Compass-and-Straightedge Constructions

We learn exactly what compass-and-straightedge constructions are, and what they can do.

So, what exactly is a compass-and-straightedge construction? What can we do—and not do? Let’s first be very careful about determining exactly what actions we can accomplish, and what it means to “construct” something.

First off, it is obvious that any thing we construct must consist of circles, arcs, and line segments. We will not restrict ourselves in any way as to size. That is, we’ll pretend that we have an arbitrarily large piece of paper to work with and can draw precisely at any scale necessary. So if we want to construct a regular 10,000-sided figure, we will allow ourselves to do so, either by pretending that we have a very big piece of paper or by pretending that we can make very small, very fine, and completely accurate marks at any smaller scale. Provided we can prove that at the end our figure has 10,000 sides all the same size and angle.

Here are some things that we can obviously do:

1. Mark an arbitrary, random point on the plane.
2. Mark an arbitrary, random point on some line segment that has already been drawn.
3. Mark an arbitrary, random point on some circle or arc that has already been drawn.
4. Given two points, draw the line that passes through both.
5. Given two points, draw the circle centered at one, passing through the other.
6. Given two lines, a line and a circle, or two circles, we may mark all intersections.

This is actually a complete list. For convenience, we will also allow ourselves any constructions that we have already proven to be possible, for example:

7. Given a point and a distance between two other points, draw the circle centered at the given point with radius equal to the given distance.
8. Given a line and a point, draw the perpendicular to the line through the point.
9. Given a line and a point, draw the parallel to the line through the point.
11. Find the midpoint of a segment.

We are only going to work in Euclidean geometry for now, where there is exactly one line parallel to a given line through a given point.

It should be noted that the points marked in 1–3 are really arbitrary and may not have any special properties. For instance, we can’t issue an instruction such as “now mark the point that is 2/3 of the way from $A$ to $B$.” That has to be specifically found using $A$ and $B$ and other known reasoning.
So how does construction work? We start with some points and use the rules to construct other points, then prove the points we have found have desirable relationships with each other and the original points. We’ll consider a figure to be “constructed” if we can mark enough key points on it so the when we connect them with line segments and arcs we’ll have the figure we desire.

**Example**  Given two points, can we construct a regular pentagon with the segment between them as one of the sides?

The answer is “yes” and here’s how it’s done. First, through similar triangles, we can determine that if the given points are (say) distance 1 apart, then the diagonals of the desired pentagon have length \( \sqrt{\frac{5+1}{2}} \). Now it is not hard to construct a right triangle whose legs are 1 and 2, so its hypotenuse would be \( \sqrt{5} \). We could extend this hypotenuse and mark off a distance of 1 from one of the ends so that we have found a segment of length \( \sqrt{5} + 1 \). Then we can find the midpoint of this segment, and we’ll have a segment of the right length. From our original segment, draw a circle of radius 1 from one endpoint, and of \( \sqrt{\frac{5+1}{2}} \) from the other, and we’ll have found a third vertex of the pentagon. Continue around until the pentagon is drawn.

Apparently, it is useful to know what lengths can be constructed. In fact, let’s make the following definition:

**Definition**  Given two points in the plane. A positive real number \( r \) is **constructible** if, from the two given points, two (not necessarily different) points may be constructed where the distance between them is (provably!) \( r \) times the distance between the original two points.

Now let’s say that we’re told that the distance between two given points is one unit. We can easily create a coordinate system, with the origin anywhere we want, and with the axes oriented any way we like, using the given unit distance as one unit along the \( x \)-axis and along the \( y \)-axis. If you have two constructible numbers, \( a \) and \( b \), you can certainly mark point \((a,0)\) on the \( x \)-axis, and \((0,b)\) on the \( y \)-axis. Then by drawing perpendiculars to the axes through these points, we can locate \((a,b)\). Thus, any pair of constructible numbers gives you a constructible point. Conversely, given a constructible configuration of points, all the distances between them must be constructible numbers, since we can transfer those distances to the origin and mark them off along either axis. Thus, there is a direct correspondence between constructible points and constructible numbers. So this clarifies the above “apparently” statement! Knowing what lengths can be constructed tells you exactly what points can be constructed.

Before we go further, we’d better prove the Pythagorean Theorem.
Theorem  Given a right triangle with legs of length $a$ and $b$ and hypotenuse of length $c$, $a^2 + b^2 = c^2$. To prove this, in the diagram below the large square has sides of length $c$ so its area is $c^2$. The four triangles each have legs $a$ and $b$, so have area $ab/2$. The four together have area $2ab$. The small square has area $(a - b)^2 = a^2 - 2ab + b^2$ so by adding areas $c^2 = a^2 + b^2$.

So now we have all the tools to determine what numbers, and hence which figures, are constructible.

Theorem  All positive integers are constructible. To do so, draw a line and mark a point on it. Transfer the unit distance to this line starting at the point. Go to the end of the unit segment, and mark another one, going in the same direction. Continue as needed (technically, do an induction argument).

Theorem  All positive rational numbers are constructible. To do so, let the rational numbers be $m/n$ in lowest terms. Mark off $(1, 0), (2, 0), \ldots (m, 0)$ along the $x$-axis and $(0, 1), (0, 2), \ldots (0, n)$ along the $y$-axis. Draw the line from $(m, 0)$ to $(0, n)$. Now draw the line parallel to this line that passes through the point $(0, 1)$. By similar triangles, it intersects the $x$-axis at $(m/n, 0)$.

So we can get all rational numbers. There are clearly others we can construct, such as $\sqrt{2}$ which is the distance between $(0, 1)$ and $(1, 0)$. To allow us to see the patterns, let’s be more general, and prove the ways in which we can combine constructible numbers.

Theorem  If $a$ and $b$ are constructible, so are $a + b$, $a - b$, $ab$, and $a/b$ (assuming $b \neq 0$).

The proof of this is pretty easy. Since $a$ is constructible, we can mark off the point $(a, 0)$ on the $x$-axis. Then set the compass to length $b$, place the center at $(a, 0)$ and draw a circle. It intersects the $x$-axis at $(a + b, 0)$ and $(a - b, 0)$. To multiply, plot the points $(1, 0), (a, 0)$, and $(0, b)$. Now draw the line from $(1, 0)$ to $(0, b)$. Draw the parallel line through $(a, 0)$ which intersects the $y$-axis at $(0, ab)$ by similar triangles. We can divide by reversing this process (just like we created rational numbers above).

So we can add, subtract, multiply, and divide. This creates a mathematical object called a field. Fields offer the nicest kind of arithmetic. We started with the integers and did basic
arithmetic and ended with the rational numbers. But as we know we can get more. It turns out that there is only one more thing we can do.

**Theorem** If $a$ is constructible, so is $\sqrt{a}$.

The construction here is a little more complicated. Starting at the origin $O$, mark the points $A(1,0)$ and $B(1+a,0)$. Find the midpoint $(\frac{1+a}{2},0)$ and draw a circle with radius $\frac{1+a}{2}$ so that $(0,0)$ and $(1+a,0)$ are on opposites ends of a diameter. Now draw a line perpendicular to the $x$-axis through $(1,0)$. It intersects the circle at $C$. From similar triangles $\triangle OAC$ and $\triangle CAB$, we learn that $\frac{1}{AC} = \frac{AC}{a}$ so $AC = \sqrt{a}$.

So we can start with the rationals, and build a “square root tower” of extensions. We can continue adding, subtracting, multiplying, dividing, and taking square roots as much as we like, and the result will be constructible. Thus, for instance, the pentagon is constructible because all we needed to do was find its diagonals, and their length $\frac{1+\sqrt{5}}{2}$ is constructible.

It’s also not too hard to prove that this is all we can do. In other words, we have proven that if we take constructible numbers and add, subtract, multiply, divide, and square root them, we get more constructible numbers. We’ll prove the converse, that a constructible number can be arrived by a finite number of steps of arithmetic. That is, with lines and circles, there’s nothing else we can get.

**Theorem** Let $a$ be a constructible number. Then it can be determined by a finite sequence of additions, subtractions, multiplications, divisions, and square root extractions.

A number is constructible if it is the distance between two points after a finite sequence of operations with lines and circles, starting with just two points a known unit distance apart. These lines and circles are themselves determined by earlier-found constructible points. Thus, a line is of the form $y = mx + b$ or $x = c$ (c a constant) and $m$, $b$, and $c$ are constructible numbers. A circle must be of the form $(x-h)^2 + (y-k)^2 = r^2$ with $a$, $b$, and $r$ constructible. Where do these things intersect to create more constructible points?

- Two lines of the first type, $y = m_1 x + b_1$ and $y = m_2 x + b_2$ intersect only if $m_1$ and $m_2$ are different (otherwise they are parallel or the same line). In this case, solving gives $x = (b_1 - b_2)/(m_2 - m_1)$. So the new $x$ coordinate is found by subtractions and divisions. The $y$ coordinate is found from $x$ by adding and multiplying.

- A slanted line and a vertical line intersect at $(c, mc + b)$, so the new coordinates are basic arithmetic from the old ones.

- Two vertical lines don’t intersect.

- A line and a circle intersect at points that can be found by substitution: $(x - h)^2 + (mx + b - k)^2 = r^2$. If this is multiplied out and like terms collected, it turns into a quadratic equation, whose roots can be found from the quadratic formula—which only involves only the $+ - \times \div \sqrt{\cdot}$ operations.

- To find the intersection between two circles, we subtract their equations, leaving a linear equation. Substituting this back into the circle leaves the same case as intersecting a line and a circle.
So, to summarize. A number is constructible if it is the distance between two constructible points. A point is constructible if and only if both of its coordinates are constructible numbers. You can do ordinary arithmetic and square roots with constructible numbers and you get other constructible numbers, and that’s the only way to get constructible numbers. Pretty nice little self-contained system! So next, we’ll explore numbers that are not constructible.