CHAPTER 1

The Fundamental Theorem of Algebra

Our first excursion into the topology of the plane will be in the proof of the Fundamental Theorem of Algebra:

THEOREM 1.1 (Fundamental Theorem of Algebra). *If $f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0 = 0$ is a polynomial equation in the unknown x and if the constant coefficients a_{n-1}, ..., a_1, a_0 are complex numbers, then there is a complex number $x = \alpha$ that satisfies the equation: $f(\alpha) = 0$. (Of course, since the real numbers are a subset of the complex numbers (the line lies in the plane), these coefficients are allowed to be real numbers.)*

The Greeks solved linear and quadratic equations geometrically. They rejected solutions that were not real numbers as being of no application or interest. Complex numbers were first acknowledged as important in the solution of cubic equations. General solution formulas for the cubic and quartic equations were found only by great effort and cleverness.

Georg Pólya wrote to me when I was a young Mormon missionary in Austria. He said that I should solve a hard mathematical problem every week so that I wouldn't rust ("Wer rastet, der rostet"). He also gave me a list of books that I might find in a used bookstore. I learned the following argument from one of them:

Here is a general solution to the cubic equation, $f(x) = x^3 + ax^2 + bx + c = 0$: We simplify the equation by translation, setting $x = z + k$. By subsititution, we find

$$f(x) = z^3 + (3k + a)z^2 + (3k^2 + 2ak + b)z + (k^3 + ak^2 + bk + c) = 0.$$

If we choose $k = -a/3$, the equation has the form

$$z^3 + pz + q = 0.$$

Setting $z = u + v$ and substituting, we find

$$(u^3 + v^3 + q) + (3uv + p)(u + v) = 0.$$

If we are able to choose u and v so that

$$u^3 + v^3 + q = 0 \quad \text{and} \quad 3uv + p = 0,$$

the equation will be satisfied. We can solve this pair of equations for u and v in the standard way by substitution:

$$u = -p/3v \text{ and, consequently,}$$

$$\left(\frac{-p}{3v}\right)^3 + v^3 + q = 0.$$

Multiplying by v^3,

$$v^6 + qv^3 + \left(\frac{-p}{3}\right)^3 = 0.$$

Though this equation has degree 6 in the unknown v, it is only quadratic in the unknown v^3, so that by the quadratic formula,

$$v^3 = -\frac{q}{2} \pm \frac{1}{2}\sqrt{q^2 + 4\frac{p^3}{27}} = -\frac{q}{2} \pm \sqrt{\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3}.$$

Assuming that we know how to take cube roots, we obtain v, from which we find $u^3 = -v^3 - q$ or $u = -p/3v$, $z = u + v$, and $x = z - (a/3)$.

For example, if $f(x) = x^3 - 15x - 4$ (already simplified), we have $p = -15$, $q = -4$, $v^3 = 2 + \sqrt{4 - 125} = 2 + 11i$, and $u^3 = 2 - 11i$. But $(2 + i)^3 = 2 + 11i$ and $(2 - i)^3 = 2 - 11i$. The sum of these cube roots is $(2 + i) + (2 - i) = 4$, which is a real root of the original equation. The fact that real roots could be mediated by formulas involving complex numbers was a huge motivating factor for the acceptance of complex numbers.

We have ignored some of the obvious difficulties: How do we take cube roots? Further, if we take the three cube roots of v^3 and three of u^3, we would be able to put together 9 possibilities for z, and there can only be three solutions to the original equation. To choose those that are in fact roots, we must satisfy the more restrictive equation, $u = -p/3v$. In the example,

$$\frac{-p}{3(2+i)} = \frac{15}{6+3i} \cdot \frac{6-3i}{6-3i} = \frac{15 \cdot (6-3i)}{36+9} = 2 - i.$$

Mathematicians hoped to find similar solutions for polynomial equations of higher degree, solutions that only required the arithmetic operations of addition, subtraction, multiplication, and division together with the extraction of roots. This hope was dashed by the work of N. H. Abel and E. Galois, who showed that the four arithmetic operations and extraction of roots were inadequate in general for expressing the roots of a quintic equation in terms of its coefficients.

I would love to include their proofs here, but Abel's proof required twenty pages of work and Galois's development requires a substantial development of the theory of fields and finite groups. Instead, we shall only prove the Fundamental Theorem of Algebra. We first need to review some fundamentals from the arithmetic of complex numbers.

1.1. Complex Arithmetic

Complex numbers were fully accepted when mathematicians learned to interpret them as the points of the plane, with $a + bi$ represented by the pair (a, b). **Addition** is then simply vector addition or parallelogram addition, with $(a+bi)+(c+di) = (a+c)+(b+d)i$ corresponding to $(a, b)+(c, d) = (a+c, b+d)$. See Figure 1.

Multiplication also has a beautiful geometric interpretation that is based on the polar representation of a complex number (see Figure 2) and on the sum formulas for the sin and cos. We review these trigonometric sum formulas and their proofs here.

The sum formulas for sin and cos are consequences of a simple projection principle, that is essentially the definition of sin and cos. See Figure 3. It is helpful

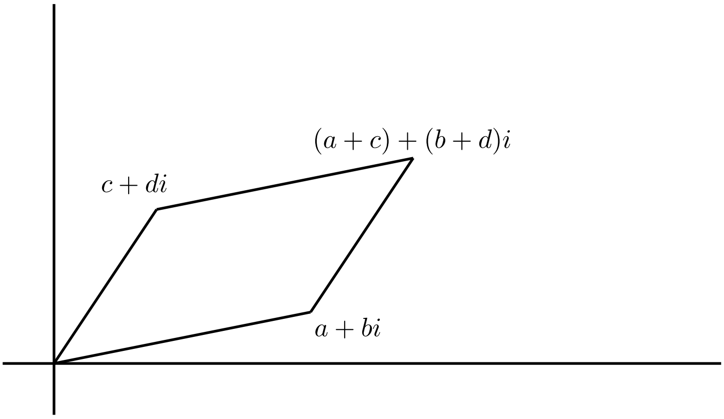


FIGURE 1. Addition of complex numbers

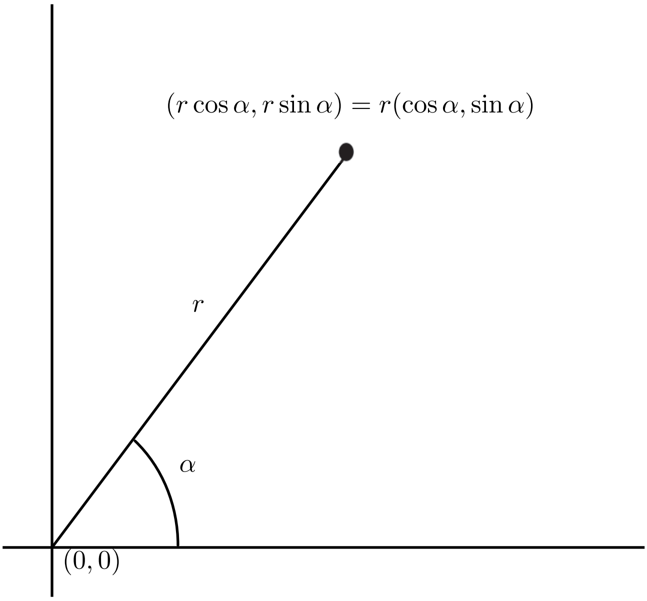


FIGURE 2. Polar coordinates

to think of the projection principle as determining the effect on lengths when one length is projected orthogonally onto another.

The projection principle. Consider a right triangle with one of its acute angles equal to α, and suppose that the length of the hypotenuse is r. Then the leg adjacent to the angle α has length $(\cos\alpha)\cdot r$, and the leg opposite the angle α has length $(\sin\alpha)\cdot r$.

To obtain the sum formulas for sin and cos, we apply the projection principle to the following two diagrams; see Figure 4. Each term of the sum formulas represents a well-defined geometric segment in the diagrams.

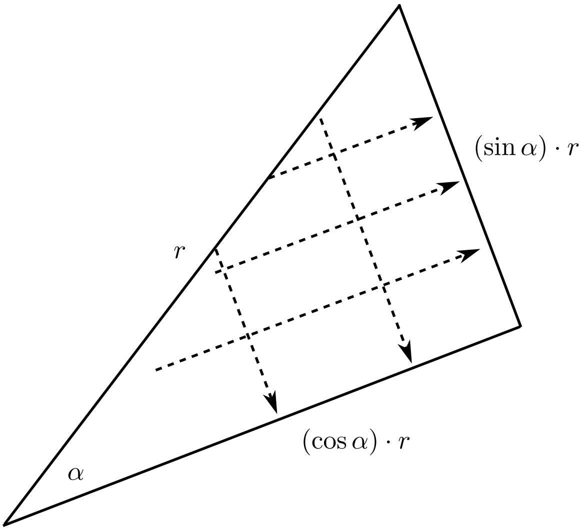


FIGURE 3. The projection principle

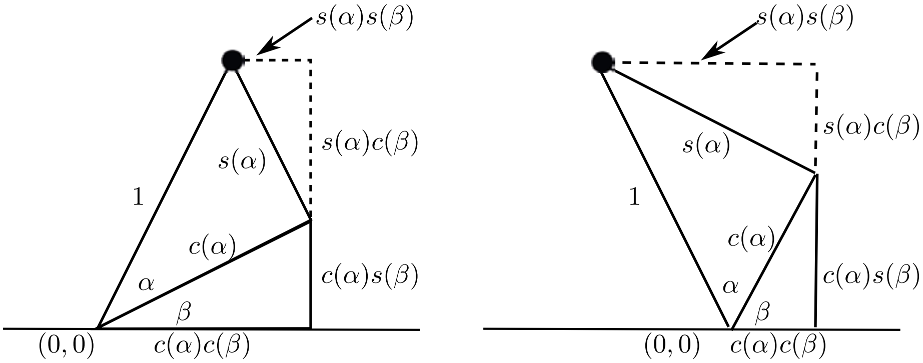


FIGURE 4. The addition formulas

We use c and s as shorthand for $\cos(\alpha+\beta)$ and $\sin(\alpha+\beta)$. We use $c(\alpha)$ and $s(\alpha)$ as shorthand for $\cos(\alpha)$ and $\sin(\alpha)$, and similarly for $\cos(\beta)$ and $\sin(\beta)$. We have drawn two right triangles in each of the diagrams, one with angle α, the other with angle β. We have scaled the triangles so that the larger one has hypotenuse 1 so that the vertex emphasized by the large dot has coordinates (c,s). All of the other entries are consequences of the projection principle. From the figures, it is clear that

$$c = c(\alpha)c(\beta) - s(\alpha)s(\alpha), \text{ or}$$

$$\cos(\alpha+\beta) = \cos(\alpha)\cos(\beta) - \sin(\alpha)\sin(\beta), \text{ and similarly}$$

$$\sin(\alpha+\beta) = \cos(\alpha)\sin(\beta) + \sin(\alpha)\cos(\beta).$$

These are the **sum formulas** for sin and cos.

Admittedly, the diagrams deal only with positive angles α and β whose sum is $\leq \pi$, but all other angles can readily be reduced to these cases.

If we add to these two formulas the **Pythagorean Theorem**,

$$\cos^2(\alpha) + \sin^2(\alpha) = 1,$$

we have the three basic identities of trigonometry.

If we apply these identities to the multiplication of complex numbers, as expressed in polar coordinates, we obtain the geometric interpretation of complex multiplication.

We multiply two complex numbers by applying the standard principles of multiplication and addition, then set $i^2 = -1$:

$$\begin{aligned} r(\cos\alpha + & i\sin\alpha) \cdot s(\cos\beta + i\sin\beta) \\ &= rs\big((\cos\alpha \cdot \cos\beta - \sin\alpha \cdot \sin\beta) + i(\cos\alpha \cdot \sin\beta + \sin\alpha \cdot \cos\beta)\big) \\ &= rs\big(\cos(\alpha+\beta) + i \cdot \sin(\alpha+\beta)\big). \end{aligned}$$

The second equality is a consequence of the addition formulas for sin and cos. In other words, **to multiply, we multiply lengths and add angles.**

To raise to the nth power,

$$\big(r(\cos\alpha + i \cdot \sin\alpha)\big)^n = r^n\big(\cos(n\alpha) + i\sin(n\alpha)\big),$$

we raise length to the nth power and multiply the angle by n.

Division is just as easy: divide lengths and subtract angles.

Taking the nth root is more sophisticated. Of course, we may take the nth root of length and divide the angle by n. But the angle of a complex number is not well-defined, for we can add any multiple of 2π to a given angle without changing the complex number. When we divide the angle by n, we actually find n possible angles:

$$\alpha/n, \quad \alpha/n + 2\pi/n, \quad \alpha/n + 4\pi/n, \quad \ldots, \quad \alpha/n + (n-1)2\pi/n.$$

In other words, for every nonzero complex number there are exactly n nth roots, and they are equally spaced on a circle of radius $r^{1/n}$ centered at the origin.

1.2. First Proof of the Fundamental Theorem

We recall the statement of the Fundamental Theorem of Algebra:

THEOREM 1.2. *Suppose that $p(z)$ is a nonconstant polynomial whose coefficients are complex numbers. Then there is a complex number z_0 such that $p(z_0) = 0$.*

I learned the first proof from [**7**, H. Dörrie]. The first proof is based on three simple facts:

LEMMA 1.3. *If $c \in \mathbb{C}$ and if $n > 0$ is an integer, then there is a number d such that $d^n = c$.*

PROOF OF THE LEMMA. We use the basic fact about multiplication of complex numbers a and b: Express a and b in polar coordinates, so that $a = (|a|, \theta(a))_{pol}$ and $b = (|b|, \theta(b))_{pol}$. Then $a \cdot b = (|a| \cdot |b|, \theta(a) + \theta(b))_{pol}$. It follows immediately that $a^n = (|a|^n, n \cdot \theta(a))_{pol}$.

Consequently, the nth roots of $c = (|c|, \theta(c))_{pol}$ are the complex numbers

$$(|c|^{1/n}, \frac{\theta(c)}{n} + \frac{i}{n} \cdot 2\pi)_{pol} \quad \text{for } i = 0,\ 1,\ \ldots,\ n-1.$$

□

Lemma 1.4. *If $p(z)$ is a nonconstant complex polynomial, then* $\lim_{z\to\infty} p(z) = \infty$.

Proof. If $p(z) = a_n z^n + a_{n-1}^{n-1} + \cdots + a_1 z + a_0$, with $a_n \neq 0$ and $|z| > 1$, then

$$|p(z)| \geq (|z|^n \cdot |a_n| - |z|^{n-1} \cdot (|a_{n-1}| + \cdots + |a_1| + |a_0|) > M,$$

provided that

$$|z| \cdot |a_n| - (|a_{n-1}| + \cdots + |a_1| + |a_0|) > M/|z|^{n-1}.$$

□

Corollary 1.5. *If $p(z)$ is a complex polynomial, then there is a complex number z_0 so that $m = |p(z_0)|$ is the minimum value of $|p(z)|$.*

Proof. Let $m = \inf\{|p(z)| : z \in \mathbb{C}\}$. Let $z_1, z_2, \ldots$ denote a sequence of complex numbers such that $m = \lim_{n\to\infty} |p(z_n)|$. Then the sequence must be bounded because of the previous lemma. Hence there is a convergent subsequence, so that we may assume $z_n \to z_0 \in \mathbb{C}$. By continuity, $|p(z_0)| = m$. □

Lemma 1.6. *If $p(z)$ is a complex polynomial and $z_0 \in \mathbb{C}$, then there is a complex polynomial of the form $q(Z) = q(z - z_0)$ such that $p(z) = q(z - z_0)$.*

Proof. The coefficients of $q(z-z_0) = a_n(z-z_0)^n + \cdots + a_0$ are easily calculated by successive divisions (most efficiently implemented as synthetic division):

$$\begin{aligned} p(z) &= p_1(z) \cdot (z - z_0) + a_0, \\ p_1(z) &= p_2(z) \cdot (z - z_0) + a_1, \\ p_2(z) &= p_3(z) \cdot (z - z_0) + a_2, \\ p_3(z) &= p_4(z) \cdot (z - z_0) + a_3, \quad \text{etc.} \end{aligned}$$

Reversing the steps demonstrates the desired equality. For example,

$$\begin{aligned} 2z^3 + 17z^2 + 50z + 56 &= (2z^2 + 11z + 17) \cdot (z + 3) + 5, \\ 2z^2 + 11z + 17 &= (2z + 5) \cdot (z + 3) + 2, \\ 2z + 5 &= (2) \cdot (z + 3) - 1, \quad \text{and} \\ 2 &= (0) \cdot (z + 3) + 2, \end{aligned}$$

so that

$$2z^3 + 17z^2 + 50z + 56 = 2(z + 3)^3 - 1(z + 3)^2 + 2(z + 3) + 5.$$

□

First proof of the Fundamental Theorem of Algebra. Suppose that, to the contrary, the minimum value (Corollary 1.5) of $|p(z)|$ is $m = |p(z_0)| > 0$. By Lemma 1.6, we may translate the domain so that $z_0 = 0$. Multiplying $p(z)$ by a nonzero complex constant, we may assume that $|p(z_0)| = p(0) = 1$, so that $p(z)$ has the form

$$p(z) = 1 + b_k z^k + b_{k+1} z^{k+1} + \cdots + b_n z^n,$$

where b_k is the first nonzero coefficient after the constant 1.

The (nonzero) complex number $-1/b_k$ has a kth root d (1.3). Let λ denote a positive number very close to 0. Then $b_k(\lambda d)^k = -\lambda^k$ so that

$$|p(\lambda d)| \leq 1 - \lambda^k + \lambda^{k+1} \cdot (|b_{k+1} d^{k+1}| + \cdots + |b_n d^n|).$$

If λ is chosen so small that $\lambda(|b_{k+1}d^{k+1}| + \cdots + |b_n d^n|) < 1$, we conclude that $|p(\lambda d)| < 1$, a contradiction. We conclude that $m = 0$ so that $p(z)$ has a root. □

1.3. Second Proof

Our second proof is based on the geometrically intuitive idea of deforming or dragging a closed curve in the plane. The technical term for deforming or dragging is *homotopy.* The technical parts of the proof are more involved than our first proof of the theorem, but the concepts are powerful and allow significant generalization.

Definition 1.7. Maps $f, g : X \to Y$ are said to be *homotopic* if there is a continuous function $F : X \times [0,1] \to Y$ such that, for each $x \in X$, $F(x,0) = f(x)$ and $F(x,1) = g(x)$. The mapping F is called a *homotopy* from f to g.

The curves in question are the images of circles under the polynomial mapping $p : \mathbb{C} \to \mathbb{C} : z \mapsto p(z)$. Recall that $S_1 = \mathbb{S}^1$ denotes the circle of radius 1 in the complex plane centered at the origin. In complex notation,

$$\mathbb{S}^1 = \{e^{i\theta} \mid \theta \in \mathbb{R}\} = \{e^{i\theta} \mid \theta \in [0, 2\pi)\}.$$

Using inner products, we obtain for each positive number r a concentric circle $S_r = r \cdot \mathbb{S}^1 = \{r \cdot z \ : \ z \in \mathbb{S}^1\}$ of radius r. We may think of the point $S_0 = 0 \cdot \mathbb{S}^1$ as a degenerate curve that has been collapsed to a point. The polynomial mapping p provides a homotopy deforming the image of any one of these circles S_r to the point $p(0)$ via the definition

$$P_r : S_r \times [0,1] \to \mathbb{C} : (x,t) \mapsto p((1-t)x).$$

Note that P_r restricted to $S_r \times \{0\}$ is the same as p restricted to S_r, while P_r restricted to $S_r \times \{1\}$ is the constant map to $p(0)$.

We also use the fact already noted:

Lemma 1.8. *If $c \in \mathbb{C}$ has polar coordinates $c = (|c|, \theta(c))_{pol}$ and if n is a positive integer, then $c^n = (|c|^n, n\theta(c))_{pol}$.* □

Corollary 1.9. *The mapping $q : \mathbb{C} \to \mathbb{C} : z \mapsto z^n$, with $n > 0$ a positive integer, wraps the circle S_r n times around the circle S_s, where $s = r^n$.* □

We are now prepared to give the intuitive outline of our second proof of the Fundamental Theorem of Algebra. We start with an arbitrary nonconstant polynomial, and, dividing by the leading coefficient, find that we may assume it has the form $p(z) = z^n + a_{n-1}z^{n-1} + \cdots + a_1 z + a_0$. We assume that, contrary to the theorem, $p(z)$ is never 0. Then the homotopy $P_r : S_r \times [0,1] \to \mathbb{C}$ actually has image in $\mathbb{C} \setminus \{0\}$. That is, the curve $p : S_r \to \mathbb{C} \setminus \{0\}$ can be dragged to a point without hitting the origin. But we shall see that, for r sufficiently large, the curve $q : S_r \to \mathbb{C} \setminus \{0\}$ of the previous corollary is homotopic in $\mathbb{C} \setminus \{0\}$ to $p : S_r \to \mathbb{C} \setminus \{0\}$. That is, $q|S_r$ can be dragged to $p|S_r$ missing the origin, and from there, using the polynomial p, can be dragged to the constant map missing the origin. That is, the curve that maps n times around the origin can be dragged to a point without going through the origin. This is physically absurd since the curve $q|S_r$ is most thoroughly hooked about the origin. This completes the intuitive proof of the theorem.

The difficulty that remains is that of turning these statements into technical mathematics. The necessary details are supplied by the following two theorems.

THEOREM 1.10. *If $q(z) = z^n$ and $p(z) = z^n + a_{n-1}z^{n-1} + \cdots + a_1 z + a_0$, then, for all sufficiently large $r > 1$, the curves $q|S_r$ and $p|S_r$ are homotopic in $\mathbb{C} \setminus \{0\}$.*

PROOF. Define $F : S_r \times [0,1] \to \mathbb{C}$ by the formula $F(s,t) = (1-t) \cdot q(s) + t \cdot p(s)$. Then F is definitely a homotopy from $q|S_r$ to $p|S_r$ in $\mathbb{C}$. We need to show that $F(s,t)$ is never 0, provided that r is big enough:

$$\begin{aligned} |F(s,t)| &= |(1-t)s^n + t(s^n + a_{n-1}s^{n-1} + \cdots + a_1 s + a_0)| \\ &\geq r^n - r^{n-1}(|a_{n-1}| + \cdots + |a_1| + |a_0|) > 0, \end{aligned}$$

provided that $r > (|a_{n-1}| + \cdots + |a_1| + |a_0|)$. □

THEOREM 1.11. *The map $q|S_r$ is not homotopic to a constant map in $\mathbb{C} \setminus \{0\}$.*

PROOF. This theorem is the technically complex bit of our second proof of the Fundamental Theorem of Algebra.

Suppose that, contrary to the theorem, there is a continuous function $Q : S_r \times [0,1] \to \mathbb{C} \setminus \{0\}$ such that, for each $s \in S_r$, $Q(s,0) = q(s) = s^n$ and $Q|S_r \times \{1\}$ is a constant map.

We divide the product $S_r \times [0.1]$ into tiny triangles, as, for example, in Figure 5.

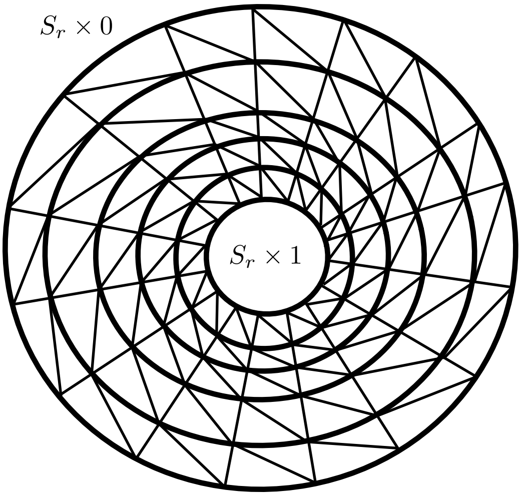


FIGURE 5. Dividing a ring into layers of triangles

Since the set $S_r \times [0,1]$ is compact and its image misses 0, there is a positive distance from the image to the origin. The map Q is uniformly continuous, so that, if the triangles are small, so are the images of those triangles. We may move the images of the vertices a tiny bit so that the images of the vertices of each triangle are vertices of a rectilinear triangle in the plane. We replace each (possibly curvilinear and singular) image triangle by the corresponding rectilinear triangle spanned by the new image vertices.

We take care in the adjustment that the following conditions are satisfied. The vertex images from $S_r \times \{0\}$ are not moved. The image of the entire set $S_r \times \{1\}$ is still very small. No adjusted triangle hits the origin 0.

We now pick a ray R from 0 to ∞ in the plane in such a way that it misses all of the new vertex images and misses the (very small) image of $S_r \times \{1\}$. We lose no generality in assuming that R is the positive x-axis.

We complete the proof by counting the number of times various triangle edges intersect and cross the ray R. We orient the edges of $S_r \times \{0\}$ in the counterclockwise direction. The image of $S_r \times \{0\}$ intersects R exactly n times, and each time crosses R from the lower side to the upper side. We assign the crossing number $+1$ to each oriented edge that crosses R from the bottom to the top. When we later come upon oriented edges that cross R from top to bottom, we assign such crossings the crossing number -1. See Figure 7. If an edge misses R, we assign it number 0. The sum of crossing numbers for the image of $S_r \times \{0\}$ is n.

We delete triangles one at a time, layer by layer, as indicated by Figure 6. We first remove the green triangles, then the orange triangles, etc.

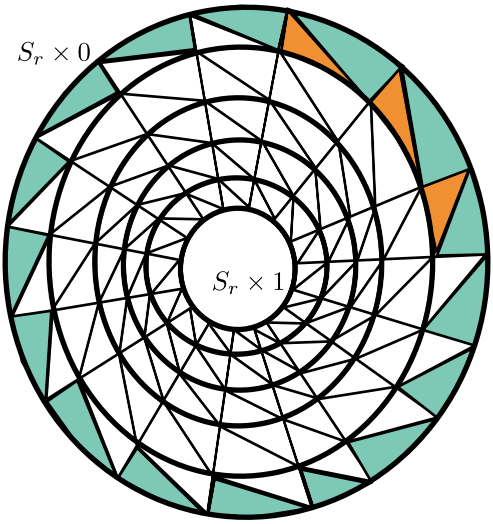


Figure 6. Peeling away the layers

Each time a triangle is deleted, we obtain a new boundary curve. Either one oriented boundary edge is replaced by two new oriented boundary edges or two oriented boundary edges are replaced by one new oriented boundary edge. It is easy to see that the algebraic sum of assigned crossing numbers remains unchanged with each triangle deletion since the sum obviously remains unchanged on each triangle. In order to see this, examine Figure 7.

Eventually, one moves through the diagram from the image of $S_r \times \{0\}$, which has algebraic sum n, to the image of $S_r \times \{1\}$, which has algebraic sum 0, a contradiction since the algebraic sum remains unchanged.

With this contradiction, the proof of the theorem is complete. □

This argument completes our second proof of the Fundamental Theorem of Algebra.

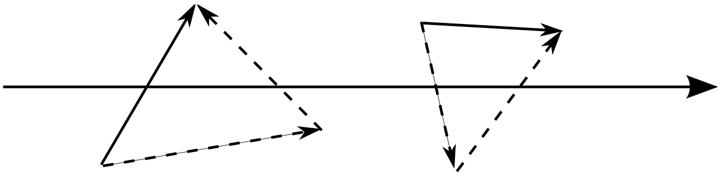


FIGURE 7. Calculating intersection numbers

1.4. Exercises

Abel and Galois proved that there are no formulas like the quadratic formula for finding the roots of polynomials with real or complex coefficients of degrees ≥ 5. The Fundamental Theorem of Algebra, for which we have just given two proofs, shows that there are always roots, in fact, if we count possible multiplicity of roots, exactly as many roots as the degree of the polynomial. The problem becomes one of finding the roots, or at least approximating the roots to a desired degree of accuracy. There is a large literature concerning efficient ways of finding the roots, some of which have been programmed into most pocket calculators. In the following exercises, we suggest exploring one of the standard methods, namely, Newton's method.

1.1. (The equation of the tangent line) Let $p(x)$ be a polynomial with real coefficients and let x_n be some real number. Find the equation of the tangent line to the graph of $p(x)$ at the point $(x_n, p(x_n))$.

Answer: $y = p(x_n) + p'(x_n) \cdot (x - x_n)$.

1.2. Find the point x_{n+1} at which the tangent line crosses the x-axis.

Answer: $x_{n+1} = x_n - p(x_n)/p'(x_n)$.

Newton's method: Newton notes that if x_n is a root of $p(x)$, then $x_{n+1} = x_n$. That is, a root of $p(x)$ is a fixed point of the operation $x_n \mapsto x_{n+1}$, and when x_n is very close to a root r of $p(x)$, then x_{n+1} is much closer to r.

1.3. Use Newton's method to approximate $\sqrt{5}$. (That is, find an approximate root to the equation $x^2 - 5 = 0$.)

1.4. Apply Newton's method in an attempt to find a root of the equation $x^2 - 2x + 10 = 0$. Why does the method fail?

1.5. Use the quadratic formula to find the two roots of $x^2 - 2x + 10$.

Newton's method often works even when there are no real roots. Use the same iterative formula $x_{n+1} = x_n - p(x_n)/p'(x_n)$, but start the process with x_0 equal to some nonreal complex number. This requires the ability to do complex arithmetic.

1.6. Apply Newton's method to the polynomial $x^2 - 2x + 10$ but with x_0 equal to the complex number i.

1.7. Construct a computer program to color the plane. Starting with an initial complex number x_0 representing a pixel in the plane, iterate Newton's process enough times that the result seems to be getting close to one of the roots of the polynomial and color the original pixel a color assigned to that root. If the method doesn't seem to converge given that initial value, leave that pixel uncolored. Iterate the process with different initial pixels.

CHAPTER 2

The Brouwer Fixed Point Theorem

As the climax of this chapter, we will prove the Brouwer Fixed Point Theorem. Our proof will be complete only in dimension 2, but the technique we use is valid in all dimensions and lacks only a tiny bit of technique to complete, technique that we will outline in exercises. This theorem is one of the two basic tools used in the proof of existence of solutions to differential equations, the other being the contraction mapping principle. A generalized version of the theorem is used in game theory to prove the existence of optima.

In order to completely understand this chapter, the reader should have a reasonable grasp of limits, continuity, open sets, and closed sets in Euclidean space.

2.1. Statement of the Theorem

First we recall some definitions.

Definition 2.1. The n-dimensional ball $\mathbb{B}^n$ in n-dimensional Euclidean space $\mathbb{R}^n$ is the set

$$\mathbb{B}^n = \{x = (x_1, x_2, \dots, x_n) \in \mathbb{R}^n \,:\, |x|^2 = x_1^2 + x_2^2 + \cdots + x_n^2 \leq 1\}.$$

The boundary of the n-ball $\mathbb{B}^n$ is the $(n-1)$-sphere $\mathbb{S}^{n-1}$, where

$$\mathbb{S}^{n-1} = \{x = (x_1, x_2, \dots, x_n) \in \mathbb{R}^n \,:\, |x|^2 = x_1^2 + x_2^2 + \cdots + x_n^2 = 1\}.$$

Thus the 1-ball is the unit interval of “radius 1” and length 2, the 0-sphere is a pair of points, the 2-ball is the circular disk of radius 1, and the 1-sphere is the circle of radius 1. The 3-ball is the solid ball of radius 1 in Euclidean 3-dimensional space. The 2-sphere is the surface of that ball. Up to *homeomorphism* (topological equivalence) there are many realizations of the ball and sphere. Every (finite closed) interval is homeomorphic to the unit interval. The planar triangle is homeomorphic to the circle, the planar triangle together with its interior is homeomorphic to the circular disk of radius 1, etc.

Theorem 2.2 (Brouwer Fixed Point Theorem). *If $f : \mathbb{B}^n \to \mathbb{B}^n$ is a continuous function, then there is at least one point $x \in \mathbb{B}^n$ such that $f(x) = x$. Such a point is called a* fixed point *of the function f.*

Exercise 2.3. Prove the 1-dimensional version of the fixed point theorem: If $f : [0, 1] \to [0, 1]$ is a continuous function, then there is at least one real number $x \in [0, 1]$ such that $f(x) = x$.

A version of the 1-dimensional theorem was exhibited as a puzzle in a magazine story, related to me, that apparently appeared many, many years ago: A monk set out at 6:00 AM to take the trail to a retreat at the top of the mountain, which he reached early in the afternoon and where he prayed and worshiped through the night. The next morning he left the mountain top, again at 6:00 AM and returned down the trail to his monastery at the foot of the mountain. Explain why there was a time of day and a place on the trail so that the monk was at that place on the trail at that same time on each of the two days. [The answer to the puzzle given in the story was "The monk had to meet himself."]

The proof of the Brouwer Fixed Point Theorem, as well as the proof of many theorems in topology, relies on the introduction of extra structure to the problem. We will illustrate the value of additional structure in a series of arguments. The first will be an elementary puzzle involving the checkerboard. The second will use a childhood puzzle to motivate the use of the sign of a permutation; this notion will be important when we treat the mathematical notion of right and left, or *orientation* in classifying 2-dimensional surfaces. The third will show how to justify the (?) obvious (?) fact that a polygonal closed curve in the plane separates the plane into two pieces, an inside and an outside; this argument will introduce the powerful topological tool of *general position*. The idea of general position, together with a trick called *the one-ended arc trick*, will be used to prove the famous No Retraction Theorem. Finally, we will easily deduce the Brouwer Fixed Point Theorem from the No Retraction Theorem.

2.2. Introducing Extra Structure into a Problem

Often we understand a problem by imposing extra structure. The simplest extra structure we can impose is that of *counting*. Many theorems in 2-dimensional topology are proved by clever yet sophisticated counting arguments or by introducing extra structure. After considering these two elementary problems in this subsection, we shall move on to some rather famous graduate level problems that can be solved by fairly elementary, yet sophisticated, techniques that involve imposing extra structure and observing the geometric consequences of that structure.

Sherlock Holmes: *How often have I said to you that when you have eliminated the impossible, whatever remains,* however improbable*, must be the truth?* - The Sign of Four, Chapter 6.

2.3. Two Elementary Problems

Problem 2.4. I learned about this problem when I was a high school student. It appeared in a paperback book that I read and reread. I thank that unknown author and very interesting book. I lost the book many decades ago. Consider a square 8x8 board of squares. See Figure 1. Remove two corner squares from the board, diagonally across the board from each other.

Is it possible to tile the remaining squares with dominoes of the shape 1x2 (Figure 2, using those dominoes both vertically and horizontally?

We will give a solution after we describe the second problem.

Problem 2.5. Before the 3-dimensional Rubik's Cube, there was the 2-dimensional Fifteen Puzzle.

The puzzle consists of fifteen tabbed and slotted plastic tiles set in a plastic case, with one blank slot (pictured as black) into which adjacent tiles can be slid. See Figures 3 and 4.

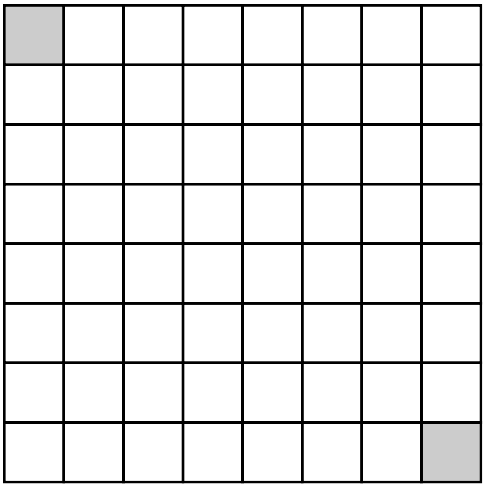


Figure 1. The uncolored board with two corners deleted

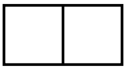
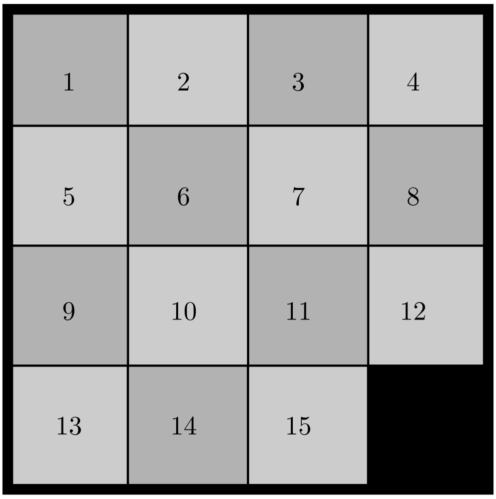


Figure 2. The uncolored domino

The goal was to transform a scrambled arrangement of tiles into the unscrambled version. This puzzle was, for a time, the rage — just as the Rubik's Cube had its period — and the puzzle is still manufactured and sold as a children's diversion. I played with the puzzle at church when I was a child when I was bored with the sermon. Puzzle enthusiasts began to ask whether all arrangements of tiles could be realized. In particular, they tried unsuccessfully to transform the standard arrangement into the reverse arrangement. See Figure 5.

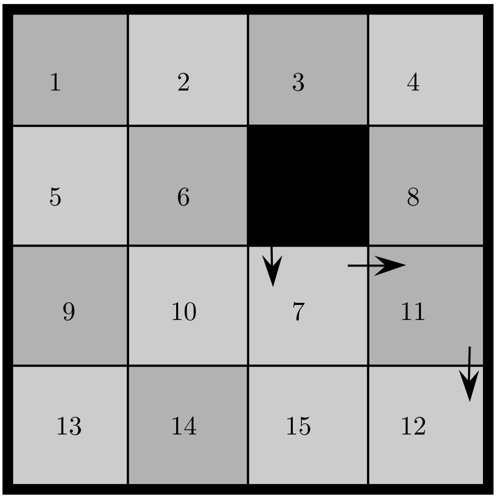
Books were written about the puzzle. A museum at the Indiana University, Bloomington, collects and displays the wonderful artifacts concerning this puzzle. Why, or why not, can the reverse position be realized?

2.3.1. Solution to the Tiling Problem. First approximation to a solution for the domino tiling problem. The board with corners deleted has 62 squares, which is even. The domino has 2 squares, which is even. Thus, any solution would require exactly $31 = 62/2$ dominoes.



1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	

Figure 3. The 15 Puzzle



1	2	3	4
5	6		8
9	10	7	11
13	14	15	12

Figure 4. Tile slides

Second approximation for the tiling problem. Introduce extra structure to both the board and to the domino by coloring every other square black. See Figures 6 and 7.

Solution to the tiling problem: Every domino, whether horizontally placed or vertically placed, will cover one white and one black square. Thus, if the tiling exists, there must be the same number of black and white squares. But there are 32 black squares and only 30 white squares. Therefore the tiling is impossible.

2.3.2. Solution to the 15 Puzzle. Solution to the 15 puzzle: As a matter of fact, exactly half of the potential arrangements of tiles can be realized, and the situation is explained by a simple, but useful, mathematical notion, namely the

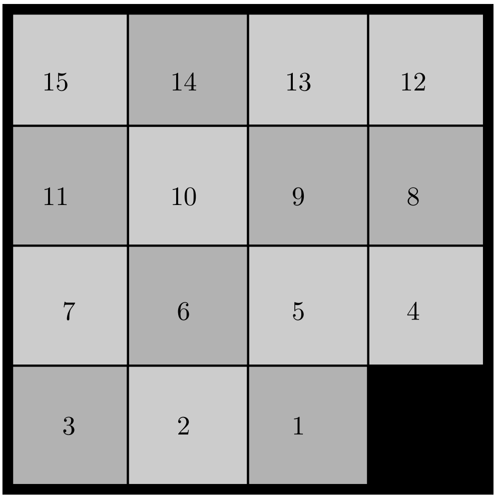


Figure 5. The 15 puzzle with numbers reversed

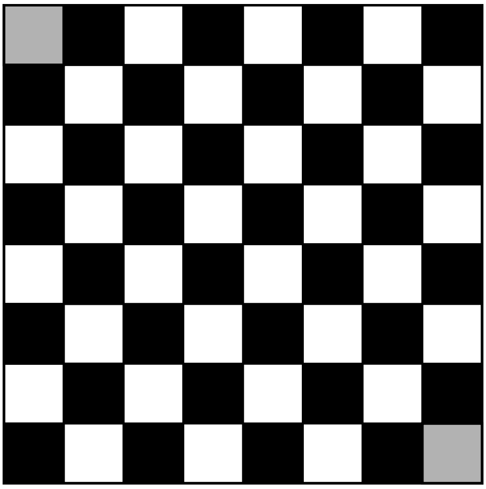


Figure 6. The colored board with two corners omitted

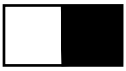


Figure 7. The colored domino

sign or parity of a permutation — is a permutation *even* (plus sign) or *odd* (minus sign)?

Definition 2.6. Let $S_n = \{1, 2, \ldots, n\}$ denote a finite set with n elements. A *permutation* of S_n is a reordering p of the set. Equivalently, a permutation is a bijection $p : S_n \to S_n$.

There are standard ways of picturing a permutation p. For example, the 2-*row form*

$$p = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 7 & 2 & 4 & 3 & 1 & 8 & 5 & 6 \end{pmatrix}$$

exhibits each i above its image $p(i)$, so that, for example, $p(1) = 7$ and $p(7) = 5$. The *disjoint cycle form*

$$(1\,7\,5)(2)(3\,4)(6\,8)$$

indicates that p naturally divides the elements of S_n into *circles or cycles*:

$$p : 1 \mapsto 7 \mapsto 5 \mapsto 1,$$

$$p : 2 \mapsto 2,$$

$$p : 3 \mapsto 4 \mapsto 3,$$

$$p : 6 \mapsto 8 \mapsto 6.$$

The cycle $(1\,7\,5)$ is called a 3-cycle; the cycle (2) is a 1-cycle, and the cycles $(3\,4)$ and $(6\,8)$ are 2-cycles. The 1-cycles of a permutation are often omitted in the disjoint-cycle form. A 2-cycle is also called a *transposition*.

Each cycle can itself be considered as a permutation (a function), where all elements of S_n not explicitly mentioned are to be considered as (omitted) 1-cycles. Since functions are typically composed from right to left, cycles (as functions) can also be composed, whether they are disjoint or not, to form new permutations. For example,

$$(1\,3\,5\,7)(2\,5\,3) = (2\,7\,1\,3)(5).$$

Theorem 2.7. *Every permutation of S_n can be realized as a product of 2-cycles (transpositions).*

Proof. It suffices to show that a cycle $(a_1\,a_2\,\ldots\,a_k)$ is a product of 2-cycles, but

$$(a_1\,a_2\,\ldots\,a_k) = (a_1\,a_k)(a_1\,a_{k-1})\cdots(a_1\,a_3)(a_1\,a_2).$$

□

Definition 2.8. A permutation is *even* if it can be expressed as the product of an even number of 2-cycles. A permutation is *odd* if it can be expressed as the product of an odd number of 2-cycles.

Definition 2.9. Suppose that p is a permutation of S_n. Consider the unordered pairs (i, j). We say that (i, j) is *inverted* by p if p changes the order of i and j; that is, (i, j) is inverted by p if either ($i < j$ and $p(j) < p(i)$) or ($j < i$ and $p(i) < p(j)$). Let $\mathrm{inv}(p)$ denote the number of unordered pairs (i, j) that are inverted by p.

Here is an example, namely the permutation

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 5 & 1 & 6 & 3 & 7 & 4 & 2 \end{pmatrix}.$$

If we draw arrows from each integer to its image spot, then the number of inversions is the number of crossings of these arrows. See Figure 8.

Theorem 2.10. *The permutation p is even iff $\mathrm{inv}(p)$ is even. The permutation p is odd iff $\mathrm{inv}(p)$ is odd. (Consequently, no permutation is both even and odd.)*

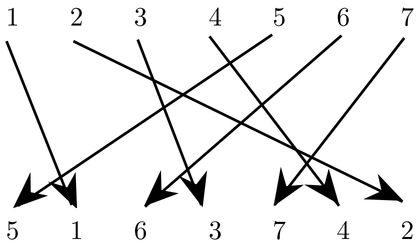


FIGURE 8. The inversions of a permutation

PROOF. It suffices to show that each 2-cycle changes the number of inversions by an odd number. We start with the 2-row representation

$$\begin{pmatrix} 1 & \dots & i & (i+1) & \dots & (k-1) & k & \dots & n \\ p(1) & \dots & p(i) & p(i+1) & \dots & p(k-1) & p(k) & \dots & p(n) \end{pmatrix}$$

and compose with the transposition $(p(i)\, p(k))$, with $i < k$. The result, after the composition is

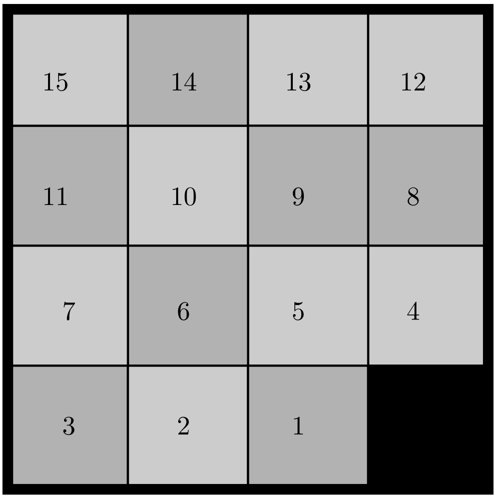
$$\begin{pmatrix} 1 & \dots & i & (i+1) & \dots & (k-1) & k & \dots & n \\ p(1) & \dots & p(k) & p(i+1) & \dots & p(k-1) & p(i) & \dots & p(n) \end{pmatrix}.$$

If $k = i + 1$, so that a_i and a_k are adjacent before the move, then the only pair whose status of inversion is affected by the transposition is the pair (i, k): if this pair is inverted before the transposition, then it is in order after; if it is in order before the transposition, then it is inverted after. Thus the parity of $\text{inv}(p)$ is changed by this adjacency transposition. If there are $\ell = k - i - 1$ elements between a_i and a_k, then ℓ adjacency moves can move a_i so that it is in position $k-1$ adjacent to a_k, with ℓ changes in parity. Then a_i and a_k can be interchanged, with a change of 1 in parity. Finally, a_k can be moved back to position i by another ℓ adjacency switches, with ℓ changes in parity. The result is $2\ell + 1$ parity changes, an odd number of changes. Hence, the original transposition resulted in a change in the parity of $\text{inv}(p)$. □

We are now in the position to show that the reverse position cannot be reached from the initial position.

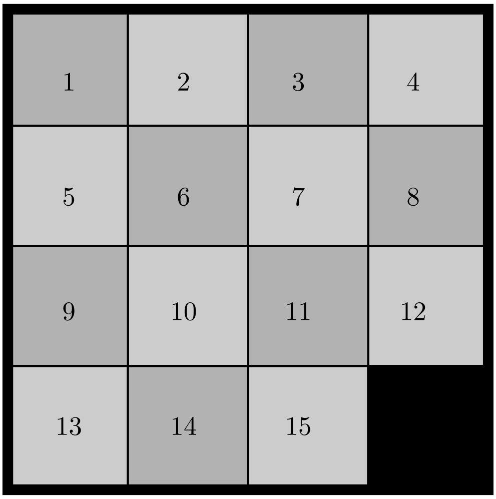
We think of the blank slot as a tile labelled 16. We can view each position of the puzzle as a permutation of the numbers 1 through 16. Each move in the puzzle is a transposition that interchanges 16 with another tile. Hence, each move changes the parity of the permutation. The tile 16 begins in a position in the checkerboard pattern that is dark. Each move takes 16 from a dark position to a light position, or vice versa. Hence, the position of 16 in the checkerboard pattern indicates whether the permutation is even or odd: dark position = even permutation; light position = odd permutation.

In the reverse position, tile 16 is in dark position (Figure 9); hence the permutation, if attainable, must be even. We count the number of inversions. (Compare with the original position, Figure 10. The number 15 is inverted with fourteen tiles, the number 14 with 13 tiles, the number 13 with 12 tiles, etc. That is, the number



15	14	13	12
11	10	9	8
7	6	5	4
3	2	1	

Figure 9. The reversed position again



1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	

Figure 10. The initial position again

of inversions is

$$14 + 13 + 12 + 11 + 10 + 9 + 8 + 7 + 6 + 5 + 4 + 3 + 2 + 1,$$

which is odd. Thus this position is not attainable.

Exercise. *Show that a tile position can be attained iff the parity of the permutation agrees with the parity dictated by the position of tile* 16*.*

2.4. Three Advanced Problems

We give a sequence of applications to geometry, all referring to important properties of the 2-dimensional plane $\mathbb{R}^2$. We are interested in the ways a circle, which we denote by $\mathbb{S}^1$, can be placed in the plane. **For the remainder of this section, it is necessary to know about continuous functions.**

2.4.1. Polygonal Simple Closed Curves in the Plane.

Definition 2.11. A simple closed curve in the plane is the image of a continuous function $J : \mathbb{S}^1 \to \mathbb{R}^2$ from the circle $\mathbb{S}^1$ into the plane $\mathbb{R}^2$, where the function J is one to one (that is, has no self-intersections).

For the next few theorems, we shall assume that the simple closed curve is polygonal, that is, formed by a finite sequence of straight line segments. Figure 11 gives two examples, the first very simple, the second rather elaborate.

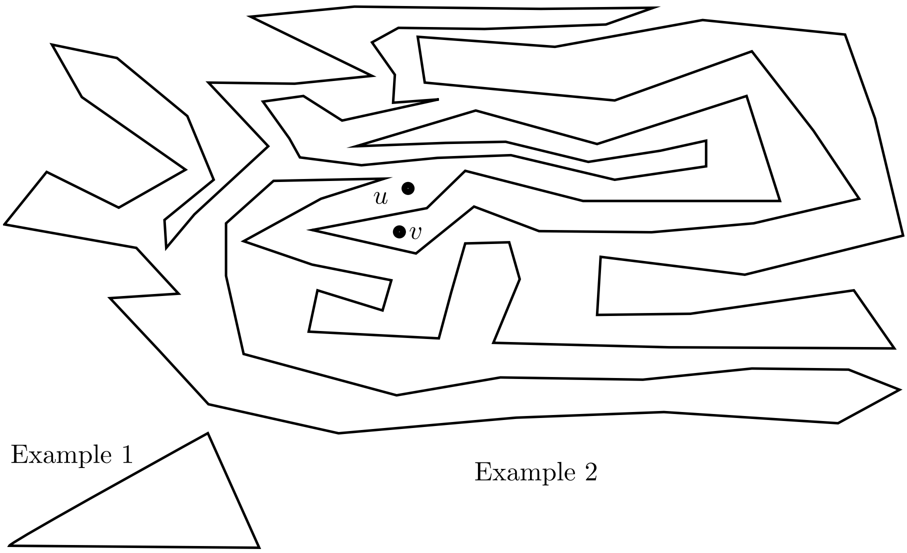


Figure 11. The simple and the complex simple closed curve

For Example 1, the triangle, there is evidently both an inside and an outside to the curve. This is likewise true for a round circle. It is intuitively obvious that every convoluted "circle" (simple closed curve) in the plane should have an inside and an outside, but the proof that it is so becomes complex as we observe Example 2 and ask ourselves whether the points u and v that we have marked are inside the curve or outside the curve.

Exercise 2.12. Is the point v inside the curve? The point u?

For a polygonal simple closed curve, there are at most two pieces in the complement $\mathbb{R}^2 \setminus J$ of J, for, locally, there are only two sides (Figure 12), and, as one traverses the curve, the local sides near one point are connected to local sides of surrounding points.

The question is only whether, when we follow those two local sides around the full length of J, does the black Side 1 come back to connect again with the Side 1, or does the black Side 1 come back to connect with grey Side 2 to form a twisted Möbius band. We will give a proof of the following theorem. Although the theorem seems obvious, the proof is not trivial and the techniques of its proof will be used to prove the very important No Retraction Theorem and Brouwer Fixed Point Theorem.

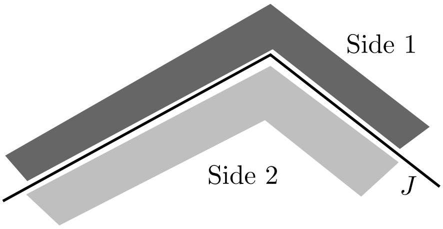


Figure 12. The two local sides of a polygonal simple closed curve

Theorem 2.13. *Each polygonal simple closed curve J in the plane $\mathbb{R}^2$ separates the plane. That is, the complement $\mathbb{R}^2 \setminus J$ of J in $\mathbb{R}^2$ is not connected. In fact, it has exactly two pieces (called* components*), an inside (called the* interior*) and the outside (called the* exterior*).*

Remark. The theorem is completely obvious for the triangle, and our proof will make use of that fact. The proof that we will give is elementary in the sense that it uses only basic geometric ideas, but the logic involved invokes ideas that, in fully developed form, are very important in the development of geometric and algebraic topology. We will try to point out the critical ideas as we go along. *End remark.*

Here are a few of the basic ideas:

Idea (1) A triangle T in the plane has both an inside and an outside. Every time you cross the boundary of the triangle you pass from outside to inside, or from inside to outside. See Figure 13. If we traverse a polygonal simple closed curve J that crosses ∂T at every intersection point, then every crossing point where J crosses into T is followed by a subpath of J in $J \cap T$ that is terminated by a point where J crosses the boundary ∂T out of T.

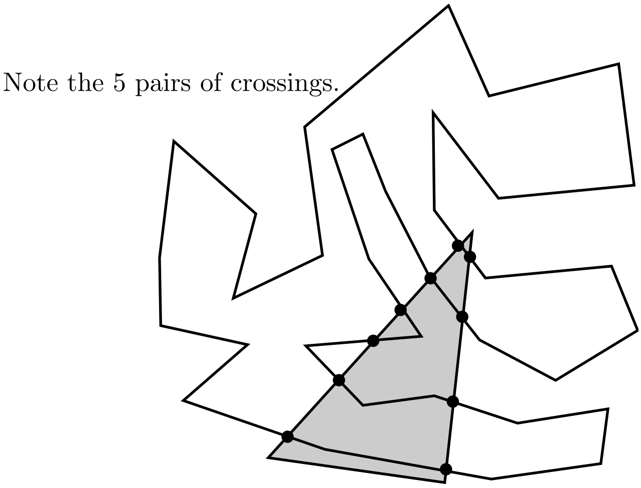


Figure 13. Curve crossing a triangle

The components (pieces) of $J \cap T$ pair the points of $J \cap \partial T$.

Idea (2) If T_1, T_2, …, T_n are triangles in the plane $\mathbb{R}^2$, and if J is a polygonal simple closed curve in the plane, then J can be translated an arbitrarily small amount so that it misses each vertex of each T_i and crosses the boundary of each T_i at each point of $J \cap \partial T_i$. This is a very simple case of an exceedingly important property called *general position.*

Idea (3) Every finite edge path that has at least one end point has, in fact, two end points. The only other alternative for the path is that it form a closed path (initial point = terminal point) with no end points.

Proof (theorem). The proof is called the *one-ended arc trick.* We will show that, unless the theorem is true, there is a finite edge path that has only one endpoint, in contradiction to Idea (3) above. We therefore begin the proof by assuming the ridiculous fact that the curve J does not separate the plane.

The proof requires that we introduce *extra structure* in our picture. This structure will take place in two different copies of the plane. We call the plane that contains our polygonal simple closed curve J the *image plane.* The other plane that we introduce we will call the *model plane.*

If our curve J contains n vertices, then we shall first construct n model triangles in the model plane and then n corresponding image triangles in the image plane. See Figures 14 and 15. After we have constructed the image triangles, we will construct a second polygonal simple closed curve K in the image plane that is in general position with respect to each of the n image triangles. See Figure 16. Corresponding to each intersection arc of K with an image triangle, there is a corresponding arc in the model triangle. It is in this collection of model arcs that we will discover a one-ended arc, an obvious contradiction. We will conclude that J must separate the image plane. (I find such an argument completely wild! I love it.)

We first introduce the triangles in the model plane. If our first polygonal simple closed curve has n vertices, then we pick n vertices on the unit circle $\mathbb{S}^1$ in the model plane and join them by straight arcs to form a polygonal simple closed curve. We then add a radius of the circle from each vertex A to the center V of the circle. We obtain a model array of triangles ABV inside the disk of unit radius. It is in this model that we will trace certain important paths:

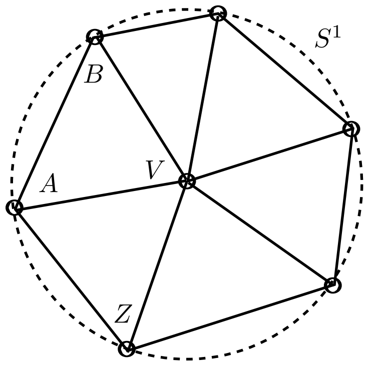


Figure 14. The model disk

We map this model disk of triangles into the plane in the following way. We map the outer boundary of the model to our original polygonal simple closed curve, vertex to vertex, edge to edge. We map the center V of our model to some rather arbitrary point V' in the plane, subject only to the condition that it not be in any of the lines defined by the edges of J. Then, given any edge e of J with vertices A' and B', the points A', B' and V' are the vertices of a triangle $A'B'V'$ in the plane. The corresponding points A, B, and V of our model bound a triangle ABV in our model. We map the image triangle $A'B'V'$ to the model triangle ABV linearly.

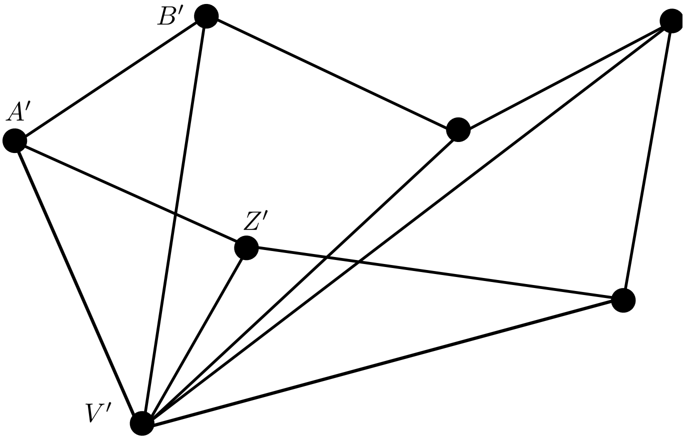


FIGURE 15. The image of the model disk

While the model triangles intersect only along edges, the image triangles do intersect along the corresponding common edge but may also fold back over each other along that edge. For example, the triangles $A'B'V'$ and $Z'A'V$ fold over each other in the figure.

We have now completed the introduction of triangles into our problem. We have still one additional structure to add to our picture. This is the addition of a second polygonal simple closed curve in the image plane.

Since we are assuming that J does not separate the image plane, it does not separate points just opposite each other across the arc $A'B'$. We may thus construct a polygonal simple closed curve K in the image plane that intersects J only at one point $X' \in A'B'$ where it crosses J.

This construction completes the added structure to our problem: n triangles in the model plane, n triangles in the image plane, and a polygonal simple closed curve K in the image plane.

2.4.2. The One-sided Arc Trick.

We now examine the way K intersects each of the image triangles. After a slight translation, we may assume that K misses all of the image vertices and crosses each triangle edge at each point of intersection (general position, Idea (2)).

Thus by Idea (1), if K intersects a triangle, it does so in a finite collection of polygonal arcs, each with its endpoints on the boundary of the triangle. The corresponding model triangle likewise is crossed by a corresponding finite collection of polygonal arcs (Figure 17), pushed over from the image triangle by the linear correspondence between the two triangles.

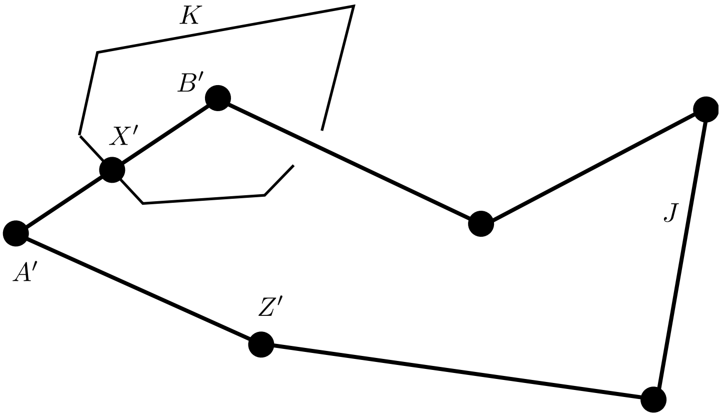


Figure 16. The curve K

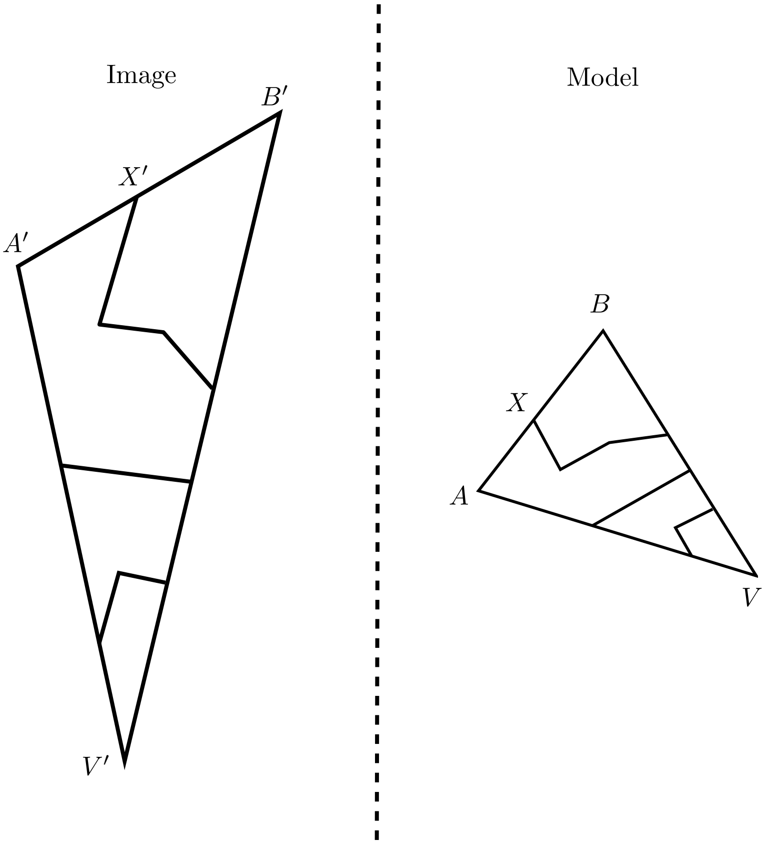


Figure 17. Pull-back arcs in the model disk

The important property of these crossing arcs is the following. Whether adjacent triangles in the image plane fold back on each other or not, any intersection arc from the one triangle that meets the common edge extends into the neighboring triangle at that point. If the triangles did not fold on each other, then the arc is simply an extension into the adjacent triangle. If the triangles did fold on each other, then the arc enters the common edge in one triangle and folds right back on top of itself in the adjacent triangle.

As a conseqence, in the model plane the intersection arcs form nonsingular curves, which can only be finite edge paths or polygonal simple closed curves.

The key observation is this: Look at the intersection path in the model that begins at the point X. Because X is an endpoint of that path, the path must have a second endpoint by Idea (3). That endpoint must lie on a triangle edge and that edge cannot be an interior edge of the model because intersection paths continue at every interior edge. But no other boundary point is available since X' is the only point at which K hits a boundary edge.

We have discovered a one-ended arc, a contradiction. □

2.4.3. The Brouwer Theorem and the No Retraction Theorem. The theorem we have just proved is both "obvious" and "uninteresting"—obvious because we have a hard time imagining that it could be false, and uninteresting because there is a much more general theorem that does not assume the simple closed curve is polygonal. Much of the content in a beginning algebraic topology course was developed in order to prove the general case of this theorem and other related facts. One generalization is the Riemann Mapping Theorem in the theory of complex variables. Riemann's proof had some gaps that required about 50 years of effort by a cadre of mathematicians to fill. Riemann's theorem is an analytic generalization of this theorem. We will return to general versions of this theorem using the techniques of geometric, set theoretic, and algebraic topology in later chapters. At that point we will prove theorems with famous names: the Jordan Curve Theorem, the Schoenflies Theorem, the Zippin Characterization of the Sphere, the R. L. Moore Decomposition Theorem, and others.

But at this point we want to enjoy the beauty of the one-ended arc proof for a while and use it to prove other results. I remember asking R. H. Bing what he could possibly do with a certain unresolved problem called the Free Surface Conjecture if he could resolve it. He said, "Oh, I think that I'd enjoy it for a while."

Here are a couple of exercises for the reader. The first is rather direct from what we have just done. The second is a generalization to dimension 3 and requires some imaginative thought. In particular, it requires that we understand what general position would mean in dimension 3, and that we understand how two triangular disks in dimension 3 could be expected to intersect.

Exercise 2.14. Prove that there is no polyhedral Möbius strip in the Euclidean plane.

Exercise 2.15. Show that every polyhedral 2-dimensional surface S in 3-dimensional Euclidean space $\mathbb{R}^3$ separates $\mathbb{R}^3$.

It requires less imagination to prove the 2-dimensional version of the famous Brouwer Fixed Point Theorem.

Theorem 2.16 (Brouwer Fixed Point Theorem). *If D is the unit square disk, and $f : D \to D$ is a continuous function that maps D into itself, then there is at least one point $x \in D$ such that $f(x) = x$.*

We will deduce the 2-dimensional Brouwer Fixed Point Theorem from the 2-dimensional version of an equally famous theorem called the No Retraction Theorem. We will prove the No Retraction Theorem by the one-ended arc trick.

THEOREM 2.17 (No Retraction Theorem). *If* $\mathbb{B}^n$ *is the unit* n*-dimensional ball (the ball of radius* 1*), then it is impossible to find a continuous function* $f : \mathbb{B}^n \to (\mathbb{S}^{n-1} = \partial\mathbb{B}^n)$ *such that, for each* $x \in \mathbb{S}^{n-1}$*,* $f(x) = x$*. Such a map, if it existed, would be called a* retraction.

PROOF OF THE NO RETRACTION THEOREM IN DIMENSION 2. In place of the round disk $\mathbb{B}^2$ we use the square disk D to which it is topologically equivalent. As in our previous proof using the one-ended arc trick, we make the ridiculous assumption that the theorem is false. Hence there is a continuous function $f : D \to \partial D$ such that, for each $x \in \partial D$, $f(x) = x$.

Again, we consider two planes: the model plane and the image plane. In the symbol $f : D \to \partial D$, we view the first D to be in the model plane and the second D to be in the image plane. As before, we add structure to the model disk D by subdividing it into very tiny triangles. See Figure 18.

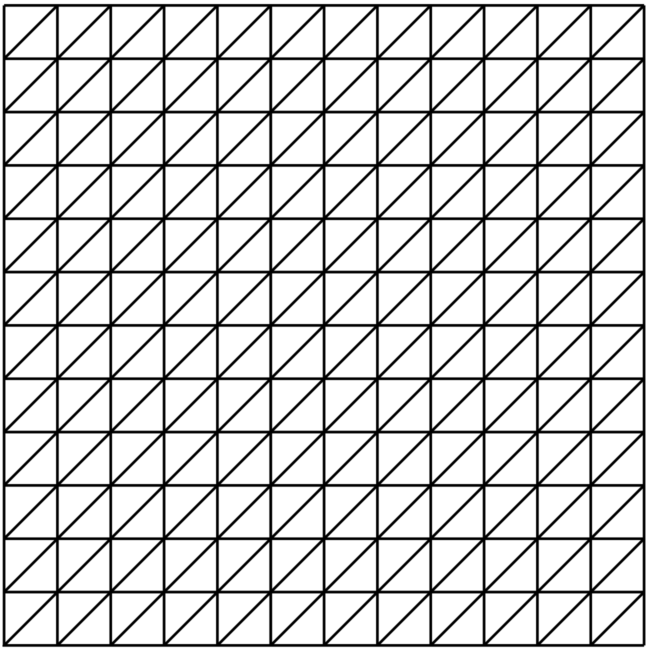


FIGURE 18. The model triangulated square

We now use the continuous function f to carry the triangles of the square in the model plane over to tiny triangles in the image plane. If $T = ABC$ is any of the tiny triangles, with vertices A, B, and C, then the images $f(A)$, $f(B)$, and $f(C)$ will be three points of ∂D that are very close to one another. We may move them a tiny bit so that they are still near one another and still near ∂D, yet they are the vertices of an actual triangle in the image plane. This process does not require that we move any vertices that were originally in ∂D. In this manner, we modify all those image vertices that must be moved. We may replace the original map f with a new continuous function $g : D \to \mathbb{R}^2$ that takes each triangle of the model space linearly to the corresponding triangle of the image space. The new map perhaps does not take D into ∂D, but it certainly takes D into the complement of the center of the disk and does not move any point of ∂D.

We now have to add to the structure a path R analogous to the polygonal simple closed curve K in the separation theorem. For R we take an infinite ray that begins at the center of D in the image plane and misses every image vertex under the map g, and passes onward to infinity. Then R either misses an image triangle $g(T)$ entirely or else intersects $g(T)$ precisely in an arc that enters $g(T)$ through one edge and exits $g(T)$ through another edge. See Figure 19.

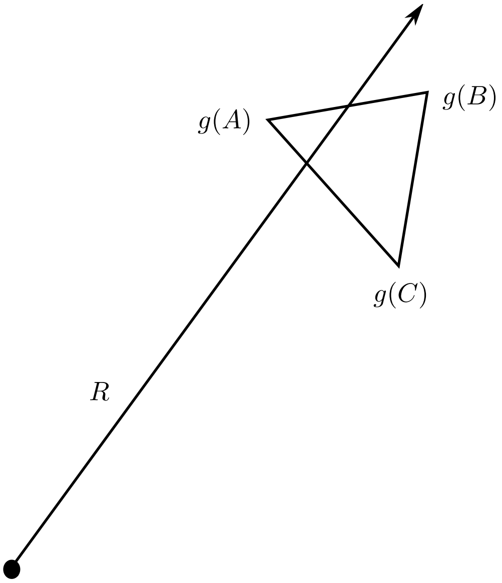


FIGURE 19. The intersection of a ray with a triangle

The intersection arc can be carried back to an intersection arc in the model triangle $T = ABC$ by the linear correspondence between T and $g(T)$.

As in the previous proof, we concentrate on the union of the intersection arcs as viewed in the model disk. As before, this union must be a finite collection of finite edge paths, each having two endpoints, and a finite collection of polygonal simple closed curves. As before, we consider that component that contains the only intersection point of R with ∂D. That component must be an arc with two endpoints by Idea (3). The second endpoint can only lie on the boundary of D, and there is no other intersection with ∂D, a contradiction. That is, this component is a one-ended arc.

As before, we conclude that the theorem is true, that our original assumption was truly ridiculous. □

PROOF OF THE BROUWER FIXED POINT THEOREM. Since the round disk E and the square disk D are topologically equivalent, we may replace D by E in both the No Retraction Theorem and the Brouwer Fixed Point Theorem. We assume that, contrary to the latter, there is a continuous function $g : E \to E$ such that, for each $x \in E$, $g(x) \neq x$. We then define a function $f : E \to \partial E$ as follows.

Consider the ray $R(x)$ that begins at $g(x)$ in E and passes through x on its way to ∞. Define $f(x)$ to be the last point of E in the ray $R(x)$. Then the function $f : E \to \partial E$ is a continuous function that fixes each point of ∂E, in contradiction to the No Retraction Theorem. □

2.5. Exercises

2.1. Solve Exercise 2.3 on page 11.

2.2. Solve Exercise 2.14 on page 24.

2.3. Solve Exercise 2.15 on page 24.

2.4. Generalize the No Retraction Theorem to dimension 3. Consider the unit cube. Show how to divide the unit cube into tiny tetrahedra. View these tetrahedra as lying in the model cube. Show how to carry those tetrahedra to tiny tetrahedra in the image cube. Determine what the appropriate general position property is for the intersection of a ray with a tetrahedron. Pull the intersection segments back into the model cube. Show that you obtain a one-ended arc.

2.5. Contemplate what would have to be done to prove a No Retraction Theorem in every dimension. Think how you would deduce a generalized Brouwer Fixed Point Theorem in that dimension.

The Brouwer Fixed Point Theorem does not suggest a method for finding a fixed point. If the map $f : \mathbb{B}^2 \to \mathbb{B}^2$ is a *contraction mapping*, a particularly nice property that we will now explain, then it is an easy matter to approximate a fixed point. (Contraction mappings defined on other suitable spaces are one of the main tools used to find solutions to differential equations.)

Definition 2.18. A map $f : \mathbb{B}^2 \to \mathbb{B}^2$ is a *contraction mapping* if there is a number $0 < \lambda < 1$ such that, for each two points $x, y \in \mathbb{B}^2$, the distances $d(f(x), f(y))$ from $f(x)$ to $f(y)$ and $d(x, y)$ from x to y satisfy the inequality

$$d(f(x), f(y)) \leq \lambda \cdot d(x, y).$$

In the following exercise you will need to use one of the fundamental properties of a compact metric space: Every Cauchy sequence converges. (Recall that a sequence $x_1, x_2, \ldots$ is a Cauchy sequence if, for each $\epsilon > 0$, there is an integer N such that $n, m \geq N$ implies $d(x_n, x_m) < \epsilon$.)

2.6. Prove that, if $f : \mathbb{B}^2 \to \mathbb{B}^2$ is a contraction mapping, then f has a unique fixed point x_0. If $x_1 \in \mathbb{B}^2$ and if $x_{n+1} = f(x_n)$ for each $n \geq 1$, then the sequence $x_1, x_2, \ldots$ is a Cauchy sequence converging to x_0.

2.7. Show that a tile position can be attained iff the parity of the permutation agrees with the parity dictated by the position of tile 16.

2.8. Consider puzzles like the 15 puzzle, but with dimensions $m \times n$. Which configurations can be realized by such a puzzle?

2.9. Suppose that J is a polygonal simple closed curve in the plane bounding a disk D. Show that D can be divided into triangles using only the vertices of J as vertices of the triangles used.

2.10. By induction on the number of triangles used in the previous exercise, show that D is homeomorphic to a single triangle T. (That is, you must prove the existence of a continuous bijection $f : D \to T$.)