomial in $l\alpha_1, \ldots, l\alpha_n$, i.e., interchanging these numbers with each other does not change the polynomial. It is a result of the fundamental theorem of symmetric functions that this sum will be an integer. By the same argument as in theorem 6, we have $f^{(k)}(\alpha_j) = 0$ when $j < p$ so it is trivially divisible by $p!$, $f^{(k)}(0)$ an integer divisible by $p!$ when $j \neq p-1$, and

$$f^{(p-1)}(0) = (p-1)!(-l)^{np}(\alpha_1 \ldots \alpha_n)^p$$

an integer divisible by $(p-1)!$ but not by $p!$ if p is large enough. So for $p > 2^d - n$ we have $|J| \geq (p-1)!$. But by 3 we have that

$$|J| \leq e^t |t| F(t) \leq c^p$$

for some constant c independent of p. For sufficiently large p we have $(p-1)! > c^p$, which is a contradiction, so π is transcendental.

□

This allows us to dispose of the last problem of the classical Greeks.

THEOREM 9. *It is impossible to square the circle.*

Proof. To square the circle means to construct a square of area equal to the area of the unit circle. The unit circle has area π, so this would mean constructing a square with sides of length $\sqrt{\pi}$. But $\sqrt{\pi}$ cannot be constructable, for then $\sqrt{\pi}\sqrt{\pi} = \pi$ would be algebraic, and we have shown that this is not the case. □

References

[1] Van Assche, William. *Padé and Hermite-Padé approximation and orthogonality.* Surveys in Approximation Theory, Vol. 2, pp. 61-91. 2006.

[2] Baker, Alan, *Transcendental Number Theory*, Cambridge University Press, Cambridge, 1975.

[3] Cohn, Henry. *A Short Proof of the Simple Continued Fraction Expansion of e.* The American Mathematical Monthly, Vol. 113, No. 1, pp. 57-62. 2006.

[4] Ireland, Kenneth; Michael Rosen. *A Classical Introduction to Modern Number Theory.* Springer-Verlag, New York, NY. 1990.

[5] Martin, George. *Geometric Proofs.* Springer-Verlag, New York, NY. 1997.

[6] Niven, Ivan. *Irrational Numbers*, The Mathematical Association of America, Rahway, NJ, 1956.

[7] Olds, C.D. *The Simple Continued Fraction Expansion of e.* The American Mathematical Monthly, Vol. 77, No. 9, pp. 968-974. 1970.

[8] Rotman, Joseph, *Galois Theory*, Springer-Verlag, New York, NY. 1990.