FIELD THEORY FOR COMPASS AND STRAIGHTEDGE IMPOSSIBILITY PROOFS

TANNER STRUCK

Abstract. This paper builds up the field theory necessary to prove some famous compass and straightedge impossibility results. Basic knowledge of rings, vector spaces, and geometric constructions with a compass and straightedge is assumed. Theorem 3.3 especially makes use of ring theory. We begin by formalizing the notion of compass and straightedge construction, and then introduce the theory of field extensions. Then, we apply field extensions to compass and straightedge constructions to reveal the field theoretic properties of these constructions. Finally, we show that some geometric constructions violate those field theoretic properties, and are thus impossible to construct with only a compass and straightedge.

Contents

1. Compass and Straightedge Constructions 1
2. A Note on Irreducibility 5
3. Introduction to Field Extensions 5
4. Compass and Straightedge Constructions Revisited 10
5. Impossibility Proofs 13
Acknowledgments 15
References 16

1. Compass and Straightedge Constructions

The treatment of compass and straightedge constructions here follows Chapter 30 of [2] and Chapter 15 of [3].

Even though compass and straightedge constructions were problems of great interest to mathematicians for millennia, many such problems, even ones that appear simple at first glance, went unsolved until the creation of modern algebra. After learning the proofs of such problems, the reason why these problems went unsolved for so long becomes clear. Abstract algebra provides advanced tools for placing restrictions on which compass and straightedge constructions are possible, and this enabled mathematicians to prove once and for all that the constructions they had struggled with were, in fact, impossible. This had gone unproven for so long because the language of abstract algebra is necessary to articulate what these restrictions even are, let alone prove them.

In order to apply algebra to compass and straightedge constructions, we have to put these constructions into an algebraic context. We can think of compass and straightedge constructions as a finite sequence of steps using the compass and

straightedge to create some geometric figure starting from a single line segment of unit length. We imagine all of this taking place in $\mathbb{R}^2$, where the initial unit line segment is the segment connecting $(0,0)$ and $(1,0)$. To formalize the idea of which points can be constructed, we will consider what points can be constructed when starting with a set $\mathcal{M}$ of "marked" points. Each step in the compass and straightedge construction constitutes "marking" a point by appending it to the set of already marked points. The following definition captures which points are "markable" from a set of already marked points.

Definition 1.1. Given a set $\mathcal{M} \subseteq \mathbb{R}^2$, a point $P \in \mathbb{R}^2$ is *constructible in one step from* $\mathcal{M}$ if P is the intersection of two distinct figures where each figure is either:

(1) a line $\overline{AB}$ where $A, B \in \mathcal{M}$.
(2) a circle with radius AB centered on C where $A, B, C \in \mathcal{M}$.

Remark 1.2. The line $\overline{AB}$ for points $A, B \in \mathbb{R}^2$ refers to the line passing *through* points A and B and extending infinitely in both directions, not just the line segment passing between A and B.

We can define a compass and straightedge construction to be a sequence of "marking" points.

Definition 1.3. Given a set $\mathcal{M} \subseteq \mathbb{R}^2$, a point $P \in \mathbb{R}^2$ is *constructible from* $\mathcal{M}$ if there exists a finite sequence of points $P_1, P_2, \ldots, P_n \in \mathbb{R}^2$ such that $P_n = P$ and for each $1 \leq i \leq n$, P_i is constructible in one step from $\mathcal{M} \cup \{P_1, P_2, \ldots, P_{i-1}\}$.

While the above definition is more general, we would like a way to refer to points that can be constructed from nothing but the unit line segment.

Definition 1.4. A point $P \in \mathbb{R}^2$ is *constructible* if P is constructible from $\{(0,0),(1,0)\}$

For the sake of brevity in our proofs, we will assume that the following geometric constructions are possible with only a compass and straightedge: constructing a perpendicular bisector of any given segment, constructing a line that passes through any given point and is parallel to any given line, and bisecting any given angle. Methods for these constructions were known even to the Ancient Greeks.

Thanks to the following proposition, we can actually translate the problem of figuring out which points are constructible in two dimensions into a problem of figuring out which coordinates are possible to construct, which is much simpler to analyze with algebra.

Proposition 1.5. *A point $(a,b) \in \mathbb{R}^2$ is constructible if and only if $(a,0)$ and $(b,0)$ are constructible.*

Proof. Using the straightedge, one can draw the line through $(0,0)$ and $(1,0)$ to construct the x-axis. One can construct $(-1,0)$ and draw a perpendicular bisector through $(-1,0)$ and $(1,0)$ to construct the y-axis.

Suppose $(a,b) \in \mathbb{R}^2$ is constructible. Construct the x and y axes as described and then do the steps necessary to construct (a,b). Construct a line parallel to the y-axis that passes through (a,b) and mark where this line intersects the x-axis. This point is $(a,0)$, so $(a,0)$ is constructible. Now, construct a line parallel to the x-axis that passes through (a,b) and mark where this line intersects the y-axis. This point is $(0,b)$. Now, draw a circle of radius $|b|$ centered on $(0,0)$ and mark where this intersects the x-axis. One of these points is $(b,0)$, so $(b,0)$ is constructible (see Figure 1).

Now, suppose instead that $(a,0),(b,0) \in \mathbb{R}^2$ are constructible. Construct the x and y axes as described and then do the steps necessary to construct $(a,0)$ and $(b,0)$. Draw a circle of radius $|b|$ centered on $(0,0)$ and mark where this intersects the y-axis. One of these points will be $(0,b)$. Construct a line parallel to the y-axis that passes through $(a,0)$. Then, construct a line parallel to the x-axis that passes through $(0,b)$, and mark where it intersects the previous line. This point is (a,b), so (a,b) is constructible (see Figure 1).

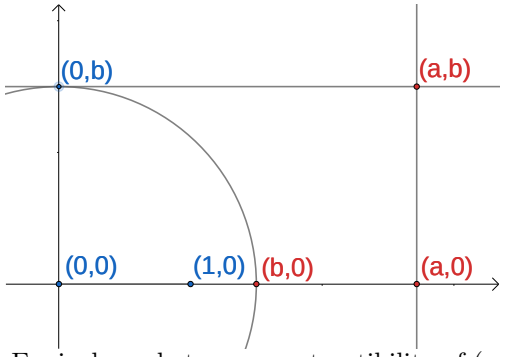

FIGURE 1. Equivalence between constructibility of (a,b) and constructibility of $(a,0)$ and $(b,0)$. Image created with [4].

□

The above proposition allows us to characterize every single constructible point in the plane just by characterizing the constructible points on the x-axis, which motivates the following definition.

Definition 1.6. A real number $x \in \mathbb{R}$ is constructible if $(x,0)$ is constructible.

Now, we can narrow our focus down to studying a subset of $\mathbb{R}$. It happens that this set actually has very good algebraic properties. Recall that a field is a commutative ring with identity $1 \neq 0$ such that every nonzero element has a multiplicative inverse. This means that a field is some set equipped with addition, subtraction, multiplication, and division. Because we can carry out all of these operations with a compass and straightedge (see below), the constructible numbers form a field.

Lemma 1.7. *Let $x \in \mathbb{R}$. Then, x is constructible if and only if $-x$ is constructible.*

Proof. If 0 is constructible, then -0 is constructible. If $x \in \mathbb{R}$ with $x \neq 0$ is constructible, then one can construct $(x,0)$, draw a circle around $(0,0)$ of radius $|x|$, and then mark the remaining intersection with the x-axis to construct $(-x,0)$. Also, by the above, if $-x$ is constructible, then $-(-x) = x$ is constructible. Thus, $x \in \mathbb{R}$ is constructible if and only if $-x$ is constructible. □

Proposition 1.8. *Let $a,b \in \mathbb{R}$ be constructible numbers. Then, $a+b$, $a-b$, and ab are all constructible numbers. If $b \neq 0$, then a/b is a constructible number.*

Proof. Construct $(a,0)$ and $(b,0)$. Draw a circle centered on $(a,0)$ with radius $|b|$ and mark the circle's intersection points with the x-axis. If $b \neq 0$, one intersection

is $(a-b,0)$ and the other intersection is $(a+b,0)$. If $b=0$, then $a-b=a+b=a$, which is constructible. Hence, $a-b$ and $a+b$ are constructible numbers.

Suppose a and b are positive. Since 0, 1, a, and b are constructible numbers, by Proposition 1.5, $(0,1)$, $(0,a)$, and $(0,b)$ are constructible points. First, construct $(a,0)$ and $(0,b)$. Draw the line through $(0,0)$ and $(0,b)$, and draw the line connecting $(0,b)$ and $(1,0)$. Draw a line parallel to the latter line that passes through $(a,0)$. Mark point P where this line intersects the line through $(0,0)$ and $(0,b)$. By similar triangles, the distance between P and $(0,0)$ is a times the length of the distance between $(0,0)$ and $(0,b)$. Therefore, $P=(0,ab)$ (see Figure 2). By Proposition 1.5, ab is a constructible number.

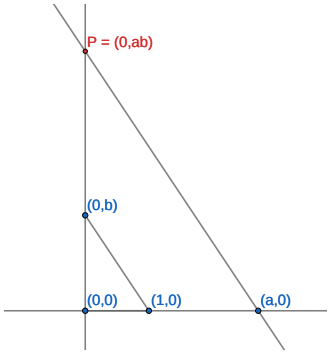

FIGURE 2. Construction of $(0,ab)$. Image created with [4].

Now, still assuming $a,b>0$, construct $(0,a)$, and $(b,0)$. Draw the line connecting $(0,a)$ and $(b,0)$. Construct a line parallel to the previous one that passes through $(1,0)$ and mark the point P where it intersects the line through $(0,0)$ and $(0,a)$. By similar triangles, the distance between P and $(0,0)$ is $1/b$ times the distance between $(0,a)$ and $(0,0)$. Therefore, $P=(0,a/b)$ (see Figure 3). By Proposition 1.5, a/b is a constructible number.

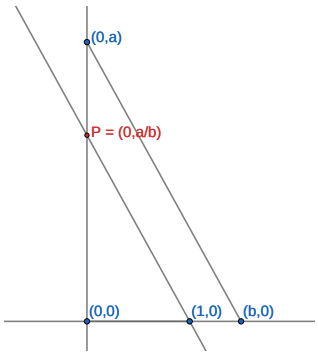

FIGURE 3. Construction of $(0,a/b)$. Image created with [4].

If $a=0$, then $ab=a/b=0$ is constructible. If $b=0$, then ab is constructible. If a or b is negative, then that will only result in changing the sign of ab and a/b, so by Lemma 1.7, ab and a/b are constructible numbers. □

By Proposition 1.8, the constructible numbers form a subfield of the real numbers, which begs to be analyzed through the lens of modern algebra. In order to prove that some compass and straightedge constructions are impossible, we must place some sort of restriction on what can be in the field of constructible numbers. Definition 1.3 suggests that we should investigate the process by which a new point is appended to a set of previously marked points. Since the constructible numbers form a field, this may lead one to wonder if the process by which a point is marked is related to the process by which a field is enlarged to contain a new element (think of enlarging $\mathbb{R}$ to contain i). Definition 1.1 suggests that we should look for a special property that pertains to the intersections of lines and circles. One thing we can already note is that the equations of lines and circles are first and second degree polynomials, which will be helpful to keep in mind as we transition toward our study of fields.

2. A Note on Irreducibility

Irreducible polynomials play an extremely important role in the theory of field extensions and in the impossibility proofs for compass and straightedge constructions. A method for proving that some polynomials are irreducible will be given here.

Definition 2.1. Given a polynomial ring $F[x]$ over a field F, a nonconstant polynomial $p(x) \in F[x]$ is *irreducible in* $F[x]$ if whenever $p(x) = a(x)b(x)$ for $a(x), b(x) \in F[x]$, either $a(x) \in F$ or $b(x) \in F$.

Remark 2.2. This definition essentially states that a polynomial is irreducible when it cannot be written as a product of polynomials of nonzero degree. Furthermore, this definition provides a natural definition of *reducible*, which means not irreducible. Also, note that this is certainly not the most general definition of irreducibility, but for our purposes, it will suffice.

Theorem 2.3 (Eisenstein's Criterion applied to $\mathbb{Z}$)**.** *Let $f(x) \in \mathbb{Z}[x] \subset \mathbb{Q}[x]$ be such that the coefficient of the highest degree term is 1. For some $n \in \mathbb{N}$, $f(x)$ can be expressed in the form*

$$f(x) = x^n + a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \cdots + a_0.$$

Suppose that for all $0 \leq i \leq n-1$, a_i is divisible by some prime number p, but a_0 is not divisible by p^2. Then, $f(x)$ is irreducible in $\mathbb{Q}[x]$.

Proof. For proof, see Chapter 9 Corollary 14 of [1]. □

Example 2.4. Consider $x^2 - 2 \in \mathbb{Q}[x]$. The leading coefficient is 1, every other coefficient is divisible by 2, and the constant term is not divisible by $2^2 = 4$. Therefore, by Theorem 2.3, $x^2 - 2$ is irreducible in $\mathbb{Q}[x]$.

3. Introduction to Field Extensions

The treatment of field theory here follows Chapter 13 of [1].

A desire to study how elements are marked in compass and straightedge constructions may lead one to study how fields are enlarged to contain certain elements. In order to investigate this, we will need to introduce more language to talk about fields and their relationships to one another. When given two fields F and K such

that $F \subseteq K$, we say that F is a subfield of K. We describe the reverse relationship as a field extension.

Definition 3.1. Let F and K be fields. We say K is a *field extension* over F if F is a subfield of K. When K is a field extension over F, we write K/F, which is read as "K over F."

For example, since $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ are all fields and $\mathbb{Q} \subseteq \mathbb{R} \subseteq \mathbb{C}$, by Definition 3.1, $\mathbb{C}$ is a field extension of $\mathbb{R}$, $\mathbb{R}$ is a field extension of $\mathbb{Q}$, and $\mathbb{C}$ is a field extension of $\mathbb{Q}$. Written in symbols, $\mathbb{C}/\mathbb{R}$, $\mathbb{R}/\mathbb{Q}$, and $\mathbb{C}/\mathbb{Q}$ are all field extensions. While the idea of a field extension may initially seem like a mere restatement of the definition of a subfield, we later find that it allows us to turn our attention toward studying the extensions of a particular field, which are distinct from the subfields of that field.

Given a field extension K/F and an element $\alpha \in K$, we want to know what new field is "generated" by appending α to F and "closing" the set under multiplication, addition, and inverses. Specifically, we want to study what happens when you append elements to subfields of the constructible numbers in hopes of understanding the constructible numbers themselves, but there are other reasons why one might want to investigate this. For instance, one may want to study the field generated by appending $\sqrt{2}$ to $\mathbb{Q}$. The following definition captures this idea.

Definition 3.2. Let K be a field extension of F and let $\alpha_1, \alpha_2, \ldots, \alpha_n \in K$. Let $F(\alpha_1, \alpha_2, \ldots, \alpha_n)$ denote the minimal subfield of K such that

$$F \subseteq F(\alpha_1, \alpha_2, \ldots, \alpha_n)$$

and

$$\alpha_1, \alpha_2, \ldots, \alpha_n \in F(\alpha_1, \alpha_2, \ldots, \alpha_n).$$

Using this definition, we can express "the field generated by appending $\sqrt{2}$ to $\mathbb{Q}$" as $\mathbb{Q}(\sqrt{2})$. Let $a, b \in \mathbb{Q}$. By Definition 3.2, $\mathbb{Q}(\sqrt{2})$ is a field and $a, b, \sqrt{2} \in \mathbb{Q}(\sqrt{2})$, so $a + b\sqrt{2} \in \mathbb{Q}(\sqrt{2})$. We will see later that every element of $\mathbb{Q}(\sqrt{2})$ can be expressed in this form. Definition 3.2 also gives another way to express the complex numbers, as $\mathbb{C} = \mathbb{R}(i)$. Note that if F is any field and $\alpha \in F$, then $F(\alpha) = F$ because F is the minimal field such that $F \subseteq F$ and $\alpha \in F$.

Using some ring theory, there is a way to give an explicit construction for $F(\alpha)$ up to isomorphism that is completely independent of any specific field extension of F (for ring theory definitions and results, see [1] chapters 7, 8, and 9).

Theorem 3.3. *Let K be a field extension of F and let $\alpha \in K$. Suppose there is a polynomial $p(x) \in F[x]$ such that $p(x)$ is irreducible in $F[x]$ and $p(\alpha) = 0$. Then,*

$$F(\alpha) \cong F[x]/(p(x)).$$

Proof. Consider the ring homomorphism $\varphi\colon F[x] \to F(\alpha)$ sending $f(x)$ to $f(\alpha)$. It follows that

$$\varphi(p(x)) = p(\alpha) = 0,$$

so $p(x) \in \ker \varphi$. Therefore, we obtain an induced homomorphism

$$\varphi\colon F[x]/(p(x)) \to F(\alpha).$$

Since $(p(x))$ is a maximal ideal, $F[x]/(p(x))$ is a field. Hence, $\ker \varphi = (0)$, so φ is injective. Therefore, φ forms an isomorphism between $F[x]/(p(x))$ and $\varphi(F[x]/(p(x)))$,

so $\varphi(F[x]/(p(x)))$ is a field. Also,

$$\varphi(F) = F \subseteq \varphi(F[x]/(p(x)))$$

and

$$\varphi(x) = \alpha \in \varphi(F[x]/(p(x))),$$

so $F(\alpha) \subseteq \varphi(F[x]/(p(x)))$. Since $F(\alpha)$ is the codomain of φ, $\varphi(F[x]/(p(x))) \subseteq F(\alpha)$. Therefore,

$$\varphi(F[x]/(p(x))) = F(\alpha),$$

so φ is surjective. Thus, φ is an isomorphism and

$$F[x]/(p(x)) \cong F(\alpha).$$

□

The above theorem essentially states that we can construct $F(\alpha)$ by taking the polynomial ring $F[x]$ and defining $p(x) = 0$. This construction is algebraically identical to $F(\alpha)$ as there is a structure-preserving correspondence between the two fields given by replacing x with α and vice versa. In fact, the reader may already be familiar with one such construction. Recall the earlier claim that $\mathbb{C} = \mathbb{R}(i)$. Since i is a root of $x^2 + 1$, which is an irreducible polynomial in $\mathbb{R}[x]$, by Theorem 3.3,

$$\mathbb{R}(i) \cong \mathbb{R}[x]/(x^2 + 1).$$

Essentially, the field $\mathbb{R}[x]/(x^2 + 1)$ is the polynomial ring $\mathbb{R}[x]$ but with $x^2 + 1$ defined to be equal to 0. Defining $x^2 + 1 = 0$ is equivalent to defining $x^2 = -1$, thus the field $\mathbb{R}[x]/(x^2 + 1)$ is the polynomial ring $\mathbb{R}[x]$ with x^2 defined to be -1. This is exactly the way in which $\mathbb{C}$ is commonly defined, only with x replaced with a formal variable i which is defined to satisfy the equation $i^2 = -1$.

Theorem 3.3 also proves why every element in $\mathbb{Q}(\sqrt{2})$ must be of the form $a+b\sqrt{2}$. Since $\sqrt{2}$ is a root of the polynomial $x^2 - 2$, which was shown to be irreducible in $\mathbb{Q}[x]$ in Example 2.4, by Theorem 3.3,

$$\mathbb{Q}(\sqrt{2}) \cong \mathbb{Q}[x]/(x^2 - 2).$$

In $\mathbb{Q}[x]/(x^2 - 2)$, since $x^2 - 2$ is defined to be 0, x^2 is defined to be 2. Hence, every term x^k for some $k \in \mathbb{N}$ is equal to $2^{k/2}$ or $2^{\frac{k-1}{2}}x$. Therefore, every element of $\mathbb{Q}[x]/(x^2 - 2)$ can be written as $a + bx$ for $a, b \in \mathbb{Q}$.

In fact, we can use Theorem 3.3 make an even stronger statement about the structure of $\mathbb{Q}(\sqrt{2})$. It turns out that each element of $\mathbb{Q}(\sqrt{2})$ can be expressed *uniquely* in the form $a + b\sqrt{2}$. This allows us to think of each element of $\mathbb{Q}(\sqrt{2})$ as a vector with two components, where the elements 1 and $\sqrt{2}$ form a basis for $\mathbb{Q}(\sqrt{2})$ when it is thought of as a vector space over $\mathbb{Q}$. This is entirely analogous to how the complex numbers can be thought of as a 2-dimensional vector space over $\mathbb{R}$ with the basis $1, i \in \mathbb{C}$, and that analogy is no coincidence. Given any field extension K/F, K forms a vector space over F because all of the vector addition and scalar multiplication axioms follow directly from the field axioms. This corresponds to the fact that $\mathbb{Q}(\sqrt{2})$ forms a vector space over $\mathbb{Q}$ and $\mathbb{C}$ forms a vector space over $\mathbb{R}$. This fact also means that both $\mathbb{R}$ and $\mathbb{C}$ form vector spaces over $\mathbb{Q}$, but these spaces are somehow different from the previous two examples. No matter how many elements we append to $\mathbb{Q}$ (provided that number is finite) we will never be able to "generate" $\mathbb{R}$, and by extension, $\mathbb{C}$. This means that $\mathbb{R}$ and $\mathbb{C}$, as a vector spaces over $\mathbb{Q}$, have no finite basis. So, $\mathbb{R}/\mathbb{Q}$ and $\mathbb{C}/\mathbb{Q}$ form infinite-dimensional

vector spaces, as opposed to the finite-dimensional vector spaces formed by $\mathbb{C}/\mathbb{R}$ and $\mathbb{Q}(\sqrt{2})/\mathbb{Q}$.

Now that we can consider field extensions as vector spaces, the dimension of those vector spaces can give us a sense of the "size" of various field extensions.

Definition 3.4. Let K/F be a field extension. Then, the *degree* of the field extension K/F is the dimension of K when interpreted as a vector space over F. If the degree of K/F is n, we write

$$[K : F] = n.$$

Since $\mathbb{C}$ is a 2-dimensional vector space over $\mathbb{R}$, the degree of $\mathbb{C}/\mathbb{R}$ is 2. Written in symbols,

$$[\mathbb{C} : \mathbb{R}] = 2.$$

As previously stated, $1, \sqrt{2} \in \mathbb{Q}(\sqrt{2})$ forms a basis for $\mathbb{Q}(\sqrt{2})/\mathbb{Q}$, so $[\mathbb{Q}(\sqrt{2}) : \mathbb{Q}] = 2$ as well. The following theorem will prove these facts.

Theorem 3.5. *Let K/F be a field extension and let $\alpha \in K$. Suppose there exists an irreducible polynomial $p(x) \in F[x]$ such that $p(\alpha) = 0$. Let n be the degree of $p(x)$. Then, $1, \alpha, \ldots, \alpha^{n-1}$ forms a basis for $F(\alpha)$ as a vector space over F, and*

$$[F(\alpha) : F] = n.$$

Proof. For the remainder of this proof, we will think of $F(\alpha)$ as a vector space over F. By Theorem 3.3,

$$F(\alpha) \cong F[x]/(p(x)).$$

Let $\varphi\colon F[x]/(p(x)) \to F(\alpha)$ be the isomorphism given in the proof of Theorem 3.3.

Let $t \in F(\alpha)$. Since φ is an isomorphism, φ is surjective, so there exists an element $f \in F[x]/(p(x))$ such that $\varphi(f) = t$. There exists a $g(x) \in F[x]$ such that $f = g(x) + p(x)F[x]$. By the Euclidean algorithm for $F[x]$, there exist $h(x), r(x) \in F[x]$ with $\deg r(x) < n$ such that

$$g(x) = h(x)p(x) + r(x).$$

Hence,

$$g(x) + p(x)F[x] = r(x) + p(x)F[x].$$

Since $\deg r(x) < n$, there exist $r_0, r_1, \ldots, r_{n-1} \in F$ such that

$$r(x) = r_0 + r_1 x + \cdots + r_{n-1}x^{n-1}.$$

Therefore,

$$t = \varphi(f) = \varphi(g(x) + p(x)F[x]) = \varphi(r(x) + p(x)F[x]) = r(\alpha),$$

which means

$$t = r_0 + r_1\alpha + \cdots + r_{n-1}\alpha^{n-1}.$$

Therefore, the elements $1, \alpha, \ldots, \alpha^{n-1}$ span $F(\alpha)$.

Suppose the elements $1, \alpha, \ldots, \alpha^{n-1}$ are not linearly independent. Then, there exist $c_0, c_1, \ldots, c_{n-1} \in F$ where there is at least one $c_i \neq 0$ such that

$$c_0 + c_1\alpha + \cdots + c_{n-1}\alpha^{n-1} = 0. \tag{3.6}$$

Let $c(x) \in F[x]$ be given by

$$c(x) = c_0 + c_1 x + \cdots + c_{n-1}x^{n-1}.$$

Since there is one $c_i \neq 0$ with $0 \leq i \leq n-1$, $c(x)$ is a nonzero polynomial. Also,

$$\deg c(x) < n.$$

By (3.6), $c(\alpha) = 0$. Hence,

$$\varphi(c(x) + p(x)F[x]) = c(\alpha) = 0 = \varphi(0 + p(x)F[x]).$$

Since φ is an isomorphism, φ is injective, so

$$c(x) + p(x)F[x] = 0 + p(x)F[x].$$

Therefore, $c(x) \in p(x)F[x]$, so $p(x)$ divides $c(x)$. Since $c(x) \neq 0$ and $p(x)$ divides $c(x)$, it follows that $\deg c(x) \geq n$. However, this is a contradiction, because $\deg c(x) < n$. Therefore, $1, \alpha, \ldots, \alpha^{n-1}$ are linearly independent.

Because $1, \alpha, \ldots, \alpha^{n-1}$ are linearly independent and span $F(\alpha)$, these elements form a basis for $F(\alpha)$. Since the basis has n elements, $F(\alpha)$ is an n-dimensional vector space. By Definition 3.4,

$$[F(\alpha) : F] = n.$$

□

The above theorem is the culmination of our study of fields so far because it allows us to instantly grasp the structure of any field extension of the form $F(\alpha)$ as soon as we find an irreducible polynomial $p(x) \in F[x]$ such that $p(\alpha) = 0$. It will prove essential in our study of compass and straightedge constructions. While this theorem is powerful on its own, it becomes many times more powerful when used in tandem with the following theorem.

Theorem 3.7. *Let L/K and K/F be field extensions. Then, L/F is a field extension. The degrees $[L : K]$ and $[K : F]$ are finite if and only if $[L : F]$ is finite. If those degrees are finite, then*

$$[L : F] = [L : K][K : F].$$

Proof. Suppose $[L : F]$ is finite. Hence, L/F has a finite basis. Since $F \subseteq K$, this means L/K is spanned by a finite number of vectors. Therefore, there exists a finite basis for L/K. Thus, $[L : K]$ is finite. Since K/F is a subspace of L/F, which is a finite dimensional vector space, K/F is also a finite dimensional vector space. Thus $[K : F]$ is finite.

Now, instead suppose $[L : K] = n$ and $[K : F] = m$. Then, L/K is an n-dimensional vector space and K/F is an m-dimensional vector space. Therefore, there exists a basis $\alpha_1, \alpha_2, \ldots, \alpha_n \in L$ for L/K and a basis $\beta_1, \beta_2, \ldots, \beta_m \in K$ for K/F. It is easy, but somewhat tedious to check that the set

$$\mathcal{B} = \{\beta_i \alpha_j \mid 1 \leq i \leq m \text{ and } 1 \leq j \leq n\}$$

forms a basis for L/F. Since $\mathcal{B}$ has nm elements, the dimension of L/F is nm. Thus,

$$[L : F] = nm = [L : K][K : F].$$

□

There are many nontrivial results that this theorem, when combined with Theorem 3.5, can prove with surprising ease. Some of these results are the compass and straightedge proofs that we are searching for, as these theorems allow us to place a tight bound on what degree field extensions are possible in the subfields of the constructible numbers.

4. Compass and Straightedge Constructions Revisited

The treatment of compass and straightedge constructions here follows Chapter 30 of [2].

Now that we have some powerful tools from field theory, we are ready to use them to analyze compass and straightedge constructions.

Proposition 4.1. *Let F be a subfield of $\mathbb{R}$ and let $(a,b) \in \mathbb{R}^2$. Let K be the minimum subfield of $\mathbb{R}$ such that $F \subseteq K$ and $(a,b) \in K \times K$. Then,*

$$K = F(a,b).$$

Proof. We have $(a,b) \in K \times K$ if and only if $a, b \in K$, so the proposition follows directly from Definition 3.2. □

The above shows how we might use field extensions to analyze how a new point is marked. The following theorems will use the field theoretic tools we have built up to carry out that analysis.

Theorem 4.2. *Let F be a subfield of $\mathbb{R}$ and let $(a,b) \in \mathbb{R}^2$ be a point constructible in one step from $F \times F$. Then,*

$$[F(a,b) : F] = 2^m$$

for some $m \in \mathbb{N} \cup \{0\}$.

Proof. By Definition 1.1, (a,b) is the intersection of two distinct figures where each figure is either:

(1) a line $\overline{AB}$ where $A, B \in F \times F$.
(2) a circle with radius AB centered on C where $A, B, C \in F \times F$.

Hence, (a,b) is either the intersection of a line and a line, the intersection of a line and a circle, or the intersection of a circle and a circle. For the remainder of this proof, if there is a point $P \in \mathbb{R}^2$, we will refer to the coordinates of P as x_P and y_P so that $P = (x_P, y_P)$. Note that if $P \in F \times F$, we have $x_P, y_P \in F$.

Suppose (a,b) is the intersection of two distinct lines $\overline{AB}$ and $\overline{CD}$ with $A, B, C, D \in F \times F$. Therefore,

$$(x_B - x_A)(b - y_A) = (y_B - y_A)(a - x_A)$$

and

$$(x_D - x_C)(b - y_C) = (y_D - y_C)(a - x_C).$$

Solving these equations simultaneously gives expressions for a and b in terms of elements of F combined using the field operations, which means $a, b \in F$. Hence, $F(a,b) = F$, so

$$[F(a,b) : F] = [F : F] = 1 = 2^0.$$

Suppose (a,b) is the intersection of a line $\overline{AB}$ and a circle with radius CD centered on E for $A, B, C, D, E \in F \times F$. Let $r = CD$. Then,

$$r = \sqrt{(x_D - x_C)^2 + (y_D - y_C)^2}.$$

Hence,

$$r^2 = (x_D - x_C)^2 + (y_D - y_C)^2 \in F.$$

Then,

$$(x_B - x_A)(b - y_A) = (y_B - y_A)(a - x_A)$$

and

$$(a - x_E)^2 + (b - y_E)^2 = r^2.$$

Solving the linear equation either gives the value of a or b in terms of elements of F, or it gives a linear relationship between a and b. In the first case, applying the result to the circle equation gives a polynomial equation for the other value of a or b. In the second case, applying the result twice gives a polynomial equation for a and a different polynomial equation for b. In all cases, there are polynomials in $F[x]$ (and by extension, $F(a)[x]$) of degree less than or equal to two which a and b are roots of. In each polynomial ring, each of these polynomials must either be irreducible or have irreducible factors that a and b are roots of, which must also be of degree less than or equal to two. Therefore, by Theorem 3.5,

$$[F(a) : F] \leq 2 \quad \text{and} \quad [F(a, b) : F(a)] \leq 2.$$

Thus, $[F(a) : F], [F(a, b) : F(a)] \in \{1, 2\}$, so $[F(a) : F], [F(a, b) : F(a)]$ are both powers of two. Hence,

$$[F(a, b) : F] = [F(a, b) : F(a)][F(a) : F] = 2^m$$

for some $m \in \mathbb{N} \cup \{0\}$.

Suppose (a, b) is the intersection of a circle with radius AB centered on C and a circle of radius DE centered on G for $A, B, C, D, E, G \in F \times F$. Let $r_C = AB$ and $r_G = DE$. By a similar argument to the previous case, $r_C^2, r_G^2 \in F$,

$$(a - x_C)^2 + (b - y_C)^2 = r_C^2,$$

and

$$(a - x_G)^2 + (b - y_G)^2 = r_G^2.$$

It follows that

$$\begin{aligned} (a - x_C)^2 + (b - y_C)^2 &= r_C^2 \\ \implies (a^2 - 2x_C a + x_C^2) + (b^2 - 2y_C b + y_C^2) &= r_C^2 \\ \implies a^2 + b^2 - 2x_C a - 2y_C b + (x_C^2 + y_C^2 - r_C^2) &= 0 \\ \implies a^2 + b^2 + c_1 a + c_2 b + c_3 &= 0 \end{aligned}$$

for $c_1, c_2, c_3 \in F$. By a similar process, one can use the other equation to find $d_1, d_2, d_3 \in F$ that depend on D, E, and G such that

$$a^2 + b^2 + d_1 a + d_2 b + d_3 = 0.$$

Taking the difference between these equations, we see that (a, b) lies on the line

$$(c_1 - d_1)x + (c_2 - d_2)y + (c_3 - d_3) = 0.$$

Because the two circles are distinct, $c_i - d_i \neq 0$ for at least one $i \in \{1, 2, 3\}$, so the graph of the above equation is not the entire plane. Also, (a, b) satisfies the above equation, so the graph of the equation is not empty. Hence, the graph of the equation is a non-degenerate line. Furthermore, all the coefficients are in F, so this line must pass through two points whose coordinates are in F. Therefore, (a, b) must lie on a line that passes through two points in $F \times F$. Thus, by applying the previous case,

$$[F(a, b) : F] = 2^m$$

for some $m \in \mathbb{N} \cup \{0\}$. In all cases, $[F(a, b) : F]$ is a power of two. □

Corollary 4.3. *Let F be a subfield of $\mathbb{R}$ and let $(a,b) \in \mathbb{R}^2$ such that (a,b) is constructible from $F \times F$. Then,*

$$[F(a,b) : F] = 2^m$$

for some $m \in \mathbb{N} \cup \{0\}$.

Proof. By Definition 1.3, there exists a finite sequence of points $P_1, P_2, \ldots, P_n \in \mathbb{R}^2$ such that $P_n = (a,b)$ and for each $1 \leq i \leq n$, P_i is constructible in one step from $(F \times F) \cup \{P_1, \ldots, P_{i-1}\}$. Let $K_0 = F$, and for each $1 \leq i \leq n$, let $P_i = (a_i, b_i)$ and $K_i = K_{i-1}(a_i, b_i)$. Note that since $P_n = (a,b)$, we have $(a_n, b_n) = (a,b)$ and $K_n = K_{n-1}(a,b)$.

By definition, we have $K_0 = F$, so $F \times F \subseteq K_0 \times K_0$. Suppose that for some $0 \leq j \leq n-1$,

$$(F \times F) \cup \{P_1, \ldots, P_j\} \subseteq K_j \times K_j.$$

Then, by Proposition 4.1,

$$\begin{aligned} (F \times F) \cup \{P_1, \ldots, P_{j+1}\} &\subseteq (K_j \times K_j) \cup \{P_{j+1}\} \\ &\subseteq K_j(a_{j+1}, b_{j+1}) \times K_j(a_{j+1}, b_{j+1}) \\ &= K_{j+1} \times K_{j+1}. \end{aligned}$$

Therefore, by induction, for all $0 \leq i \leq n$,

$$(F \times F) \cup \{P_1, \ldots, P_i\} \subseteq K_i \times K_i.$$

Let $1 \leq i \leq n$. Then, P_i is constructible in one step from $(F \times F) \cup \{P_1, \ldots, P_{i-1}\}$. Therefore, P_i is constructible in one step from $K_{i-1} \times K_{i-1}$. By Theorem 4.2,

$$[K_i : K_{i-1}] = [K_{i-1}(a_i, b_i) : K_{i-1}] = 2^{m_i}$$

for some $m_i \in \mathbb{N} \cup \{0\}$. Finally, by Theorem 3.7,

$$\begin{aligned} [K_n : F] &= [K_n : K_0] \\ &= [K_n : K_{n-1}][K_{n-1} : K_{n-2}] \ldots [K_1 : K_0] \\ &= (2^{m_n})(2^{m_{n-1}}) \ldots (2^{m_1}) \\ &= 2^{m_n + m_{n-1} + \cdots + m_1} \\ &= 2^m \end{aligned}$$

for some $m \in \mathbb{N} \cup \{0\}$.

Recall that $K_n = K_{n-1}(a,b)$, so $a, b \in K_n$. Also, since

$$F = K_0 \subseteq K_1 \subseteq \cdots \subseteq K_n,$$

we have $F \subseteq K_n$. Therefore, $F(a,b) \subseteq K_n$. By Theorem 3.7,

$$2^m = [K_n : F] = [K_n : F(a,b)][F(a,b) : F],$$

so $[F(a,b) : F]$ divides 2^m. Therefore, there exists some $0 \leq k \leq m$ such that

$$[F(a,b) : F] = 2^k.$$

□

This is the powerful theorem that makes all the impossibility proofs in this paper possible. For ease of use, we will put it into a form that can be more readily applied to constructible numbers.

Proposition 4.4. *A point $P \in \mathbb{R}^2$ is constructible if and only if P is constructible from $\mathbb{Q} \times \mathbb{Q}$.*

Proof. Suppose $P \in \mathbb{R}^2$ is constructible. By Definition 1.4, P is constructible from $\{(0,0),(1,0)\}$. Since

$$\{(0,0),(1,0)\} \subseteq \mathbb{Q} \times \mathbb{Q},$$

P is constructible from $\mathbb{Q} \times \mathbb{Q}$.

Suppose $P \in \mathbb{R}^2$ is constructible from $\mathbb{Q} \times \mathbb{Q}$. Since the constructible numbers form a subfield of $\mathbb{R}$, and $\mathbb{Q}$ is the minimal subfield of $\mathbb{R}$, $\mathbb{Q}$ is a subset of the constructible numbers. Therefore, every point used in constructing P from $\mathbb{Q} \times \mathbb{Q}$ is constructible, which implies that P is constructible. □

Corollary 4.5. *Let $\alpha \in \mathbb{R}$ be a constructible number. Then,*

$$[\mathbb{Q}(\alpha) : \mathbb{Q}] = 2^m$$

for some $m \in \mathbb{N} \cup \{0\}$.

Proof. By Definition 1.6, $(\alpha, 0)$ is a constructible point. By Proposition 4.4, $(\alpha, 0)$ is constructible from $\mathbb{Q} \times \mathbb{Q}$. By Corollary 4.3,

$$[\mathbb{Q}(\alpha, 0) : \mathbb{Q}] = 2^m$$

for some $m \in \mathbb{N} \cup \{0\}$. Since $0 \in \mathbb{Q}$, $\mathbb{Q}(\alpha, 0) = \mathbb{Q}(\alpha)$. Therefore,

$$[\mathbb{Q}(\alpha) : \mathbb{Q}] = 2^m.$$

□

5. Impossibility Proofs

The treatment of the compass and straightedge impossibility proofs here follows Chapter 30 of [2].

We now have everything we need to do the compass and straightedge impossibility proofs. The first of these proofs is very quick and will communicate the general strategy used in this type of proof. Do not be fooled, though: this problem went unsolved for millennia. The simplicity of its proof is a testament to the beauty and power of field theory.

Proposition 5.1. *It is not possible, using only a compass and straightedge, to construct a cube with double the volume of any given cube.*

Proof. Suppose that, given the side length of any cube, it is possible to construct the side length of a cube with double the volume. Then, given a length of 1, it is possible to construct the side length of a cube with volume 2. This side length would have a length of $\sqrt[3]{2}$, so one could use this line segment to construct $(\sqrt[3]{2}, 0)$. Therefore, $\sqrt[3]{2}$ would be a constructible number. We know that $\sqrt[3]{2}$ is a root of $x^3 - 2$. Since the leading coefficient is 1, every other coefficient is divisible by 2, and the constant term is not divisible by 4, by Theorem 2.3, $x^3 - 2$ is irreducible in $\mathbb{Q}[x]$. By Theorem 3.5,

$$[\mathbb{Q}(\sqrt[3]{2}) : \mathbb{Q}] = 3,$$

which is not a power of 2. This contradicts Corollary 4.5. Therefore, it is not possible, using only a compass and straightedge, to construct a cube with double the volume of any given cube. □

The following proof is for a considerably more famous result: the problem of angle trisection, which means using a compass and straightedge to divide a given angle into three equal angles. The proof is somewhat more involved than the previous proof, but it still follows the same structure, and it is far simpler than one might expect given the infamy of this problem throughout history.

Proposition 5.2. *It is not possible, using only a compass and straightedge, to trisect any given angle.*

Proof. Given the unit segment, one can construct an equilateral triangle by drawing two circles of radius 1 centered on the points $(0,0)$ and $(1,0)$, and then marking the top intersection. Connecting these three points with lines gives an equilateral triangle of side length 1. Let $A = (0,0)$, $B = (1,0)$, and let C be the third point of the triangle. Since $\triangle ABC$ is equilateral, $\angle BAC = \frac{\pi}{3}$.

Suppose it is possible to trisect any arbitrary angle. Then, it is possible to divide $\angle BAC$ into three angles where each angle is $\frac{\pi}{9}$ radians. This will create two lines that will each intersect $\overline{BC}$. Let the lower of these points of intersection be D. Thus, $\angle BAD = \frac{\pi}{9}$. Finally, mark the point E where $\overline{AD}$ intersects a circle of radius 1 centered on A. Since E is on the unit circle and is $\frac{\pi}{9}$ radians from the x-axis,

$$E = \left(\cos\frac{\pi}{9}, \sin\frac{\pi}{9}\right)$$

(see Figure 4). Therefore, $\cos\frac{\pi}{9}$ is a constructible number. Since the constructible numbers form a field, $2\cos\frac{\pi}{9}$ is also a constructible number.

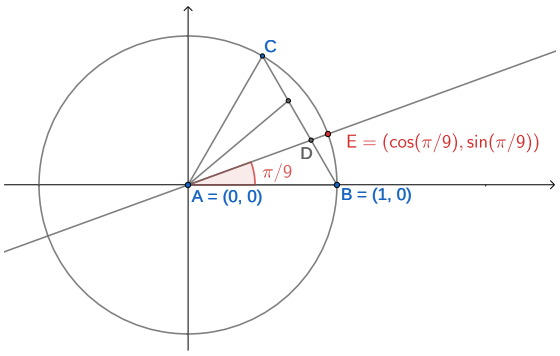

FIGURE 4. Construction of point E. Image created with [4].

By elementary trigonometry,

$$\cos(3\theta) = 4\cos^3\theta - 3\cos\theta$$

for any angle θ. It follows that

$$\begin{aligned}
\cos\frac{\pi}{3} &= 4\cos^3\frac{\pi}{9} - 3\cos\frac{\pi}{9} \\
\implies \frac{1}{2} &= 4\cos^3\frac{\pi}{9} - 3\cos\frac{\pi}{9} \\
\implies 1 &= 8\cos^3\frac{\pi}{9} - 6\cos\frac{\pi}{9} \\
\implies 0 &= 8\cos^3\frac{\pi}{9} - 6\cos\frac{\pi}{9} - 1 \\
\implies 0 &= \left(2\cos\frac{\pi}{9}\right)^3 - 3\left(2\cos\frac{\pi}{9}\right) - 1.
\end{aligned}$$

Therefore, $2\cos\frac{\pi}{9}$ is a root of $x^3 - 3x - 1$. If $x^3 - 3x - 1$ is reducible in $\mathbb{Q}[x]$, then $(x+1)^3 - 3(x+1) - 1$ is reducible in $\mathbb{Q}[x]$. Expanding the latter polynomial gives

$$\begin{aligned}
(x+1)^3 - 3(x+1) - 1 &= \left(x^3 + 3x^2 + 3x + 1\right) + (-3x - 3) - 1 \\
&= x^3 + 3x^2 + (3x - 3x) + (1 - 3 - 1) \\
&= x^3 + 3x^2 - 3.
\end{aligned}$$

In the above polynomial, the leading coefficient is 1, every other coefficient is divisible by 3, and the constant term is not divisible by 9. Therefore, by Theorem 2.3, $x^3 + 3x^2 - 3$ is irreducible in $\mathbb{Q}[x]$. Hence, $(x+1)^3 - 3(x+1) - 1$ is irreducible in $\mathbb{Q}[x]$, so $x^3 - 3x - 1$ is irreducible in $\mathbb{Q}[x]$. By Theorem 3.5,

$$\left[\mathbb{Q}\left(2\cos\frac{\pi}{9}\right) : \mathbb{Q}\right] = 3,$$

which contradicts Corollary 4.5. Therefore, it is not possible to trisect any arbitrary angle. □

Note that the above proof also shows that it is impossible to construct an angle of $\frac{\pi}{9}$ radians. This also inadvertently proves that any construction resulting in an angle of $\frac{\pi}{9}$ radians is impossible. For instance, see the following corollary.

Corollary 5.3. *It is impossible to construct a regular* 9*-gon using only a compass and straightedge.*

Proof. Suppose that it is possible to construct a regular 9-gon. Then, each exterior angle on the 9-gon is $\frac{2\pi}{9}$ radians. One could then bisect one of these angles to create an angle of $\frac{\pi}{9}$ radians. However, the proof of the above proposition shows that it is impossible to construct a $\frac{\pi}{9}$ radian angle. Therefore, it is impossible to construct a regular 9-gon with only a compass and straightedge. □

Acknowledgments

I am happy to thank my mentor, Sam Quinn, for recommending this topic, guiding my learning, helping to motivate unfamiliar concepts, assisting me in the writing process, and encouraging me along the way. This paper would not have been possible without his guidance. I would like to thank Geoffrey Baring and John Naughton for our conversations on algebra, which helped answer many of my questions on the material. I would also like to thank Connor Lockhart, Jacob Fiedler, and Professor Daniil Rudenko for giving me an enriching and fun experience in the apprentice program. Finally, I would like to thank Professor Peter May for organizing the REU. Because of your hard work leading the program, I learned a lot of beautiful mathematics and made some wonderful memories. Thank you.

References

[1] David S. Dummit, Richard M. Foote. Abstract Algebra. John Wiley & Sons. 2004.
[2] Charles C. Pinter. A Book of Abstract Algebra. McGraw-Hill. 1990.
[3] Thomas W. Hungerford. Abstract Algebra: An Introduction. Brooks/Cole, Cengage Learning. 2014.
[4] GeoGebra Geometry. https://www.geogebra.org/.