Greek Impossibilities, Mathcamp 2010

Nic Ford

1 Introduction

Ancient Greek mathematics was pretty much the same thing as ancient Greek geometry, and one of the things that ancient Greek geometers were very concerned with was the task of constructing shapes using just a compass and a straightedge. The advantage of this method, which is still sometimes taught today in high school geometry classes, is its exactness: all the operations you're allowed to do involve essentially no guesswork at all. Even given these restrictions, though, it's possible to do a great many things using this method: you can construct a regular polygon with 17 sides, bisect an angle, or cut a line segment into fifths. There were some problems, though, that the ancient Greeks (and other people who studied their geometry) were completely unable to solve with a compass and straightedge. They included:

1. Squaring the circle: Given a circle, draw a square with the same area.
2. Trisecting an angle: Given an angle, cut it into thirds.
3. Doubling the cube: Given a face of a cube, draw the face of a cube with twice the volume.

These problems remained open for over two thousand years, until, in the 1800's, it finally became clear why they were having so much trouble: these three tasks are actually impossible! The reason it took so long to figure this out is that the solution requires some study of the theory of fields, which was only developed shortly before the proof of impossibility. These notes are dedicated to proving these surprising facts, and developing the field theory necessary to understand the proof.

2 Fields, Vector Spaces, and Dimension

We start with a review of the definition of a field, a vector space, and the dimension of a vector space. While all the relevant facts will be laid out here, we will go rather quickly, and it will be helpful if you have had some exposure to these ideas before.

By a *field* we will mean a collection F of complex numbers with the following properties:

- Both 0 and 1 are in F.
- F is *closed under addition and multiplication*, that is, if a and b are in F, then so are $a+b$ and ab.
- F has *inverses* of nonzero elements, that is, if a is in F and $a \neq 0$, then $1/a$ is in F.

A field can be thought of as a set of numbers in which the basic operations of arithmetic can be performed. You can add, subtract, multiply, and divide inside F without ever leaving F. Most of these notes will be devoted to studying the properties of fields in an attempt to answer the questions about compass-and-straightedge constructions discussed in the introduction.

- All the complex numbers themselves form a field, called $\mathbb{C}$.
- The rational numbers, $\mathbb{Q}$, and the real numbers, $\mathbb{R}$, both form fields.
- Consider the set of all numbers of the form $a+b\sqrt{3}$ where a and b are rational numbers. This set is clearly closed under addition and contains 0 and 1. Since $(a+b\sqrt{3})(c+d\sqrt{3}) = (ac+3bd)+(bc+ad)\sqrt{3}$, it's also closed under multiplication. You can find an inverse for a number of the form $a+b\sqrt{3}$ using "conjugates":

  $$\frac{1}{a+b\sqrt{3}} \cdot \frac{a-b\sqrt{3}}{a-b\sqrt{3}} = \frac{a-b\sqrt{3}}{a^2-3b^2}.$$

  Since the denominator is rational, this number is in our set. So our set forms a field, which we call $\mathbb{Q}(\sqrt{3})$.
- There was nothing special about the number 3 in the previous example other than the fact that it isn't the square of a rational number. We can form a field $\mathbb{Q}(\sqrt{d})$ in an analogous way for any rational number d. If d is a square, this will be the same as $\mathbb{Q}$, otherwise it will be bigger.

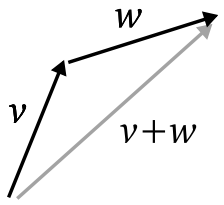
Given a field F, a *vector space over* F is a set V with an *addition law* (a way of "adding" two elements v and w to get a new element $v+w$) and a *scalar multiplication law* (a way of "multiplying" elements v by numbers a in F to get a new element av) with the following properties:

- Addition is *commutative* and *associative*, that is, $v+w=w+v$ and $v+(w+x)=(v+w)+x$.
- There's some element 0 in V with the property that $0+v=v$ for every v, and every element v in V has a *negative*, that is, some element $-v$ so that $v+(-v)=0$.
- Scalar multiplication is *associative*: $a(bv)=(ab)v$ where a and b are in F and v is in V.
- Scalar multiplication is *distributive* over addition, that is, $a(v+w)=av+aw$ and $(a+b)v=av+bv$.

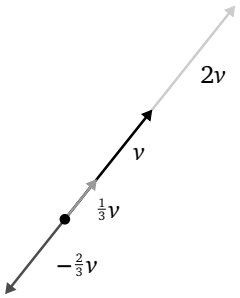
- $1 \cdot v = v$ for every element v.

Before moving on, we give a few examples:

- A field is always a vector space over itself, where addition is normal addition and scalar multiplication is normal multiplication.
- $\mathbb{C}$ is a vector space over $\mathbb{R}$ with the usual operations.
- If F is a field and n is some positive integer, we can form the set F^n of all sequences of n elements of F. For example, if $F = \mathbb{C}$ and $n = 3$, then $(4, \sqrt{2}, 0)$ is an element of $\mathbb{C}^3$. These can be added by just adding the first element to the first element, the second to the second, and so on. (So in $\mathbb{C}^3$, for example, $(1,2,3)+(0,4,-3) = (1,6,0)$.) We can define scalar multiplication by multiplying every number in the sequence by the scalar $(5(1,2,3) = (5,10,15))$.
- If $F = \mathbb{R}$ and n is 2 or 3, there is a geometric interpretation of $\mathbb{R}^n$. The numbers in the sequence can be thought of as specifying a point in the plane or in space, and the addition as the "vector addition" that may be familiar from physics:



The scalar multiplication "stretches" the vector or, if the scalar in negative, flips it around:



- Consider the collection of all polynomials with coefficients in some chosen field F. This is a vector space over F: polynomials can be added to each other and multiplied by elements of F, and the coefficients will always stay in F by the definition of a field.
- Similarly, we can just take, for example, all polynomials of degree at most 6. You can't increase the degree of a polynomial by adding them or multiplying by a constant, so this works just as well.

There is a lot that can be said about the properties of vector spaces, and a course in linear algebra would cover a lot of those things. For the purpose of these notes, however, we only need

one of those things: the notion of the *dimension* of a vector space. Before we can define that, we have to discuss a couple of related ideas.

Suppose you have a collection of elements $\{v_1, v_2, \ldots, v_n\}$ in some vector space V. A *linear combination* of these elements is something of the form $a_1v_1 + a_2v_2 + \cdots + a_nv_n$ where the a_i's are in F. If the collection has just one element, then all linear combinations are just scalar multiples of that one element. In the "geometric" vector spaces $\mathbb{R}^2$ and $\mathbb{R}^3$ mentioned above, linear combinations have a corresponding geometric interpretation: the linear combinations of two vectors v and w in $\mathbb{R}^3$, for example, consist of all the points on the plane containing v, w, and 0. For example, the linear combinations of $(1,0,0)$ and $(0,1,0)$ are all vectors of the form $a(1,0,0)+b(0,1,0)=(a,b,0)$, that is, the xy-plane. The span of a single vector in $\mathbb{R}^2$ or $\mathbb{R}^3$ is just the line passing through that vector and 0.

The collection of all linear combinations of $\{v_1, v_2, \ldots, v_n\}$ is called the *span* of the set, and if this span is the entire vector space V, we say that the set *spans* V.

- For the vector space F over F, any collection of elements spans as long as it contains a nonzero element: if a is nonzero and it's in the set, then any element b in F can be written as $\frac{b}{a}a$.
- The collection of vectors $\{(1,0),(0,1),(1,3)\}$ spans F^2 for any field F: we can write (a,b) as $a(1,0)+b(0,1)$. Notice that the third element doesn't help at all, but it does no damage to the spanning property.
- The collection $\{1, x, x^2, x^3\}$ spans the vector space of polynomials of degree at most 3, since any such polynomial is of the form ax^3+bx^2+cx+d (and $d = d \cdot 1$).
- The collection $\{x+1, x^2+1, x^2+x\}$ spans the vector space of polynomials of degree at most 2: we can, after solving a few simulatenous linear equations, write ax^2+bx+c as $\frac{1}{2}(b+c-a)(x+1)+\frac{1}{2}(a-b+c)(x^2+1)+\frac{1}{2}(a+b-c)(x^2+x)$.

Saying that a collection of elements spans V is saying that, in some sense, the collection is "big enough" to reach every element of the vector space just using the operations of scalar multiplication and addition. There is a notion of smallness corresponding to this notion of bigness: we say that the collection $\{v_1, v_2, \ldots, v_n\}$ if the only way to get a linear combination $a_1v_1 + a_2v_2 + \cdots a_nv_n$ to equal 0 (the element of V) is to have each of the a_i's equal to 0 (the element of F). Put another way, if you have such a linear combination and at least one of the coefficients is not 0, then the linear combination can't be 0. (We might sometimes say that no *nontrivial* linear combination is equal to 0; this is what we mean.)

Linear independence means that, as far as counting linear combinations is concerned, your set is not redundant, in the following sense. Suppose your set is $\{v, w\}$ and $av+bw=0$. Then for any linear combination of v and w, say $y = sv+tw$, we can add our expression for 0 to get another way of writing y: $y = (sv+tw)+(av+bw)=(s+a)v+(t+b)w$. If either a or b isn't zero, then this is a different way of expressing y as a linear combination of v and w. Linear dependence is precisely saying that this never happens. If your set is linearly independent, then every linear combination of those elements is a linear combination in exactly one way, no more: if $\sum a_iv_i = \sum b_iv_i$, then by distributivity, $\sum(a_i-b_i)v_i=0$, and if the set of v_i's is linearly independent, then each a_i must be equal to the corresponding b_i.

Of the examples above, the third and the fourth are linearly independent, and the second isn't. For the first, the collection is linearly independent if and only if it just contains a single nonzero element.

A collection of elements is called a *basis* (plural *bases*, pronounced /BAY-sees/) if it spans and it is linearly independent. A basis is essentially a system of coordinates for the vector space: there will be a unique way of writing every element of the vector space as a linear combination of basis elements, so it sets up a one-to-one correspondence between sequences of coefficients and elements of the vector space. (The spanning property makes sure that you hit every element, and the linear independence property ensures that the sequence of coefficients is unique.) For what follows, there's just one important fact about linear independence and spanning that we need to establish:

Proposition. *If $\{v_1, \ldots, v_n\}$ is a linearly independent set and $\{w_1, \ldots, w_m\}$ is a spanning set, then $n \leq m$. In other words, no linearly independent set is larger than any spanning set.*

Proof. We will show that we can replace one of the v_1's with one of the w_i's and keep the set linearly independent. If we keep replacing elements in this way one by one until every v_i is one of the w_i's, then the first set will be a subset of the second, and the result will be proved.

We claim that one of the w_i's is not a linear combination of $\{v_2, v_3, \ldots, v_n\}$. To see this, suppose they all are. Then, since the w_i's span, we can write v_1 as a linear combination of the w_i's. But since each w_i is a linear combination of the remaining v's, we get that v_1 is a linear combination $\{v_2, v_3, \ldots, v_n\}$ as well, that is, $v_1 = a_2 v_2 + \cdots + a_n v_n$, where the a_i's are in F. But then $-1 \cdot v_1 + a_2 v_2 + \cdots + a_n v_n = 0$, and this contradicts the fact that the v's were linearly independent.[1]

Then we claim that $\{w_i, v_2, \ldots, v_n\}$ is still linearly independent. To see this, suppose we have $a_1 w_i + a_2 v_2 + a_3 v_3 + \cdots a_n v_n = 0$. If $a_1 = 0$, then we actually have a linear combination of v_i's equal to 0, which means all the rest of the a_i's have to be zero. Otherwise, we have $w_i = \frac{a_2}{a_1} v_2 + \frac{a_3}{a_1} v_3 + \cdots + \frac{a_n}{a_1} v_n$, but w_i wasn't supposed to be a linear combination of the v_i's, so this is impossible.

We've now shown everything we set out to show in the first paragraph, so the result follows. □

Corollary. *All bases of a vector space have the same size.*

Proof. If $\{v_1, \ldots, v_n\}$ and $\{w_1, \ldots, w_m\}$ are both bases of a vector space V, then since the v's are linearly independent and the w's span, we have $n \leq m$. And since the w's are linearly independent and the v's span, we have $m \leq n$. So $m = n$. □

This brings us to the final definition of this section. Since all bases of a vector space have the same size, that size is a number that depends only on the vector space and not on whatever individual basis we might be looking at. That number — the size of any basis of the vector space — is called the *dimension* of the vector space, written $\dim(V)$. Some vector spaces don't have any finite bases at all. In this case, we write $\dim(V) = \infty$ and say that V is *infinite-dimensional*.

Exercises

1. Do the polynomials $1 + x$, $1 - x^3$, $2x + x^4$, and $x^2 + x^3 + 5x^4$ span the vector space (over $\mathbb{R}$) of polynomials with real coefficients of degree at most 4?

2. What is the dimension of the vector space of polynomials of degree at most d with coefficients in $\mathbb{C}$ as a vector space over $\mathbb{R}$ (not over $\mathbb{C}$)? Find a basis.

[1] To make this argument run, you actually need to prove that $-1 \cdot v = -v$, a fact that looks simple, but actually requires a short argument. One such argument is that $-1 \cdot v + v = (-1 + 1)v = 0v = 0$, so $-1 \cdot v$ is the additive inverse of v. This, in turn, requires the similarly simple fact that $0 \cdot v = 0$, which can be proved using a similar distributivity argument. Such concerns are not the focus of these notes.

3. In the vector space $\mathbb{R}^3$ over $\mathbb{R}$, give an example of each of the following things or explain why one can't exist:

   - A collection of three vectors which is linearly independent but doesn't span.
   - A collection of four vectors which spans but isn't linearly independent.
   - A collection of two vectors which is linearly independent and spans.
   - A collection of two vectors which neither spans nor is linearly independent.

4. Just as $\mathbb{C}$ is a vector space over $\mathbb{R}$, $\mathbb{R}$ is a vector space over $\mathbb{Q}$. What is its dimension?

5. Given some collection of elements $v_1, \ldots, v_n$ of a vector space V, show that the span of $\{v_1, \ldots, v_n\}$ is a vector space over the same field, call it W. If $\{v_1, \ldots, v_n\}$ was a linearly independent set in V, show that $\dim(W) = n$.

6. Take a vector space V. Adapt the proof the proposition to show that, as long as you have a basis for V, any spanning set contains a basis.

3 Field Extensions and Degree

We can now start building up the concepts that will be used directly in our examination of the geometric questions discussed in the introduction. The first of these tools is the concept of a *field extension*. The definition is very simple: a field extension is simply a pair of fields with one contained inside the other. If the fields are E and F, with E containing F, we will sometimes say that E *is an extension of* F and we might write the extension as E/F or E *over* F. (The slash is meant to suggest the word "over" rather than any sort of division.)

The central observation to make here is that if E/F is a field extension, then E is actually a vector space over F. Indeed, elements of E can be added to each other and multiplied by elements of F. The dimension of this vector space is called the *degree* of the extension, written $[E:F]$. If the vector space is infinite-dimensional, we'll write $[E:F]=\infty$. Field extensions will be the only vector spaces we consider for the rest of these notes.

- Any field F is an extension of itself. Since $\{1\}$ is a basis for F over F (as is any other nonzero element of F), we have $[F:F]=1$.
- The complex numbers $\mathbb{C}$ form an extension of the real numbers $\mathbb{R}$. Since $\{1,i\}$ is a basis, we have $[\mathbb{C}:\mathbb{R}]=2$.
- The field $\mathbb{Q}(\sqrt{3})$ discussed in the previous section forms a degree-2 extension of $\mathbb{Q}$, since $\{1,\sqrt{3}\}$ is a basis.
- Consider the set of numbers of the form $a+b\sqrt{2}$ with a and b in $\mathbb{Q}(\sqrt{3})$. The same proof that we used to show that $\mathbb{Q}(\sqrt{3})$ is a field will show that this is a field as well, written $\mathbb{Q}(\sqrt{3},\sqrt{2})$. For the same reason as in the previous example, we have $[\mathbb{Q}(\sqrt{3},\sqrt{2}):\mathbb{Q}(\sqrt{3})]$.
- We can try to form a field out of a cube root like we did for square roots, say $\sqrt[3]{2}$. But we find that if we just try to take numbers of the form $a+b\sqrt[3]{2}$, we lose closure under multiplication, since $(\sqrt[3]{2})^2$ is not of this form. It's enough, though, if we through it in: numbers of the form $a+b\sqrt[3]{2}+c(\sqrt[3]{2})$ do form a field, called $\mathbb{Q}(\sqrt[3]{2})$. The only difficult part to prove is the existence of inverses. While it's possible to come up with an explicit formula like we did for $\mathbb{Q}(\sqrt{3})$, it's going to be much easier to make use of some basic facts about polynomials which we're about to prove. Once this is done, though, we will have shown that $\mathbb{Q}(\sqrt[3]{2})$ is a degree-3 extension of $\mathbb{Q}$.

We can look at the problem of finding inverses of elements of $Q(\sqrt[3]{2})$ as a problem about manipulating polynomials in the following way. To any element $u=a+b\sqrt[3]{2}+c(\sqrt[3]{2})^2$ we can associate a polynomial $f=a+bx+cx^2$. Then to find an inverse for u, it's enough to find two polynomials p and q with rational coefficients for which $p(x)f(x)=1+(x^3-2)q(x)$. This is because, if we then plug in $x=\sqrt[3]{2}$, we get that $p(\sqrt[3]{2})u=1$, and so $p(\sqrt[3]{2})$ is our inverse. (Notice that $p(\sqrt[3]{2})$ is in our field because we already know $\mathbb{Q}(\sqrt[3]{2}$ is closed under addition and multiplication.)

First, we need a couple more definitions. Given two polynomials f and g, we say that f *divides* g, and write $f|g$, if there's some other polynomial h with $g=fh$. (That is, $f|g$ if g is a multiple of f.) For example, $(x+1)|(x^2-1)$, since $x^2-1=(x+1)(x-1)$, but $x\nmid x^3-7$, since any multiple of x has 0 as a constant term. Notice that a nonzero constant polynomial a divides any polynomial f at all, since $f=a\cdot(\frac{1}{a}f)$. Similarly, if a is a nonzero constant, then af divides f, since $f=\frac{1}{a}\cdot(af)$.

Suppose we have picked a field F. We say that a polynomial g is *irreducible over* F if the only polynomials with coefficients in F that divide g are constants and constant multiples of g. Irreducible polynomials can be thought of as the polynomial analogues of prime numbers.

We consider some examples:

- Any linear polynomial is irreducible.
- The choice of the field F can affect whether or not g is irreducible. The polynomial x^2-2 is irreducible over $\mathbb{Q}$ (which we'll talk about again momentarily), but not over $\mathbb{R}$, since we can write $x^2-2=(x+\sqrt{2})(x-\sqrt{2})$.
- If $\deg(g)\leq 3$, then g is irreducible over some field F if and only if it doesn't have any roots in F. This is because, if $g=hk$ for some polynomials h and k, and neither h nor k is a constant, then one of them has to have degree 1, say $h=ax+b$. But then $-\frac{b}{a}$ is a root of h, and therefore a root of g.
- If $\deg(g)\geq 4$, though, it's possible for g to be reducible over some field but still not have any roots in that field. For example, $g=x^4+5x^2+4=(x^2+1)(x^2+4)$ is reducible even over $\mathbb{Q}$, but its only roots are i, $-i$, $2i$, and $-2i$, and none of these are in $\mathbb{Q}$.

With these ideas in hand, we prove some results that will, among other things, give us the existence of inverses in $\mathbb{Q}(\sqrt[3]{2})$.

Proposition (The Division Algorithm)**.** *Let F be a field. Given any two polynomials f and g with coefficients in F, there are polynomials q and r, also with coefficients in F, so that $f=gq+r$ and* $\deg(r)<\deg(g)$. *We call q the* quotient *and r the* remainder *of the division of f by g.*

Proof. Among all polynomials of the form $f-gq$, where q is any polynomial with coefficients in F, let r be one with the smallest possible degree. Suppose that $\deg(r)\geq\deg(g)$, say $\deg(r)=m$ and $\deg(g)=n$. If $r=r_0x^m+\cdots$ and $g=g_0x^n+\cdots$, then we can see that $(\frac{r_0}{g_0}x^{m-n})g$ has leading term r_0x^m, and therefore $r-(\frac{r_0}{g_0}x^{m-n})g$ has degree smaller than m, since we've cancelled the leading term. But this polynomial is equal to $f-gq-(\frac{r_0}{g_0}x^{m-n})g$, and so it is a polynomial of degree smaller than $\deg(r)$ of the form $f-gq'$, which contradicts our choice of r. □

Corollary. *Suppose F is a field and f is irreducible over F. If g is a polynomial with coefficients in F and* $\deg(g)<\deg(f)$*, then there are polynomials h and k, also with coefficients in F, with $hf+kg=1$.*

Proof. Among all nonzero polynomials of the form $hf+kg$ where p and q have coefficients in F, let m be one of the smallest degree. Using the Division Algorithm, write $f=qm+r$ with $\deg(r)<\deg(m)$. Then we have to have $r=0$, because otherwise, $r=f-qm=(h+1)f+(k-q)g$ is a nonzero polynomial of smaller degree than m. So in fact, $f=qm$, meaning that $m|f$. Similarly, $m|g$. But since $\deg(g)<\deg(f)$, m can't be a constant multiple of f so, since f is irreducible, the only other option is that m is a constant. We can then write $\frac{h}{m}f+\frac{k}{m}g=1$. □

So in order to find inverses in $\mathbb{Q}(\sqrt[3]{2})$, all we need to do is prove that x^3-2 is irreducible over $\mathbb{Q}$. But if it were reducible, then, by the discussion about irreducibility of degree-3 polynomials above, x^3-2 would have a rational root, that is, there would be a rational cube root of 2, and this impossible.[2]

How was the polynomial x^3-5 chosen for the proof of the existence of inverses in $\mathbb{Q}(\sqrt[3]{2})$? The relevant facts were that (1) it's 0 when you plug in $x=\sqrt[3]{2}$, (2) it's irreducible, and (3) every element

[2]It's possible to show this in the same way as one shows that the square root of 2 is irrational: suppose $(p/q)^3=2$ for integers p and q with no common factors and derive a contradiction.

of $\mathbb{Q}(\sqrt[3]{2})$ can be expressed as a polynomial with rational coefficients with $x = \sqrt[3]{2}$. Whenever we have a situation like this, we can apply the exact same proof to conclude the existence of inverses. For example, the polynomial $x^2 - 3$ is irreducible for the same reason as $x^3 - 2$ (if it factors, it has to have a factor of degree 1 and therefore a root, but there's no rational square root of 3) so the same argument can be used to find inverses in $\mathbb{Q}(\sqrt{3})$ without the use of the conjugation trick we employed in the last section.

This technique can be used to build and analyze a large class of field extensions, including the ones we'll be using to study compass-and-straightedge constructions. Given some field F, we say that some complex number a is *algebraic over* F if there's some polynomial g with coefficients in F for which $g(a) = 0$. For example, we've shown that $\sqrt[3]{2}$ is algebraic over $\mathbb{Q}$, since we can take $g = x^3 - 2$. Similarly, the number i is algebraic over $\mathbb{R}$ (and also over $\mathbb{Q}$), since we can take $g = x^2 + 1$. In particular, any element a of F is algebraic over F, since it satisfies $x - a = 0$.

Suppose E is an extension of F and $[E : F]$ is finite, that is, E is a finite-dimensional vector space over F. Then any element a of E is algebraic over F. To see this, take the list of elements $1, a, a^2, \ldots, a^n$, where $n = [E : F]$. Since this list contains $n + 1$ elements, it can't be linearly independent, so some linear combination of these powers of a has to be 0. This gives a polynomial with coefficients in F which is 0 at a, so a is algebraic.

There is a very important fact about algebraic numbers that we will use several times in the material that follows:

Proposition. *Given a field F and a number a which is algebraic over F, there is a unique monic polynomial (that is, a polynomial with leading coefficient 1), called the* minimal polynomial of a *and written p_a, which is irreducible, has coefficients in F, and satisfies $p_a(a) = 0$.*

Proof. Since a is algebraic over F, there's some polynomial with coefficients in F which takes a to 0. If g is such a polynomial, then dividing g by its leading coefficient doesn't stop it from being 0 when you plug in a, so there's also a monic polynomial with this property. Let p_a be a nonzero monic polynomial of the smallest possible degree which satisfies $p_a(a) = 0$. We need to show that p_a is irreducible and that it's the only polynomial with the required properties.

First, suppose p_a is reducible, say $p_a = hk$ where h and k are both non-constant. Then both $\deg(h)$ and $\deg(k)$ are smaller than $\deg(p_a)$, and by multiplying both of them by the right constants, we can assume that they're both monic. But then, since $0 = p_a(a) = h(a)k(a)$, either $h(a)$ or $k(a)$ has to be 0, which contradicts our choice of p_a. So in fact, p_a is irreducible.

Suppose g is some other monic irreducible polynomial with $g(a) = 0$. Then use the Division Algorithm to write $g = qp_a + r$ with $\deg(r) < \deg(p_a)$. Then plugging in a gives us that $r(a) = 0$ so, since p_a is the smallest-degree nonzero polynomial which is 0 at a, we have to have $r = 0$. So $p_a | g$. Since g is irreducible, and p_a isn't a constant, g has to be a constant multiple of p_a which, since they're both monic, means $g = p_a$. □

Because of the technique we demonstrated for $\mathbb{Q}(\sqrt[3]{2})$, minimal polynomials are a very powerful tool for constructing field extensions. Given any field F and any number a which is algebraic over F, we can build a field called $F(a)$ which is the smallest extension of F containing a. Simply take all numbers of the form $b_0 + b_1 a + b_2 a^2 + \cdots + b_{n-1} a^{n-1}$ where the b_i's are elements of F and $n = \deg(p_a)$. We will write $F(a, b)$ for $(F(a))(b)$, $F(a, b, c) = (F(a, b))(c)$, and so on.

We first need to show that $F(a)$ is a field. The proof goes almost exactly the same as the proof for $\mathbb{Q}(\sqrt[3]{5})$. It's clearly closed under addition. To show that it's closed under multiplication, take any two elements and multiply them. You'll get some linear combination of powers of a with coefficients in F, that is, something of the form $g(a)$ for some polynomial g with coefficients in F. Using the Division Algorithm, write $g = qp_a + r$ with $\deg(r) < n$; then $g(a) = r(a)$, and $r(a)$ is

in $F(a)$. Finally, we can get inverses in $F(a)$ using the same trick as before: any element of $F(a)$ is $h(a)$ for some polynomial h with coefficients in F, and, writing $kh + l\,p_a = 1$, we see that $k(a)$ gives our inverse for $h(a)$.

In addition, we can see that $[F(a) : F] = n$. The basis we will use is $\{1, a, a^2, \ldots, a^{n-1}\}$. We already know that it spans $F(a)$ by definition, so it's enough if we can show that it's linearly independent. But if some linear combination of those powers of a were 0, then that would mean some polynomial of degree less than n is 0 when evaluated at a, and that's impossible, since the minimal polynomial has degree n.

There is only one more fact about field extensions that we need to establish: what happens to degrees of extensions of extensions? The answer turns out to be very simple.

Proposition. *Suppose K, E and F are fields, where K is an extension of E and E is an extension of F, and both extensions have finite degree. Then* $[K : F] = [K : E][E : F]$.

Proof. First we make up some notation. Write $d = [E : F]$ and $e = [K : E]$. Suppose $\{a_1, a_2, \ldots, a_d\}$ is a basis for E as a vector space over F and $\{b_1, b_2, \ldots, b_e\}$ is a basis for K over E. We claim that set of all products of the form $a_i b_j$ forms a basis for K over F, and since there are de such products, this will finish the proof.

Any element x of K can be written as $\sum_{i=1}^{e} s_i b_i$ where the s_i's are elements of E. But then each s_i can be written as $\sum_{j=1}^{d} t_{ij} a_j$, so in fact, $x = \sum_{i=1}^{e} \sum_{j=1}^{d} t_{ij} a_j b_i$. This shows that our set spans. To see that it's linearly independent, suppose you had some linear combination $\sum_{i=1}^{e} \sum_{j=1}^{d} t_{ij} a_j b_i = 0$. We can write this sum as $\sum_{i=1}^{e} \left(\sum_{j=1}^{d} t_{ij} a_j \right) b_i$, and since the b_i's are linearly independent, this means that each individual sum $\sum_{j=1}^{d} t_{ij} a_j$ has to be 0. But since the a_j's are linearly independent, this means that each t_{ij} has to be 0, which is what we wanted. □

Some examples:

- We already showed that $[\mathbb{Q}(\sqrt[3]{2}) : \mathbb{Q}] = 3$. By the result we just proved, $[\mathbb{Q}(\sqrt[3]{2}) : \mathbb{Q}(a)]$ has to be either 1 or 3. If it's 1, we get that $\mathbb{Q}(a) = \mathbb{Q}(\sqrt[3]{2})$. Otherwise, $[\mathbb{Q}(a) : \mathbb{Q}] = 1$, meaning that a was actually rational.

- Take $a = \sqrt{2} + \sqrt{3}$. It's simple to check that, if $f(x) = (x^2 - 5)^2 - 24$, then $f(a) = 0$. We claim that this is the minimal polynomial of a over $\mathbb{Q}$. There are two ways we could go about doing this. One would be to show, by checking all possible ways that f could factor, that f is irreducible. But a more conceptual method would be to prove that $[\mathbb{Q}(a) : \mathbb{Q}] = 4$. Then, since the $\deg(f) = 4$, it follows that f is the minimal polynomial of a, and in particular that f is irreducible.

  We already know by definition that $\deg(p_a) \leq \deg(f)$, and so $[\mathbb{Q}(a) : \mathbb{Q}] \leq 4]$. In fact, $a^2 - 5 = \sqrt{6}$, and $\frac{1}{a} = \sqrt{3} - \sqrt{2}$, which means that $\frac{1}{2}(a + \frac{1}{a}) = \sqrt{3}$ and $\frac{1}{2}(a - \frac{1}{a}) = \sqrt{2}$. So $\mathbb{Q}(a)$ is an extension of $\mathbb{Q}$, of degree at most 4, which contains $\mathbb{Q}(\sqrt{2})$, so its degree over $\mathbb{Q}$ is either 2 or 4, depending on whether it is equal to $\mathbb{Q}(\sqrt{2})$ or bigger. But, for example, $\mathbb{Q}(\sqrt{2})$ doesn't contain $\sqrt{3}$: the square of a general element of $\mathbb{Q}(\sqrt{2})$ looks like $(a + b\sqrt{2})^2 = a^2 + 2b^2 + 2ab\sqrt{2}$. If this is rational, then $2ab = 0$, so it's either a^2 or $2b^2$. In particular, it's not 3.

- The minimal polynomial of $\sqrt{2} + \sqrt{3}$ over $\mathbb{Q}(\sqrt{2})$ is different: the polynomial f above factors as $((x - \sqrt{2})^2 - 3)((x + \sqrt{2})^2 - 3)$, and the first of these factors is the new minimal polynomial. (The second is the minimal polynomial of $-\sqrt{2} + \sqrt{3}$.) This agrees with the calculation that $[\mathbb{Q}(\sqrt{2} + \sqrt{3}) : \mathbb{Q}(\sqrt{2})] = [\mathbb{Q}(\sqrt{2} + \sqrt{3}) : \mathbb{Q}]/[\mathbb{Q}(\sqrt{2}) : \mathbb{Q}] = 2$.

There is an interesting consequence of this fact which, though we won't need it in what follows, is still worth mentioning. Given any field F, the set of all numbers which are algebraic over F forms a field. To see this, it's enough to take elements a and b which are algebraic over F and show that $a+b$, ab, and $1/a$ are also algebraic over F. Consider the field $E = F(a,b)$. Since b is algebraic over F, it's certainly algebraic over $F(a)$, so both extensions $[E : F(a)]$ and $[F(a) : F]$ have finite degree. So by the proposition, $[E : F]$ has finite degree. Therefore, every element of E is algebraic over F. But since E is a field which contains both a and b, these elements include $a+b$, ab, and $1/a$.

Exercises

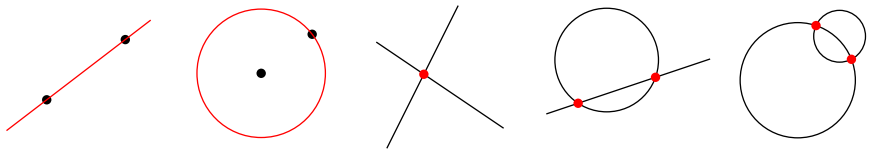
1. Prove that if f is some polynomial and $f(a) = 0$ for some number a, then $(x-a)|f$. Prove that a polynomial of degree n has at most n roots over any field.
2. Given a field F, show that any extension of F of degree 2 is of the form $F(\sqrt{a})$ for some element a of F.
3. Prove that $\sqrt[3]{2}$ is not an element of the field $\mathbb{Q}(\sqrt{3}, \sqrt{7}, \sqrt{19}, \sqrt{26}, \sqrt{\frac{11}{5}})$.
4. What is the minimal polynomial of $\sqrt{2+\sqrt{3}}$ over $\mathbb{Q}$? Over $\mathbb{Q}(\sqrt{3})$?
5. For a positive integer n, an *n'th root of unity* is some complex number a with $a^n = 1$. If p is prime and ζ is an p'th root of unity other than 1, find $[\mathbb{Q}(\zeta) : \mathbb{Q}]$.
6. Suppose F is a field and $[F(a) : F]$ is odd for some number a. Prove that a is in $[F(a^2) : F]$. Provide a counterexample when $[F(a) : F]$ is even.
7. Suppose f is an irreducible polynomial of degree n and g is any polynomial, both with coefficients in some field F. Prove that the degree of every irreducible factor of $f(g(x))$ is a multiple of n.

4 Constructible Numbers

We now finally come to the discussion of compass-and-straightedge constructions. In the introduction you were promised proofs that certain operations are impossible to perform with a compass and straightedge. In order to prove this, we first need to establish exactly what types of operations are permitted in a classical Greek compass-and-straightedge construction (which, to save typing and reading, we'll just call *constructions* from here on).

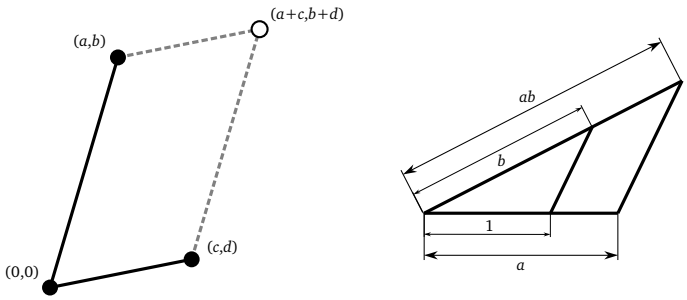
A construction consists of three types of objects: points, lines, and circles. The allowable operations are as follows:

1. Given two points, draw (with your straightedge) a line which passes through both of them.
2. Given two points, draw (with your compass) a circle whose center is the first point and which contains the second point
3. Given two lines, or two circles, or a line and a circle, draw a point anywhere where they intersect.



Of course, none of this can even get off the ground unless you start with two points drawn on the paper, so assume that that happens. We'll pick units so that the distance between these two points is 1. We can also pick a coordinate system for the plane which agrees with these units and for which one of the points is the origin and the other is $(0,1)$. Using this coordinate system, we call a number a *constructible* if there is some construction which ends up drawing a point whose x- or y-coordinate is a. The rest of this section will be concerned with figuring out which numbers are constructible and showing that accomplishing any of the tasks mentioned in the introduction would involve constructing a point with non-constructible coordinates.

First, we observe that the constructible numbers form a field. First, it's possible to add and subtract coordinates of points by constructing a parallelogram as in the diagram below on the left. Negatives can be found by simply flipping one coordinate or the other around an axis. Products can be taken using a process like the one in the diagram below on the right, and that same method, after switching the roles of b and ab, can be used to take quotients.



In any construction, we start out with just the numbers 0 and 1 as coordinates of points we have constructed. As we proceed, we consider the field that you get by starting with $\mathbb{Q}$ and adding

in coordinates of points we've constructed so far. So if at some stage in the construction we have the field F and we then construct a point with coordinates (a,b), we replace F with $F(a)$, and then with $F(a,b)$.

The only way to add numbers to the field is to take intersections of lines and/or circles, so we consider what happens when you do so. Any line is a line passing through two points that have been constructed, say (a,b) and (c,d). The line connecting these two points has the formula $(y-b)(c-a)=(x-a)(d-b)$. In particular, all the coefficients are already in our field. Similarly, the formula for a circle with center (a,b) and containing the point (c,d) is $(x-a)^2+(y-b)^2=(c-a)^2+(d-b)^2$. Here again all the coefficients are in our field already.

The coordinates of our new point will come from a simultaneous solution of two equations of this kind. If one of the equations is a line, it will be possible to solve for one of the variables in terms of the other. Plugging that into the other equation will give a quadratic or linear polynomial in the remaining variable, so in this case, the new field has degree 1 or 2 over the old one. In the case of two circles, say

$$\begin{aligned}(x-a)^2+(y-b)^2 &= R\\ (x-a')^2+(y-b')^2 &= R'\end{aligned}$$

we can subtract these two equations to get

$$\begin{aligned}(x-a)^2+(y-b)^2 &= R\\ 2(a'-a)x+2(b'-b)y &= R-a^2-b^2-R+a'^2+b'^2\end{aligned}$$

and we can use the last equation, which is now linear, to again solve for one of the variables in terms of the other and proceed as before.

Either way, we get that either we didn't enlarge the field at all, or that the new field has degree 2 over the old one. This means, by the multiplicativity of degrees of extensions that we proved in the last section, that the degree of the new field over $\mathbb{Q}$ is always a power of 2. If a is a constructible number, then $\mathbb{Q}(a)$ will be contained in one of these fields, and so, again by that multiplicativity, $[\mathbb{Q}(a):\mathbb{Q}]$ has to divide a power of 2, that is, it is a power of 2. In particular, the degree of the minimal polynomial of any constructible number has to be a power of 2.

Perhaps surprisingly, this is the only thing we need to prove what we set out to prove at the beginning of these notes. The easiest of the three problems to handle is doubling the cube. If it were possible to double the cube, then it would be possible to construct a square whose length is $\sqrt[3]{2}$. We've already shown that the minimal polynomial of this number is x^3-2, and since 3 isn't a power of 2, $\sqrt[3]{2}$ isn't constructible.

Only slightly more challenging is the problem of trisecting an angle. Some angles can be trisceted. For example, it's possible to trisect an angle of 90°, since it's possible to construct an angle of 30°. But we can show that it is not possible to trisect an angle of 60° by showing it's not possible to construct 20°. If it were possible, then $\cos(20°)$ would be a constructible number: you can construct a line perpendicular to any given line, so you would be able to build a right triangle in which one of the angles is 20°.

We can use the angle addition formula $\cos(a+b)=\cos(a)\cos(b)-\sin(a)\sin(b)$ to compare $\cos(20°)$ with $\cos(60°)=\frac{1}{2}$, and we find that, if $\cos(20°)=b$, then $4b^3-3b-\frac{1}{2}=0$ or, setting $a=2b$, $a^3-3a-1=0$. Since this is a polynomial of degree 3, it can only be reducible over $\mathbb{Q}$ if it has a rational root. But if c were a rational root of x^3-3x-1, you can check directly that $c=\pm1$ (this is a special case of the *Rational Root Theorem*), but neither 1 nor -1 is a root. We conclude that this polynomial is irreducible, and therefore that a lives in a degree-3 extension of $\mathbb{Q}$. In particular, it's not constructible.

To talk about squaring the circle, we need one fact that we won't be able to prove in these notes: π is not an algebraic number. If it were possible to square a circle, then the ratio between the side length of the square and the radius of the circle would be $\sqrt{\pi}$, but this number is not algebraic, and therefore certainly not constructible.