Society of Mathematics
Proceedings of the Colloquium

Boston University

Edited by Zachariah Zobair

Spring 2025

Editor's Note

This volume is a collection of the notes prepared alongside talks given for the Boston University Society of Mathematics Colloquium. The Society of Mathematics is a student group which prides itself on being a space for all those who are interested in math to come together and discuss, learn, and collaborate. To this end, we officially established the colloquium this Spring 2025 semester as a way for students who are excited about what they are learning to invite others to experience some of that excitement. Since speakers were given a choice of topic, the reader will see that this text spans a wide swath of mathematics, ranging from algebraic topology to number theory to geometry. The double-edged nature of this, however, is that it was nigh impossible to mandate any sort of consistency in notations and conventions used. Thus it is perhaps more beneficial to view each chapter as a standalone work.

Of course, this would not be possible to create without the work of the speakers themselves. All of the writings were prepared by the speakers and so this really is their production. I extend my deep gratitude to each contributor for their efforts both in ensuring we had interesting mathematics to listen to and learn during the semester, as well as for allowing for their work to be collected here. I also must thank Professor Margaret Beck, for the idea of creating these proceedings is due to her.

As one last remark, and really it is more bookkeeping than anything else, we note that the last two chapters, "Root Systems and Lie Groups" and "Elliptic Curve Cryptography", are both from talks given in the Fall 2024 semester, before the Society of Mathematics Colloquium officially began. We felt that for the sake of completeness they ought to be included here.

- *Zach Zobair, May 2025*

Contents

1. An Invitation to Algebraic Geometry Through Conics

—*Skyler Marks*—

Abstract

In high school algebra 1 and 2, we study the theory of single variable polynomial equations. In linear algebra, we study systems of linear equations, or polynomial equations with no exponents greater than 1. Algebraic Geometry combines these disciplines to study polynomial equations (in particular, their solutions) in many variables. This theory is useful as it is specific enough that we can compute with it, yet general enough that it applies to many problems we care about. The promised invitation will be extended by way of plane conics and cubics. After some definitions and preliminaries, we will review a family of classical results regarding plane conics (quadratic polynomials in two variables). We will begin by classifying conics into families who's members are "alike". We will then leverage this classification to study the intersections of two conics. Our conclusion to this first act will be a detailed discussion of the number of conics passing through n points in the plane, introducing Moduli spaces and projective space.

WARNING: These notes may contain errors; if you find any, please email me at `skyler@bu.edu`.

NOTE: These notes are designed for a general audience, but include remarks directed towards those with a background in algebraic geometry.

1.1 Preliminaries

We'll be working primarily with polynomials in two variables with complex coordinates:

Definition 1.1.1. The **ring of polynomials in two variables**, denoted $\mathbb{C}[x, y]$ is the set of all polynomials in two indeterminates together with the standard operations.

Remark 1.1.2. As far as I know, much of what follows generalizes to $k[x, y]$ where $k = \bar{k}$, although some things may fail in e.g. characteristic 2. No guarantees on any of these notes, but even less if you're working in positive characteristic.

Example 1.1.3. The polynomials $x^2y + 1$, $x^2 + y$, $y^2 + x^2$, and $x^2 + x$ are all elements of $C[x, y]$. Recall that:

$$(x^2 + y) + (y^2 + x^2) = 2x^2 + y^2 + y$$

$$(x^2 + y)(y^2 + x^2) = x^2y^2 + x^4 + y^3 + x^2y$$

Definition 1.1.4. Polynomials with no power higher than 1, like $x - 1$ and $y - 2$, are called **linear**.

This notion motivates the following definition:

Definition 1.1.5. The **degree** of a single term is the sum of the powers in that term. The degree of a polynomial is the highest degree of any monic term.

Definition 1.1.6. An **ideal** in the ring of polynomials is a subset of the set $\mathbb{C}[x, y]$ which is closed under addition and multiplication by any element in $\mathbb{C}[x, y]$.

Definition 1.1.7. A **morphism of rings** or **ring homomorphism** is a function f between two rings satisfying $f(a + b) = f(a) + f(b)$ and $f(ab) = f(a)f(b)$. An **isomorphism** of rings is an invertible morphism.

Definition 1.1.8. We'll call the set of all pairs (z_1, z_2) for $z_1, z_2 \in \mathbb{C}$ the **affine plane over** $\mathbb{C}$ and denote it $\mathbb{A}^2$.

Lemma 1.1.9. *Pick some subset S of affine space. The set $I(S)$ of all polynomials which vanish at S is an ideal.*

Proof. Exercise. □

Definition 1.1.10. We define the **vanishing set** of a polynomial $f(x, y)$, denoted $V(f)$, to be the set where the polynomial is zero:

$$V(f) = \{(x, y) \in A^2 | f(x, y) = 0\}$$

We broadly wish to study, in this section, the vanishing set of polynomials of degree two; well call these conics:

Definition 1.1.11. A **conic** is the zero set of a polynomial of degree 2.

Question 1.1.12. We can state our main goals for this section formally as follows:

1. What can the set $V(f)$ "look like" for f a degree 2 polynomial?

2. What does the set $V(f) \cap V(g)$ "look like" for f, g degree two polynomials? In particular, how many points does this set contain?

3. What are the elements of $I(\{p_1, ..., p_n\})$ for $p_i = (x_i, y_i)$ a point in $\mathbb{A}^2$, and for all $n \in \mathbb{N}$? (how many quadratics pass through n points).

These three questions exemplify three of the main fields of algebraic geometry: namely, birational geometry / the classification problem, intersection theory, and moduli theory. The first field seeks to classify all *algebraic varieties* (things which look locally like affine algebraic sets, more or less) up to some form of isomorphism; the second seeks to study how often and in what ways two such algebraic varieties intersect; the third seeks to describe families of mathematical objects using geometric objects.

The final concept that will allow us to answer these questions is a notion of isomorphism between vanishing sets of polynomials; that is, a way of determining when two vanishing sets are "essentially the same". This will allow us to formalize and answer Question 1.1.12, points 1 and 2. In order to do this, we need a little more algebra:

Definition 1.1.13. Let R be a ring (in particular, the ring $\mathbb{C}[x, y]$ of polynomials), and let I be an ideal. The **quotient** R/I is then the set of cosets of the form $f + I$ for an element f.

Lemma 1.1.14. *The quotient R/I is a ring under the operations $(p + I) + (q + I) = (p + q) + I$ and $(p + I)(q + I) = (pq) + I$.*

Proof. C.F. [DF08] □

Remark 1.1.15. One can consider the quotient by the ideal generated by an element f to be "evaluation at $f = 0$", or the quotient by an ideal to be the identification of everything in that ideal with zero. Indeed, we see this information captured formally in:

Theorem 1.1.16 (The First Isomorphism Theorem)**.** *Let $f : R \to S$ be a morphism of rings; that is, a map satisfying $f(a+b) = f(a) + f(b)$ and $f(ab) = f(a)f(b)$. Then* $\mathrm{Ker}(f) = \{x \in R | f(x) = 0\}$ *is an ideal, and $(f) \cong R/\mathrm{Ker}(f)$. In particular, if f is surjective, $S \cong R/\mathrm{Ker}(f)$. (Here $\cong$ denotes isomorphism).*

Proof. C.F. [DF08] □

Definition 1.1.17. We define the **affine coordinate ring** of a subset S of A^2 to be $k[x,y]/I(S)$. We say that two (algebraic) subsets of A^2 are isomorphic ("look the same") if their affine coordinate rings are isomorphic.

Definition 1.1.18. A **algebraic set** is the zero set of a polynomial in affine space. Thus the above definition associates to each algebraic set it's affine coordinate ring.

Remark 1.1.19 (Affine Schemes - Should probably be ignored)**.** This association is extremely important in algebraic geometry. Those who are well-versed in abstract algebra could benefit from convincing themselves that (1) there is a bijection between points in A^2 and maximal ideals in $k[x,y]$ (this is simple; $(a,b) \mapsto (x-a, y-b)$), (2) that this bijection also descends to algebraic subsets of A^2 (think carefully about what it means for a maximal ideal in $k[x,y]$ to remain a maximal ideal in $k[x,y]/(f)$), and (3) that this argument extends to show that there is an inclusion-reversing bijection between points in algebraic sets in A^2 and prime ideals in their coordinate rings. If one is excited by this line of reasoning, one might continue to show that if g is a function in $k[x,y]$, we have that $g/(x-a, y-b) = g(a,b)$. This discussion motivates the construction of **affine schemes**: instead of our geometric object being a set of tuples of complex numbers, we take it to be a set of prime ideals in a ring; the zero set of a function (any element of the ring) f becomes the set of all prime ideals which include f. This captures all of the geometric intuition we develop here, but is far more powerful and general.

1.2 Conics up to Isomorphism

In math, a huge portion of our work is classifying objects up to some sort of isomorphism - in our case, isomorphism of affine varieties. Such a question is a good answer to Question 1.1.12, part 1; it turns out we can classify plane conics in a very satisfying way. Furthermore, we'll see in the next section that this classification is extremely useful (in particular, for proving Theorem 1.3.4) as well as being interesting in and of itself.

Theorem 1.2.1. *Suppose* $\phi = ly^2 + axy + bx^2 + cx + dy + e$ *defines a conic* $Z(\phi)$*. Then if* $4b - a^2 = 0$*,* $A(Z(\phi)) \cong C[x, y]/(y^2 - x)$*; otherwise,* $A(Z(\phi)) \cong C[x, y]/(xy - 1)$*.*

Proof. First consider the case where $l = b = 0$. Then $\phi = axy + cx + dy + e$. Then add and subtract $\frac{cd}{a}$ ($a \neq 0$, as otherwise $Z(\phi)$ would be a degenerate conic) to obtain $\phi = axy + cx + dy + \frac{cd}{a} - \frac{cd}{a} + e$. Factor to obtain $\phi = (ax + d)(y + \frac{c}{a}) - \frac{cd}{a} + e$. Consider the transformation $\psi : Z(\phi) \to Z(xy + \tilde{e})$ by the rule $(x, y) \mapsto \left(\frac{x-d}{a}, y - \frac{c}{a}\right)$. Clearly this is a polynomial function with a polynomial inverse, and thus a regular isomorphism. Thus

$$Z(\phi) \cong Z(xy + \tilde{e})$$

Moreover, another transformation $f : Z(xy + \tilde{e}) \to Z(\tilde{e}xy + \tilde{e})$ can be constructed, by the rule $(x, y) \mapsto (\tilde{e}x, y)$ (which is well defined as $k = \bar{k}$). Thus, as multiplication by a scalar does not change the zeros of a polynomial, $Z(\phi) \cong Z(xy + 1)$.

However, if l or b is nonzero, (suppose l without loss of generality) then we can divide to obtain a monic polynomial for new coefficients $a, b, ...$

$$y^2 + ayx + bx^2 + dy + cx + e$$

Add and subtract $\frac{(ax)^2}{4}$:

$$y^2 + ayx + \frac{(ax)^2}{4} - \frac{(ax)^2}{4} + bx^2 + dy + cx + e$$

Factor

$$(y + \frac{ax}{2})^2 - \frac{(ax)^2}{4} + bx^2 + dy + cx + e$$

Use the transformation $(x, y) \mapsto \left(x, y - \frac{ax}{2}\right)$. Note this is an isomorphism with inverse $(x, y) \mapsto (x, y + \frac{ax}{2})$ and image

$$Z\left(y^2 - \frac{(ax)^2}{4} + bx^2 + d\left(y - \frac{ax}{2}\right) + cx + e\right)$$

Which simplifies to

$$Z(y^2 + b'x^2 + c'x + d'y + e')$$

If $b' = 0$, then if $c' = 0$ the polynomial is a quadratic in y and splits, so c' or b' are nonzero. First suppose c' is nonzero:

$$= Z(y^2 + c'x + d'y + e')$$

Add and subtract $\frac{d'^2}{4}$ and factor to obtain

$$= Z((y + \frac{d'}{2})^2 - \frac{d'^2}{4} + c'x + e')$$

Then another affine transformation $(x, y) \mapsto \left(y - \frac{d'}{2}, \frac{-x + \frac{d'^2}{4} - e'}{c'}\right)$ yields the intended result

$$\cong Z(y^2 - x)$$

Suppose, then, that $b' \neq 0$.

$$Z(y^2 + b'x^2 + c'x + d'y + e')$$

Then add and subtract $\left(\frac{c'}{2b'}\right)^2$ and factor:

$$Z\left(y^2 + \left(\sqrt{b'}x + \frac{c'}{2b'}\right)^2 - \left(\frac{c'}{2b'}\right)^2 + d'y + e'\right)$$

Another coordinate transform and re-labling coefficients gives

$$\cong Z\left(y^2 + x^2 + d'y + \tilde{e}\right)$$

Complete the square and transform again to obtain

$$\cong Z\left(y^2 + x^2 + \rho\right)$$

$$\cong Z\left((x + iy)(x - iy) + \rho\right)$$

. Transform $(x, y) \mapsto (x + iy, y)$

$$\cong Z\left((x + 2iy)(x) + \rho\right)$$

Transform $(x, y) \mapsto (x, \frac{-i}{2}(y - x))$.

$$\cong Z\left(xy + \rho\right)$$

Transform $(x, y) \mapsto (\rho x, y)$, and note that scalar multiplication does not change the zeros of a polynomial:

$$\cong Z\left(xy + 1\right)$$

□

This solves Question 1.1.12, part 1.

1.3 Intersections of Conics

We now address our second question, namely how many points two conics meet in. This is a classical problem; the intersections of lines and the simultaneous vanishing of higher-order polynomial functions has occupied much of the study of both algebra and geometry for quite some time. The notion that the algebraic viewpoint and the geometric viewpoint are linked, and can be used in concert, has been extremely fruitful and lead to modern-day *intersection theory.*

Lemma 1.3.1. *If T_i for each natural i are subsets of $\mathbb{C}[x, y]$, then:*

$$V(T_1) \cup V(T_2) = V(T_1T_2)$$

Where T_1T_2 is the ideal generated by all products of elements in T_1 and T_2, and

$$\bigcap_i V(T_i) = V\left(\bigcup_i T_i\right)$$

Remark 1.3.2. Those who are familiar with the topic will recognize that the above are the closure conditions necessary to specify the closed sets of a topology; indeed, the topology who's closed sets are the vanishing sets of polynomials is called the Zariski topology, and is the (main) topology used in algebraic geometry.

Lemma 1.3.3.

$$C[x, y]/(f, g) \cong (C[x, y]/(f))/(g)$$

Proof. Consider the map $\psi : C[x, y] \to (C[x, y]/(f))/(g)$ by the rule $a \mapsto (a + (f)) + (g)$. Clearly this is surjective; any element in $(C[x, y]/(f))/(g)$ is of the form $(a+(f))+(g)$. Moreover, the kernel of this map is exactly (f, g); $a \in (f, g)$ if and only if $a = xf + yg$;[1] f is killed by the first quotient and g is killed by the second, so a maps to zero if and only if it is of this form. Then we are done by Theorem 1.1.16. □

Theorem 1.3.4. *Let f and g be degree 2 polynomials in $\mathbb{C}[x, y]$ with no common factor.*[2] *Then $V(f) \cap V(g)$ contains at most four points.*

[1] This fact is due to the fact that the ideal generated by a set A is equal to RA; see [DF08], page 251.

[2] For more information on why this condition makes sense, look into the theory of Unique Factorization Domains - a broad class of rings, of which $\mathbb{C}[x, y]$ is an element.

Proof. By lemma 1.3.1, we study $S = V((f) \cup (g))$ where f, g are polynomials of degree 2 with no common factor. This will be the same as $V((f,g))$; consider the affine coordinate ring $C[x,y]/(f,g)$. By Lemma 1.3.3, this is $(C[x,y]/(f))/(g)$. We know that $C[x,y]/(f)$ is isomorphic to either $C[x,y]/(xy+1)$ or $C[x,y]/(y^2+x)$. Consider first the second case. Note that in this quotient we can replace each instance of x with one of y^2, thereby obtaining a polynomial in y; because there is a unique way to do this, we obtain unique representatives for each element in the quotient. Every polynomial in y can be obtained this way, yielding a well-defined homomorphism to $C[y]$. We consider the image of g under the map "substitute y^2 for x" - this will be, in general, a degree 4 or lower polynomial (as g may have an x^2 term, which maps to y^4) in one variable over C; as C is algebraically closed, this polynomial has at most four roots and splits into at most 4 linear factors. Then these linear factors are the maximal ideals which contain the image of (g), and thus correspond bijectively to the maximal ideals in our final quotient $C[x,y]/(f,g)$, and thus to the points in the associated affine variety.

We now consider the second case; we can view the quotient $C[x,y]/(xy+1)$ as $C[x,x^{-1}]$ under the map $y \mapsto -x^{-1}$. Then we note that g maps to a polynomial of the form $P = ax^2 + bx^{-2} + cx + dx^{-1} + e$. But now that x is invertible, the ideal generated by this polynomial is the same as the ideal generated by x^2P; this is because we can multiply x^2P by x^{-2} to get P, and vice versa, so any ideal which contains one must contain the other. But x^2P is a quadratic in x, and thus splits into at most 4 linear factors; a symmetric argument to the first case then shows that there are at most 4 points in the algebraic set associated to the coordinate ring. □

We can see that in the setting we've outlined, the maximum and minimum (zero intersections) are both attained.

Example 1.3.5. The polynomials $\frac{x^2}{3} + \frac{y^2}{1} - 1$ and $\frac{x^2}{1} + \frac{y^2}{3} - 1$ intersect in four points.

Example 1.3.6. The polynomials $yx - 1$ and $xy - 2$ do not intersect:

$$xy - 2 = xy - 1$$

$$2 = 1$$

Exercise 1.3.7. Find non-degenerate pairs of conics (i.e. Conics which have a nonzero order-two term and are not the product of two linear terms) which intersect exactly one, exactly two, and exactly three times, or prove that no such pair exists.

Remark 1.3.8. This result is somewhat unsatisfying as it doesn't really tell us how many intersections any given pair of polynomials have. We can remedy this (somewhat) by some constructions which, although beyond the scope of this talk, allow us to "fix" cases like the above so that the answer to our question "how many times do two conics intersect" is a decisive "four".

1.4 Conics Meeting n Points and Moduli Spaces

Now we address the final aspect of our question; namely, how many conics (and which) pass through n points. This may seem like the least natural question to ask; one could motivate it by saying that we wish to see how much "control" we have over a conic; in essence, how many "degrees of freedom" a conic has, or how many points are necessary to specify a conic. Indeed, we will see that the best way to answer this question is to formalize the question "how does one specify a conic lying in the plane"; the answer to that question (a so-called "moduli space") also provides the answer to Question 1.1.12, part 3. We begin with the case $n = 1$:

Theorem 1.4.1. *There are infiniely many degree* 2 *polynomials passing through a point.*

Proof. Let P be a point in the affine plane. We can represent this point as $P = V((x-a, y-b))$ for $P = (a, b)$; the set of points where the functions $x-a$ and $y-b$ both vanish is exactly P (recall also our correspondence between points and maximal ideals - this ideal is maximal). We wish to find degree two polynomials in the maximal ideal $(x-a, y-b)$; but all such polynomials are of the form $f(x-a) + g(y-b)$ where f and g are degree ≤ 1, and we have found all (infinitely many) degree two polynomials through the given point. □

Definition 1.4.2. **Affine** n**-space** over the complex (or real) numbers is the set of all n-tuples $(x_1, ..., x_n)$, for x_i in the complex (or real) numbers.

Definition 1.4.3. **Projective** n**-space** is the set of lines through the origin in affine $n+1$ space. We can view this as all points which are a scalar multiple of a nonzero point in affine $n+1$ space; as such, define projective space to be the set of sets $[x_0 : ... : x_n] = \{(\lambda x_0, ..., \lambda x_n) | \lambda \in \mathbb{C}\}$ (note the colons, square brackets, and zero indexing).

Remark 1.4.4. The numbering conventions exist for dimensional reasons that we won't touch on in this talk. Effectively, around any point, projective n space "looks like" affine n space.

Definition 1.4.5 (Loose Definition)**.** A **moduli space** is a "space[3]" in which each point corresponds to a object which you wish to study, and for which nearby points usually correspond to similar objects.

Example 1.4.6. Consider the set of all lines in $\mathbb{R}^2$. A (non-vertical) line is given uniquely by the equation

$$y = mx + b$$

meaning that we can parametrize all non-vertical lines by $\mathbb{R}^2$, where a point (m, b) in $\mathbb{R}^2$ corresponds to the line with the above equation. If we want to consider vertical lines, a line can be given by a line through the origin which is "slid" somewhere else; we have a good representation of the set of all lines through the origin in 2-space, we can consider the space $\mathbb{RP} \times \mathbb{R}$ - the set of pairs (m, b) where m is a point in projective 1-space (aka a set of the form $[x_1 : x_2]$) and b is the y-intercept of the line we want. We can't really rigorously discuss the notion of nearby points yielding similar lines, but intuitively, the idea is there - varying x_1, x_2, or b only slightly gives a line which isn't much different.

Theorem 1.4.7. *The moduli space of complex conics in the plane is P^5. Moreover, 5 points determine a conic; more general points will not lie in a single conic, and less will lie in infinitely many.*

Proof. Consider a general conic $P = ax^2+bxy+cy^2+dx+ey+f$. The scaling λP by any complex number λ yields a polynomial with the same roots, so we identify P with λP. Identifying the conic with the tuple of it's coordinates then gives a moduli space; the space of conics is P^5. Suppose we require $P(x_1, y_1) = 0$. Then

$$ax_1^2 + bx_1y_1 + cy_1^2 + dx_1 + ey_1 + f = 0$$

This defines a hyperplane through the origin in affine 6-space; indeed, each point we require the zero set of P to contain defines another hyperplane through the origin. Then, if we require the zero set of P contain all n distinct points, the coefficients must be in the intersection of n hyperplanes through the origin. Assuming no two points define the same hyperplane (which is true in general - changing the point slightly means it will define a different hyperplane), some linear algebra gives that the intersection of n

[3] A topological space usually, often with some extra structure - but we won't go into the details of topology.

distinct hyperplanes in affine 6-space is $6 - n$ dimensional. In particular, if $n = 5$, then we are left with a line; when we consider the space of all lines in this space, it is a single point. Thus we have recovered the classical result that 5 points determine a conic. If $n < 5$, we have an infinite number of lines in our space; if $n > 5$, we have none. □

Example 1.4.8. Find the conic polynomial which passes through the points $(0, 0)$, $(0, 1)$, $(1, 2)$, $(-1, 0)$ and $(0, -1)$.

Proof. Begin with $(0, 0)$. This imposes the condition $f = 0$. Next we consider the point $(0, 1)$. This imposes the condition:

$$c + e + f = 0$$

The point $(1, 1)$ imposes the condition:

$$a + 2b + 4c + d + 2e + f = 0$$

The point $(-1, 0)$ imposes the condition:

$$a - d + f = 0$$

The point $(0, -1)$ imposes the condition:

$$c - e + f = 0$$

Some linear algebra yields:

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 2 & 4 & 1 & 2 & 1 & 0 \\ 1 & -1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & -1 & 1 & 0 \end{bmatrix}$$

Row reducing:

$$\begin{bmatrix} 1 & 0 & 0 & \frac{1}{3} & 0 & 0 & 0 \\ 0 & 1 & 0 & \frac{1}{3} & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

We obtain our answer in terms of a free variable, d. That is, our solution is $c = e = f = 0$ $3d = a = b$. Identifying all scalar multiples allows us to assume $d = 1$, so we can write our polynomial:

$$P = 3x^2 + 3xy + x$$

This is then the (unique) degree 2 polynomial which is zero at each of these points! □

Exercise 1.4.9. Check that the above polynomial is indeed zero at each of these points.

Exercise 1.4.10. Pick 5 points and find a polynomial passing through those. Try to pick your points such that the resulting polynomial doesn't factor.

Remark 1.4.11. It's interesting to think about how difficult it is to complete the above exercise. In theory, there is a 100% probability that you'll pick at random such sets of five points which don't define a reducible polynomial (this fact relies on some advanced statistics that's beyond the scope of this lecture, but I still find it interesting). Why, if this is the case, is it so easy to find sets of points for which the polynomial *is* reducible?

Bibliography

[DF08] David S Dummit and Richard M Foote. *Abstract Algebra, 2Nd Ed.* 7 2008.

[Gro74] Alexandre Grothendieck. *Eléments de Géométrie Algébrique*. 1974.

[Har11] Robin Hartshorne. *Algebraic geometry*. Springer, New York ; London, 2011.

2. CW-Complexes and Cellular Approximation

—*Alice Marchant*—

Abstract

Imagine building a tower. You have to start with some sort of foundation and then build a layer up and then another layer on top of that and so on. A CW complex is a way that we can do this same building process but for (almost) any space you can think of by just adding n-disks to a previous part. We will explore in more detail how this construction works, some of the constraints on it, and why it is such a useful construction in algebraic topology.

2.1 Introduction

These notes are organized around providing the background and proof for the Cellular Approximation Theorem in algebraic topology. The facts from point-set topology can be found in any introductory topology textbook and the subsequent sections involving CW complexes and the applications of the theorem will be pretty closely adapted from Hatcher's Algebraic Topology. We will organize these notes as follows: we will begin with a quick sketch of the idea of the proof, then we will go through a quick overview of useful theorems and definitions in point-set topology, then we will define some constructions needed to prove the theorem (including CW complexes), then we will go over some applications. These notes should be fairly self-contained, only assuming some knowledge of calculus, linear algebra, and basic set theory. Except for the applications section, which requires some group theory and additional topological knowledge.

2.2 Motivation

Consider a loop on a two sphere perhaps like the one in the figure:

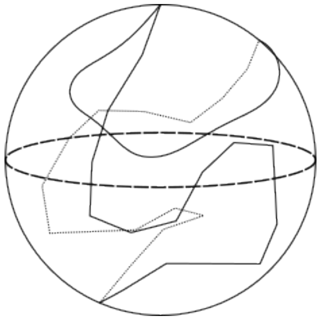


Figure 2.1: A loop on a sphere

Note that the loop is not particularly smooth, and it self-intersects. This is because the main question we wish to answer is what happens to the loop as we allow continuous deformations, and we want to consider this question for any loop on a sphere. At the start, it may be useful to consider the loop as a rubber band wrapped around an ordinary sphere. From here it may be clear that while keeping the rubber band on the sphere, we can move it around to be as tightly compressed as physics would allow. Assuming now that our rubber band is infinitely thin, we would notice that the rubber band can be continuously deformed to a point. This appears to be an obvious fact, but when we visualize a rubber band as our loop on the sphere, we are implicitly assuming that the loop has some amount of smoothness. This may not always be the case. For example, there is a way we can put a closed interval into a filled in square by the following construction called a *space-filling curve*. A reader can see [Mun14, p. 272] for details about this construction. The upshot of this is that we can a priori have a loop on the sphere that does actually covers every point of the sphere! In this case, one might worry that we cannot actually slide this curve around on the sphere so that we can compress it to a point. To see why this is indeed possible, we must look closer at what it is that allows us to deform the rubber band to a point in the nice case. But this will require us to consider the sphere in a different way. Namely, we can consider a sphere as obtained from a point by gluing the boundary of a disk to it as in the figure:

Returning to our physical rubber band, if the rubber band misses one point of the disk, we can push the rubber band outward from the point that it misses so that it is only on the boundary of the disk, which is the same as

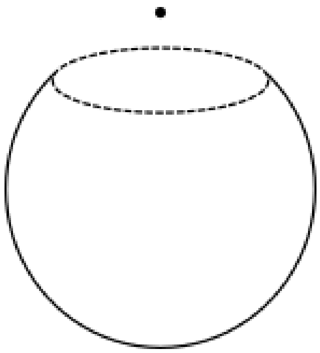


Figure 2.2: Gluing the boundary of a disk to a point

the rubber band being compressed to a point because we glued the boundary of the disk to the point. The reader would note that this method would still fail if we have a space-filling curve, but there is a result that we can actually deform any loop to miss one point! So this method still works. What we have roughly shown in more formal terms is that any continuous function $f : [0, 1] \to S^2$, with S^2 being the sphere in question such that $f(0) = f(1)$ can be deformed to a point. An interesting fact about this is that it also generalizes to general spheres S^n as the space of unit vectors in $\mathbb{R}^{n+1}$ in that *any* loop on S^n can be deformed to a point! This can also be nontrivially generalized to the following theorem:

Theorem 2.2.1 ([Hat01, p. 349])**.** *Every map $f : X \to Y$ of CW complexes is homotopic to a cellular map. If f is already cellular on a subcomplex $A \subseteq X$, the homotopy can be taken to be stationary on A.*

A priori it may not be clear why this is a generalization of what we have shown. But as we will see, a CW complex is really a collection of spheres glued together in some way. So it will be good to keep this example in mind while reading the proof of the theorem to notice the parallels.

2.3 A Quick Introduction to Topology

2.3.1 Topology

First to do topology, we need to know what a topology is. To motivate this, we will consider topology as a way to generalize limits in calculus. Recall that informally $\lim_{x \to x_0} f(x)$ is what the values of $f(x)$ approach as x gets arbitrarily close to x_0. We now want to formalize what it means to get "arbitrarily" close to a point. However, rather than considering functional limits, we will focus on sequential limits. Consider the sequence $\frac{1}{2^n}$. As n goes to infinity we say that this sequence converges to 0. One way we could make this precise is by saying that for any distance away from 0, there is an element of the sequence that is closer that said distance, or that if s_n is our sequence, for any $\varepsilon > 0$, there is an n such that $s_n \in (-\varepsilon, \varepsilon)$. We actually want something stronger than this to formalize convergence. The main reason for this is that a priori there is nothing stopping a sequence from for instance getting to our limit and then continuing on away from the limit. For example, under our current definition, the sequence $(1, 2, 3, 0, 3, 4, 5, 6, \dots)$ will converge to 0. So we actually want to require that there is an n such that anything greater than that n, will also be in this neighborhood $(\varepsilon, \varepsilon)$ for our sequence to converge to 0. But also note that because $(-\varepsilon, \varepsilon)$ is an open set, we can again rephrase the definition to be that our sequence converges to 0 if for any open set containing 0, there is a point after which the sequence is in our open set. Through this we can see open sets in $\mathbb{R}$ as capturing what it means to be near a point. And topology in general deals with generalizations of this to consider this nearness, as well as continuity and convergence for an arbitrary set.

Definition 2.3.1 ([Mun14, p. 76])**.** A *topology* on a set X is a collection $\mathcal{T}$ of subsets of X having the following properties:

1. $\emptyset$ and X are in $\mathcal{T}$.

2. The union of the elements of any subcollection of $\mathcal{T}$ is in $\mathcal{T}$.

3. The intersection of the elements of any finite subcollection of $\mathcal{T}$ is in $\mathcal{T}$.

A set X for which a topology $\mathcal{T}$ has been specified is called a *topological space*

Aside from the usual topology on the real line in terms of the distance between points, we also have some more interesting topologies we can put on an arbitrary set X.

Example 2.3.2 (Discrete Topology [Mun14, p. 77])**.** $\mathcal{T} = \mathscr{P}(X)$

Example 2.3.3 (Trivial Topology [Mun14, p. 77])**.** $\mathcal{T} = \{\emptyset, X\}$.

Example 2.3.4 (Finite Complement Topology [Mun14, p. 77])**.** A set $U \subsetneq X$ is open if and only if $X \setminus U$ is finite and we declare that $\emptyset$ is open. Note that $X \setminus X = \emptyset$ is a finite set and $\emptyset$ is open by definition. So we have satisfied (1). For (2), consider:

$$X \setminus \bigcup_{\alpha} U_{\alpha} = \bigcap_{\alpha} (X \setminus U_{\alpha})$$

from DeMorgan's so we have an intersection of finite sets, which is finite. For (3) we have:

$$X \setminus \bigcap_{i=1}^{n} U_i = \bigcup_{i=1}^{n} (X \setminus U_i)$$

which is a finite union of finite sets, and hence finite.

We leave as an exercise the statement that the collection of sets consisting of unions of sets of the form (a, b) is a topology on $\mathbb{R}$.

Definition 2.3.5 (Closed Sets [Mun14, p. 93])**.** A set $C \subseteq X$ of a topological space X is closed if its complement $X \setminus C$ is open.

Definition 2.3.6 (Closure [Mun14, p. 95])**.** Given a subset $Y \subseteq X$ of a topological space X, we define the *closure* of Y to be the set:

$$\text{Cl}(Y) = \bigcap_{C \supseteq Y; C \text{ closed}} C$$

The reader might recall that sets of the form $[a, b] \subseteq \mathbb{R}$ on the real line are considered closed, building off the intuition that they contain their boundary. But this is not quite right topologically because a set of the form $[a, \infty)$ would be considered closed because its complement is $(-\infty, a)$, an open set. But the notion of a set that has a clearly defined boundary does exist in topology. We call this compactness. And we give the topological definition:

Definition 2.3.7 (Compactness [Mun14, p. 164])**.** A subset C of a topological space X is *compact* if given any collection of open sets $\{U_{\alpha}\}_{\alpha \in A}$ such that $\bigcup_{\alpha \in A} U_{\alpha} \supseteq C$, there is a finite subset $A' \subseteq A$ such that $\bigcup_{\alpha \in A'} U_{\alpha} \supseteq C$. We call the collection U_{α} an *open cover.*

It turns out that this definition of compactness is equivalent to the definition that a subset of $\mathbb{R}$ is compact if it is closed and bounded (see [Abb15, p. 96] for details). But this is not going to be true in general. It turns out that this definition is the right one to generalize the nice things about sets of the form $[a, b]$ on $\mathbb{R}$. In particular, we have the following theorems:

Theorem 2.3.8 ([Mun14, p. 165])**.** *Every closed subspace of a compact space is compact.*

Proof. Let $B \subseteq C$ with C compact. Then let $\mathcal{O}$ be an open cover of B. Then $\mathcal{O} \cup \{C \setminus B\}$ is an open cover of C. So we take a finite subcover of this to get a finite subcover of $\mathcal{O}$. □

Theorem 2.3.9 ([Mun14, p. 166])**.** *The image of a compact space under a continuous map is compact.*

The latter theorem turns out to be very powerful, but we need to know what continuity is in a general topological sense before we can make sense of and prove it. To do this, first recall how topology is a way to generalize what it means for points to get arbitrarily close to another point. But this means we can apply this to continuity as well. Recall that informally a continuous function is one that has no sudden jumps or holes. Or another way of putting this is that $\lim_{x \to x_0} f(x) = f(x_0)$. Or that as points in the domain get closer to a target, the points under the function get closer to a corresponding target. In terms of open sets, we could say that we wish that given any open set U containing the target $f(x_0)$, we would like there to be infinitely many points x_n such that $f(x_n) \in U$. But we do not have a sequence of points so we change it to be that there is an open set V containing x_0 in the domain such that $f(V) \subseteq U$. From this it is not hard to conclude that we would like the preimage $f^{-1}(U)$ to be open. So we get the definition:

Definition 2.3.10 ([Mun14, p. 102])**.** Let X and Y be topological spaces. A function $f : X \to Y$ is continuous if for every $U \subseteq Y$ that is open in Y, $f^{-1}(U)$ is open in X. If f is continuous and f^{-1} exists and is continuous. We say that f is a *homeomorphism.* If $f : X \to Y$ is a map such that $f : X \to \operatorname{Im}(f)$ is a homeomorphism, we say that f is an *embedding.*

We will henceforth refer to continuous functions as maps. Using this, we can prove the theorem:

Proof of continuous image of compact spaces. Let $f : X \to Y$ be continuous and let X be compact. Let $\mathcal{O}$ be an open cover of $f(X)$. Then the collection:

$$\{f^{-1}(U) | U \in \mathcal{O}\}$$

is an open cover of X. Hence we may take a finite subcover $f^{-1}(U_i)$. But then U_i is a finite subcover for $f(X)$. □

Defining continuous functions also allows us to build several new topologies from existing ones:

Definition 2.3.11 (Subspace [Mun14, p. 88])**.** Let X be a topological space with topology $\mathcal{T}$. If Y is a subset of X, the collection:

$$\mathcal{T}_Y = \{Y \cap U | U \in \mathcal{T}\}$$

if a topology on Y, called the *subspace topology*. With this topology, Y is called a *subspace* of X.

Definition 2.3.12 (Product [Mun14, p. 86])**.** Let X and Y be topological spaces. The *product topology* on $X \times Y$ is the topology consisting of unions of sets of the form $U \times V$ for U open in X and V open in Y.

Remark 2.3.13. Consider $\mathbb{R}^2$ with the product topology. Then we could have a union of two rectangles, for example $(0, 1) \times (0, 1) \cup (.5, 1.5) \times (.5, 1.5)$, which looks something like:

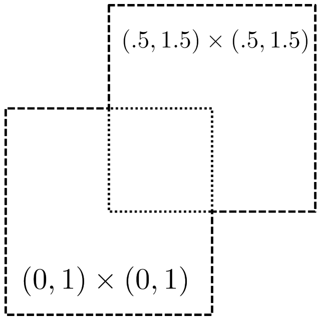


Figure 2.3: Union of rectangles

But this set looks like one that we should consider open and we cannot write it as a product of two open sets in $\mathbb{R}$. Hence the need to consider unions of sets of the form $U \times V$ for the product topology.

Remark 2.3.14 (Alternative Definition of Product Topology)**.** The product topology can also be defined as the topology $\mathcal{T}$ such that if there is another topology $\mathcal{T}'$ such that the projections $\pi_X : X \times Y \to X$ and $\pi_Y : X \times Y \to Y$ are continuous, then $\mathcal{T} \subseteq \mathcal{T}'$

Exercise 2.3.15 (Disjoint Union)**.** Using the same idea as the product topology in terms of the projections. We can define a topology on the set $X \coprod Y := X \times \{0\} \cup Y \times \{1\}$ as the topology $\mathcal{T}$ such that if $\mathcal{T}'$ is another topology such that the inclusions $i_X : X \hookrightarrow X \coprod Y$ and $i_Y : Y \hookrightarrow X \coprod Y$ are continuous, then $\mathcal{T}' \subseteq \mathcal{T}$. We leave as an exercise to show that this topology is given by $U \subseteq X \coprod Y$ is open if and only if $U \cap X$ is open in X and $U \cap Y$ is open in Y.

Definition 2.3.16 (Quotient [Mun14, p. 138])**.** If X is a space and A is a set and if $p : X \to A$ is a surjective map, then the *quotient topology* on A induced by p is the collection of sets:

$$\{U \subseteq A | p^{-1}(U) \text{ is open in } X\}$$

Example 2.3.17 ([Hat01, p. 2])**.** Given topological spaces X and Y and a map $f : X \to Y$, the *mapping cylinder* M_f is the space $X \times I \coprod Y/(x, 1) \sim f(x)$ formed by taking a "cylinder" of X and then attaching one end of the cylinder to Y via the map f.

Now that we have a notion of what a space is in a topological sense, and what it means for a map to be continuous, we would like to be able to formalize the notion of continuous deformation. Consider the following example:

Example 2.3.18 (Deformation of Punctured Disk to its Boundary Circle)**.** Consider a disk with a point removed. We ask how we could write down a mapping that would move points on the disk to its boundary. Consider the disk as a square with the puncture in the center at the origin. Then we get a map:

$$r(x,y) = \begin{cases} (x/y, 1) & |x/y| \leq 1 \text{ and } x + y \geq 0 \\ (1, y/x) & |y/x| \leq 1 \text{ and } x + y \geq 0 \\ (-x/y, -1) & |x/y| \leq 1 \text{ and } x + y \leq 0 \\ (-1, -y/x) & |y/x| \leq 1 \text{ and } x + y \leq 0 \end{cases}$$

This map can be illustrated as the map taking each point to the point on the boundary that the line through the origin and the point hits. Or the mapping demonstrated in the figure:

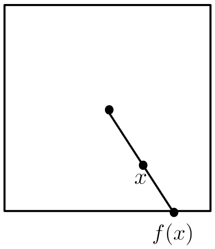


Figure 2.4: Mapping the punctured square to its boundary

And to deform this map to be the identity we would like to parameterize the path that a point takes while moving toward the boundary. This would look something like:

$$r_t(x,y) = \begin{cases} (xy^t/y, y^t) & |x/y| \leq 1 \text{ and } x+y \geq 0 \\ (x^t, yx^t/x) & |y/x| \leq 1 \text{ and } x+y \geq 0 \\ (-xy^t/y, -y^t) & |x/y| \leq 1 \text{ and } x+y \leq 0 \\ (-x^t, -yx^t/x) & |y/x| \leq 1 \text{ and } x+y \leq 0 \end{cases}$$

We call the construction r_t in the previous example a homotopy. Which is defined as follows:

Definition 2.3.19 ([Hat01, p. 3])**.** Let X and Y be topological spaces and $f : X \to Y$ and $g : X \to Y$ maps. A homotopy from f to g is a map $F : X \times I \to Y$ such that $F(x, 0) = f(x)$ and $F(x, 1) = g(x)$ for all $x \in X$ and for each $t \in I$, $F(x, t)$ is continuous as a function of X. If there is a homotopy from f to g we write $f \simeq g$ and say that f is homotopic to g.

Definition 2.3.20 (Retraction [Hat01, p. 3])**.** Given a topological space X with a subspace $A \subseteq X$, a map $r : X \to A$ is a *retraction* of X onto A if $r(X) = A$ and $r|_A = \text{id}$.

Definition 2.3.21 (Deformation Retraction [Hat01, p. 2])**.** A *deformation retraction* of a space X onto a subspace A is a homotopy from the identity map to a retraction of X onto A such that $f_t(x) = x$ for all $t \in I$ and $x \in A$.

Remark 2.3.22 ([Hat01, p. 3])**.** More generally, we call a homotopy such that $F(a, t) = F(a, t')$ for all $t, t' \in I$ and for all $a \in A$ a subspace of the space where the homotopy is defined, a homotopy relative to A. In particular, a deformation retraction is a homotopy from the identity to a retraction $r : X \to A$ a homotopy relative to A.

Definition 2.3.23 (Homotopy Equivalence [Hat01, p. 3])**.** A map $f : X \to Y$ is a homotopy equivalence if there is a map $g : Y \to X$ such that $fg \simeq \text{id}$ and $gf \simeq \text{id}$.

Remark 2.3.24. Any homeomorphism is a homotopy equivalence but not every homotopy equivalence is a homeomorphism. One way to understand the difference is that a homeomorphism is an instantaneous transformation of one space into another that is continuous, where a homotopy equivalence can be considered as allowing one space to morph before transforming it into another space. A priori this appears to be a stronger condition, but using another property called connectedness, we can show that $\mathbb{R}$ is not

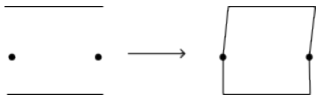


Figure 2.5: CW structure on S^1

homeomorphic to $\mathbb{R}^2$ but the are homotopy equivalent. Indeed, if we let $\pi : \mathbb{R}^2 \to \mathbb{R}$ be projection onto the first coordinate, and $i : \mathbb{R} \to \mathbb{R}^2$ be inclusion into the first coordinate, we get that πi is already the identity and $i\pi(x, y) = (x, 1)$ is homotopic to the identity via the homotopy $F(x, y, t) = (x, y^t)$. This homotopy can be seen as squishing $\mathbb{R}^2$ down to a line.

2.4 The Construction

We can now define our main construction of this talk:

Definition 2.4.1 ([Hat01, p. 5]). A CW complex X is a union of *n-skeleta* X^n such that X^0 is a discrete space and X^n is obtained from X^{n-1} by attaching cells e^n_α via maps $\varphi_\alpha : S^{n-1} \to X^{n-1}$. In other words, X^n is the quotient space $X^{n-1} \coprod_\alpha D^n_\alpha / \sim$ with $x \sim \varphi_\alpha(x)$ for $x \in \partial D^n_\alpha$. An infinite CW complex X is obtained as the union over all n-skeleta $X = \bigcup_n X^n$ and we give X the weak topology: A set $A \subseteq X$ is open if and only if $A \cap X^n$ is open in X^n for all n.

Definition 2.4.2 ([Hat01, p. 7]). Given a CW complex X and let e^n_α be an n-cell. Then there is a map $\Phi_\alpha : D^n \to X$ such that $\Phi|_{\partial D^n} = \varphi_\alpha$ and $\Phi|_{\mathrm{Int}(D^n)} : \mathrm{Int}(D^n) \to e^n$ is a homeomorphism. We call Φ_α the *characteristic map* of e^n.

Example 2.4.3 (Spheres [Hat01, p. 6]). We can now formalize how we were describing a sphere in the introduction. In the case of an n-sphere S^n, we can have multiple structures, but we will only demonstrate two of them. First we will consider the case for $n = 1$ (or a circle). For this, let $X^0 = \{x_0, x_1\}$ a discrete space of two points. It turns out that this is just the boundary of the unit interval D^1 so we can attach two disks D^1_1 and D^1_2 via the identity map $\varphi_i : S^0 \to X^0 = S^0$ as shown in the figure:
and smoothing out the corners we get that this is the same as S^1. When $n = 2$, we have a more traditional hollowed out sphere and we can consider this space as being obtained from S^1 by attaching two disks D^2_+ and D^2_- by the identity map again. Then inductively, we can obtain an n-sphere from an $n - 1$-sphere in a similar way.

Alternatively, we could take $X^0 = \{x_0\}$ to be a single point. Then for S^1, we could get this from attaching the D^1 via the constant map sending the boundary S^0 to x_0. Similarly for S^2 we get the same construction of the sphere that we had in the introduction. And in general we get a CW structure on a sphere S^n consisting of one 0-cell and one n-cell attached via the constant map.

Example 2.4.4 (Torus)**.** By definition, a torus is a product of two circles. So see this, note that a torus is a space taken from a circle being swept in a circular motion. So our torus T is the same as $S^1 \times S^1$. We already have a CW structure on S^1 so we might ask if it is possible to get a CW structure on the product. Using the fact that $D^n \times D^m \cong D^{n+m}$, we note that if Φ_i is a characteristic map for the i-cell in S^1 and Ψ_j is the characteristic map for the j-cell of S^1, we get the products:

$$\begin{aligned}\Phi_0 \times \Psi_0 &: D^0 \to T\\ \Phi_1 \times \Psi_0 &: D^1 \to T\\ \Phi_0 \times \Psi_1 &: D^1 \to T\\ \Phi_1 \times \Psi_1 &: D^2 \to T\end{aligned}$$

give a CW structure on T. In particular, we find that T has one 0-cell, two 1-cells, and one 2-cell, and the attachings for the one cells are:

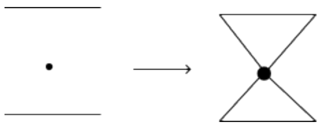


Figure 2.6: Torus 1-skeleton

because there are no interactions between the characteristic maps Φ and Ψ for the 1-cells. And through this we get a figure 8-looking space. But then we consider how the 2-cell is attached. Considering D^2 as a square and letting Φ and Ψ denote their corresponding circles. We get visually:

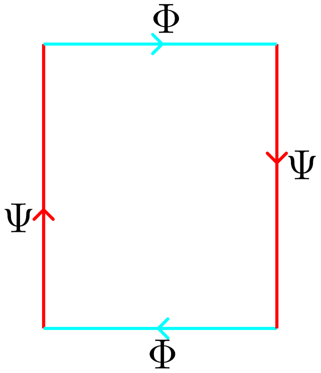


Figure 2.7: Attaching map for the 2-cell on a torus

From the fact that $\Phi_1 \times \Psi_1$ is $(x_0, \Psi_1(y))$ on the red sides of the square and Φ_1 on the blue sides. Using the same color scheme, and following this square from the bottom-left corner, we find that the boundary of D^2 maps onto the figure 8 something like: and if we imagine the red lines gluing to

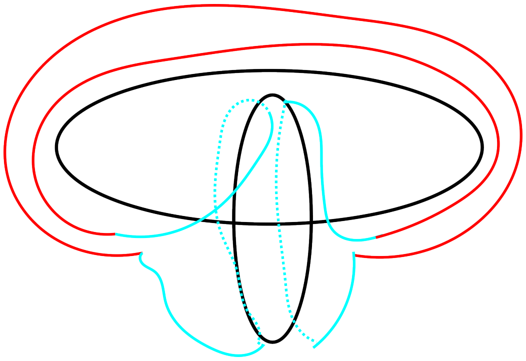


Figure 2.8: Attaching the boundary of D^2 to the 1-skeleton for the torus

the horizontal circle and the blue lines gluing to the vertical circle, as well as the space in the inside of the colored lines filled in, we can see that this will indeed give something that looks like a torus.

Remark 2.4.5 ([Hat01, p. 524])**.** More generally, given two CW complexes X and Y, we get a CW structure on their product by taking products of characteristic maps like we did for the torus in this example.

Warning 2.4.6. In this case, since S^1 is compact, the topology from the CW structure on T will agree with the product topology, but this need not be the case in general. See [Hat01, p. 524] for details.

Example 2.4.7 (Wedge of spheres [Hat01, p. 10])**.** For the torus, we have an example of a *wedge product* when we get the figure 8 for the 1-skeleton. More generally, given spheres S^n and S^m with 0-cells s_0, s_1, we form the wedge product $S^n \vee S^m$ by taking one 0-cell by identifying s_0 with s_1 and attaching D^n and D^m as normal. In low-dimensional cases, we have a figure 8 for $S^1 \vee S^1$ and for $S^1 \vee S^2$ we have:

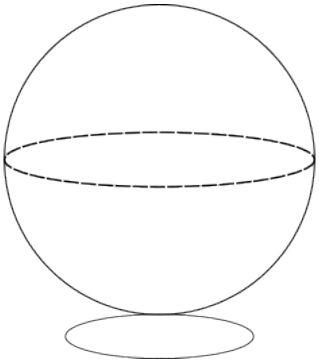


Figure 2.9: $S^1 \vee S^2$

and we leave as an exercise what $S^2 \vee S^2$ looks like.

Example 2.4.8 (Mapping Cylinder of a Cellular Map)**.** Let X and Y be CW complexes and $f : X \to Y$ a map such that $f(X^n) \subseteq Y^n$ for all n. We call such a map *cellular*. Then $X \times I$ has a product CW structure. To get a CW structure on M_f, first let $Z \coloneqq X \times I \coprod Y$. Then let $\Phi : D^n \to Z$ be a characteristic map. If $p : Z \to M_f$ is the quotient, we would like $p\Phi$ to be a characteristic map for M_f. The only thing we need to check is that $p\Phi|_{\partial D^n}$ is an attaching map. I.e. it has image in the $n-1$-skeleton of M_f, which we have constructed inductively. But this is a consequence of f being cellular. Note that $\Phi|_{\partial D^n}$ already has image in Z^{n-1} and any element in $X^{n-1} \times \{1\}$ will be identified with an element in Y^{n-1} because $f(X^{n-1}) \subseteq Y^{n-1}$. So composing characteristic maps with p does indeed give a CW structure on the mapping cylinder M_f.

Definition 2.4.9 (Subcomplex [Hat01, p. 520])**.** A subcomplex A of a CW complex X is a subset $A \subseteq X$ such that if $e^k \subseteq A$, then $\mathrm{Cl}(e^k) \subseteq A$ for all cells in A.

Example 2.4.10 (Skeleton)**.** The n-skeleton X^n is a subcomplex because the closure of an n-cell still lives in the n-skeleton.

Theorem 2.4.11 (Closure Finiteness [Hat01, p. 520])**.** *A compact subspace of a CW complex is contained in a finite subcomplex.*

Proof. Let $C \subseteq X$ be a compact subset of a CW complex X. Then suppose that there is an infinite sequence of points x_i such that each x_i lies in a different cell from the other x_i's. Then let S be the set of the x_i's. Note that S is closed. This can be seen from induction. If $S \cap X^{n-1}$ is closed in X^{n-1}, then for each cell e^n_α of X, we have that $\varphi^{-1}_\alpha(S)$ is closed in ∂D^n_α and $\Phi^{-1}_\alpha(S)$ is at most one more point in D^n_α, so $\Phi^{-1}_\alpha(S)$ is also closed in D^n. So $S \cap X^n$ is closed in X^n for each n and hence S is closed in X. But we can use a similar argument to show that any subset of S is closed, so S is discrete. But S is a closed subset of C, so it is compact, but if it is discrete, it cannot be compact. A contradiction. So we conclude that S is finite. So C is contained in a finite union of cells, so we need only show that a finite union of cells is contained in a finite subcomplex of X. Note that a finite union of finite subcomplexes is still a finite subcomplex. So we need only show that a cell e^n_α is contained in a finite subcomplex. But the image of the attaching map φ_α for e^n_α is compact, so an inductive argument tells us that the image is in a finite subcomplex A of X. So e^n_α is contained in $A \cup e^n_\alpha$. □

Remark 2.4.12 ([Hat01, p. 521])**.** We can now explain the meaning behind the CW in CW complex. The C is for closure finiteness, which we have just proven and the W is for the weak topology.

Definition 2.4.13 ([Hat01, p. 14])**.** A pair (X, A) has the homotopy extension property if given a map $f_0 : X \to Y$ and homotopy $f_t : A \to Y$, then there is a homotopy $g_t : X \to Y$ such that $g_0(x) = f_0(x)$ and $f_t(x) = g_t(x)$ for all $x \in A$.

Theorem 2.4.14 (Homotopy extension property [Hat01, p. 15])**.** *If (X, A) is a CW pair, then (X, A) has the homotopy extension property.*

Proof. We have a retraction $r : D^n \times I \to D^n \times \{0\} \cup \partial D^n \times I$ given by a similar retraction for a disk onto its boundary that we constructed before. Then we get a homotopy $r_t = tr + (1-t)\,\mathrm{id}$ to get a deformation retraction. But then this gives a deformation retraction of $X^n \times I$ onto $X^n \times \{0\} \cup (X^{n-1} \cup A^n) \times I$ because we build $X^n \times I$ from $X^n \times \{0\} \cup (X^{n-1} \cup A^n) \times I$ from $D^n \times I$ attaching along $D^n \times \{0\} \cup \partial D^n \times I$. Then we perform the deformation retraction of $X^n \times I$ onto $X^n \times \{0\} \cup (X^{n-1} \cup A^n) \times I$ in the time interval $[1/2^{n+1}, 1/2^n]$, we get a deformation retraction of $X \times I$ onto $X \times \{0\} \cup A \times I$, and this deformation retraction existing implies the homotopy extension property. □

2.5 The Cellular Approximation Theorem

Recall the theorem:

Theorem 2.5.1 (Cellular Approximation [Hat01, p. 349])**.** *Every map $f : X \to Y$ of CW complexes is homotopic to a cellular map. If f is already cellular on a subcomplex $A \subseteq X$, the homotopy can be taken to be stationary on A.*

Proof. We proceed by induction. First assume that f has been homotoped to be cellular on the $k-1$ skeleton. Then let e^n be an n-cell of X. We leave as an exercise to show that the closure of e^n is the image of the characteristic map. One important property about CW complexes to show this is that compact subsets in CW complexes are closed. But this means that $f(e^n)$ meets finitely many cells so let e^k be a cell of largest dimension that $f(e^n)$ meets. Then we would like to deform $f|_{X^{n-1} \cup e^n}$ staying fixed on X^{n-1} so that it misses a point of e^k. But to do this we need the following lemma:

Lemma 2.5.2 ([Hat01, p. 350])**.** *Let $f : I^n \to Z$ be a map with Z obtained from a subspace W by attaching a cell e^k. Then f is homotopic* $\operatorname{rel} f^{-1}(W)$ *to a map f_1 for which there is a simplex $\Delta^k \subseteq e^k$ with $f_1^{-1}(\Delta^k)$ a union of finitely many convex polyhedra, on each of which f_1 is the restriction of a linear surjection $\mathbb{R}^n \to \mathbb{R}^k$.*

Because the proof is more analytic, we will not be including it in these notes but the curious reader should see the proof in [Hat01].
To continue the proof, compose $f : X^{n-1} \cup e^n \to Y^k$ with a characteristic map for e^n to get a map $f\Phi$ as in the lemma with Y^k taking the place of Z and $Y^k \setminus e^k$ taking the place of W. Then we get a homotopy of $f\Phi \operatorname{rel}(f\Phi)^{-1}(W)$. But $(f\Phi)^{-1}(W) = \partial I^n$ so this homotopy is fixed on ∂I^n so we get a homotopy f_t of $f|_{X^{n-1} \cup e^n}$ fixed on X^{n-1} but when $k > n$, there is no surjective map $\mathbb{R}^n \to \mathbb{R}^k$ so $f_1^{-1}(\Delta^k)$ is empty and we let $p \in \Delta^k$. Then we have a deformation retraction of $Y^k \setminus \{p\}$ onto $Y^k \setminus e^k$ from something similar to our deformation retraction of a punctured disk onto its boundary. We then compose f with this deformation retraction to get a homotopy of f to a map that misses e^k. Then since $f(e^n)$ only met finitely many cells, we can iterate this process finitely many times to have $f(e^n)$ miss all cells of dimension greater than n. Then doing this for each n-cell and staying fixed where f is already cellular, we get a homotopy of $f|_{X^n} \operatorname{rel} X^{n-1} \cup A^n$ to a cellular map. Then homotopy extension allows us to extend this homotopy to all of X. □

2.6 Other Fun Things

This theorem turns out to be more of a powerful technical tool than one with many interesting direct corollaries. This section goes through the most direct corollary possible, which turns out to be a generalization of our motivating question. Then we examine what the theorem says in general and give some examples of theorems where it is used in the proof. It is important to note that this section assumes higher-level algebraic topology knowledge but we will still give some intuition for each construction and result.

2.6.1 Homotopy Groups

Like with our motivating question, a big motivating question in algebraic topology is whether we can classify maps between spaces X and Y up to homotopy. In particular, we want to know about maps between spheres S^n and S^m. Even with spaces as simple as spheres, this turns out to be a hard question. In fact, classifying maps from CW complexes relies heavily on knowing maps between spheres as they are spaces that are built from spheres. We can formalize this in the case that one of the spaces is a sphere using the following construction:

Definition 2.6.1 (Homotopy Groups [Hat01, p. 340])**.** For a space X with basepoint x_0, define $\pi_n(X, x_0)$ to be the set of homotopy classes of maps $f : (I^n, \partial I^n) \to (X, x_0)$ with the homotopies f_t satisfying $f_t(\partial I^n) = x_0$ for all t. When $n \geq 1$, we define an operation in $\pi_n(X, x_0)$ by:

$$(f+g)(s_1, \dots, s_n) = \begin{cases} f(2s_1, s_2, \dots, s_n) & s_1 \in [0, 1/2] \\ g(2s_1 - 1, s_2, \dots, s_n) & s_1 \in [1/2, 1] \end{cases}$$

This makes $\pi_n(X, x_0)$ into a group with identity the constant map and inverses given by $f(1 - s_1, s_2, \dots, s_n)$.

Remark 2.6.2. In view of this, when we say that $\pi_n(X, x_0) = 0$, this means that any map $(I^n, \partial I^n) \to (X, x_0)$, or equivalently, $(S^n, s_0) \to (X, x_0)$ is homotopic to a constant map with the homotopy fixing the basepoint.

The following proposition is a generalization of our earlier example where we showed that any map $S^1 \to S^2$ can be homotoped to a constant. [[Hat01, p. 349]] $\pi_i(S^n) = 0$ for $i < n$.

Proof. Use cellular approximation to homotope a map $S^i \to S^n$ to a constant. Details are left as an exercise. □

In general, the main power of this theorem is that as we have seen before, cellular maps behave nicely with constructions such as the mapping cylinder because it gives a CW structure on the mapping cylinder. One application of this is the following:

Theorem 2.6.3 (Whitehead's Theorem [Hat01, p. 346])**.** *If a map $f : X \to Y$ between connected CW complexes induces isomorphisms $f_* : \pi_n(X) \to \pi_n(Y)$ for all n, then f is a homotopy equivalence. In case f is the inclusion of a subcomplex $X \hookrightarrow Y$, the conclusion is stronger: X is a deformation retract of Y.*

Proof. We first prove the case for $i : X \hookrightarrow Y$. To do this, we would first need to get a retraction $Y \to X$. To do this, we will use the following lemma:

Lemma 2.6.4 ([Hat01, p. 346])**.** *Let (X, A) be a CW pair and let (Y, B) be any pair with $B \neq \emptyset$. For each n such that $X \setminus A$ has cells of dimension n, assume that $\pi_n(Y, B, y_0) = 0$ for all $y_0 \in B$. Then every map $f : (X, A) \to (Y, B)$ is homotopic* rel A *to a map $X \to B$.*

Proof. Assume inductively that f has been homotoped to take the skeleton X^{k-1} to B. If Φ is the characteristic map of a cell e^k of $X \setminus A$, the composition $f\Phi$ can be homotoped into B because $\pi_n(Y, B, y_0) = 0$ by definition. This then induces a homotopy on the quotient space $X^{k-1} \cup e^k$ relative to ∂X^{k-1}. Then we do this for all k-cells of $X \setminus A$ to get a homotopy of $f|_{X^k \cup A}$ into B and homotopy extension gives a homotopy defined on all of X. □

Since i_* is an isomorphism on all homotopy groups, from the long exact sequence of homotopy groups:

$$\cdots \to \pi_n(X) \xrightarrow{i_*} \pi_n(Y) \to \pi_n(X, Y) \to \pi_n(X) \xrightarrow{i_*} \pi_n(Y)$$

we know that $\pi_n(X, Y) = 0$ for all n. So the lemma applied to the identity $(X, Y) \to (X, Y)$ gives us a deformation retraction.
For the general case, we pass to the mapping cylinder. First note that there is a retraction of M_f onto Y by:

$$r(x) = \begin{cases} (x', 1) & x = (x', s) \in X \times I \\ x & x \in Y \end{cases}$$

then we get a homotopy from r to the identity by:

$$r_t(x) = \begin{cases} (x', s^t) & x = (x', s) \in X \times I \\ x & x \in Y \end{cases}$$

so in fact, M_f deformation retracts onto Y. But this means that f can be written as the composition $X \hookrightarrow M_f \xrightarrow{r} Y$. Since we can deform f to be cellular, we can take M_f to be a CW complex. But then the map $X \hookrightarrow M_f$

is inclusion of a subcomplex. So if we can show that $X \hookrightarrow M_f$ induces isomorphisms on all homotopy groups, we are done by the special case. If we let i be the inclusion $X \hookrightarrow M_f$, then we have that $f_* = r_* i_*$. Composing f_* with its inverse on the left we get $\mathrm{id}_* = f_*^{-1} r_* i_*$. So i_* is injective. And we also get that $\mathrm{id}_* = i_* f_*^{-1} r_*$ so i_* is surjective, hence an isomorphism. □

The following application is more technical and requires a lot more machinery than the previous applications, such as cohomology and Steenrod squares, but it demonstrates the general idea of when we would like to use cellular approximation for some nice concrete cases. Namely, that we often have a map into a CW complex that has a nice skeleton at some level, and cellular approximation allows us to only care about mapping into the nice skeleton.

Proposition 2.6.5 ([Hat01, p. 493])**.** *If $n = 2^r(2s+1)$, then the sphere S^{n-1} cannot have 2^r orthonormal tangent vector fields if $s \geq 1$.*

Proof. Let $V_{n,k}$ be the space of orthonormal k-frames in $\mathbb{R}^n$. I.e. each element of $V_{n,k}$ is k-linearly independent vectors in $\mathbb{R}^n$. Then we get a projection $p : V_{n,k} \to S^{n-1}$ from projection onto the first component. Then a section $f : S^{n-1} \to V_{n,k}$ is a map such that $pf(x) = x$ for all x. In particular, this means that $f(x) = (x, v_1(x), \ldots, v_{n-1}(x))$. In particular, each of the v_i's defines an orthonormal vector field on S^{n-1} that are all linearly independent. A section of this bundle amounts to $k-1$ linearly independent vector fields on S^{n-1}. But we can deform f to be cellular by cellular approximation and it turns out that the $n-1$-skeleton of $V_{n,k}$ is $\mathbb{R}\mathrm{P}^{n-1}/\mathbb{R}\mathrm{P}^{n-k-1}$ ([Hat01] 3.D). So we get a map $g : S^{n-1} \to \mathbb{R}\mathrm{P}^{n-1}/\mathbb{R}\mathrm{P}^{n-k-1}$. In particular, we get a map on cohomology:

$$g^* : H^{n-1}(S^{n-1}; \mathbb{Z}/2\mathbb{Z}) \to H^{n-1}(\mathbb{R}\mathrm{P}^{n-1}/\mathbb{R}\mathrm{P}^{n-k-1}; \mathbb{Z}/2\mathbb{Z})$$

and this map is surjective because $g^* p^* = \mathrm{id}^*$ so g^* has a right inverse. And since $k > 0$, we get that $H^{n-1}(\mathbb{R}\mathrm{P}^{n-1}/\mathbb{R}\mathrm{P}^{n-k-1}; \mathbb{Z}/2\mathbb{Z}) \cong H^{n-1}(\mathbb{R}\mathrm{P}^{n-1}; \mathbb{Z}/2\mathbb{Z}) \cong \mathbb{Z}/2\mathbb{Z}$ so g^* is a map $\mathbb{Z}/2\mathbb{Z} \to \mathbb{Z}/2\mathbb{Z}$ that is surjective, so it is bijective and has an inverse. In addition, we have an operation:

$$Sq^{k-1} : H^{n-k}(\mathbb{R}\mathrm{P}^{n-1}/\mathbb{R}\mathrm{P}^{n-k-1}; \mathbb{Z}/2\mathbb{Z}) \to H^{n-1}(\mathbb{R}\mathrm{P}^{n-1}/\mathbb{R}\mathrm{P}^{n-k-1}; \mathbb{Z}/2\mathbb{Z})$$

is nonzero when k is such that $\binom{n-k}{k-1} \equiv 1 \pmod 2$ but this should be zero because the operation $Sq^{k-1} : H^{n-k}(S^{n-1}; \mathbb{Z}/2\mathbb{Z}) \to H^{n-1}(S^{n-1}; \mathbb{Z}/2\mathbb{Z})$ is zero and the map g^* is an isomorphism. Then let $k = 2^r + 1$ with $n = 2^r(2s+1)$ and some algebraic manipulation from here yields the result in the proposition. □

Bibliography

[Abb15] Stephen Abbott. *Understanding Analysis.* Springer, 2015.

[Hat01] Allen Hatcher. *Algebraic Topology.* Cambridge University Press, 2001.

[Mun14] James Munkres. *Topology.* Pearson, 2014.

3. The Braid Symmetries of a Disk

—*Olivia Hu*—

Abstract

A braid in math is not too far from how it sounds — a collection of strands interlacing and intertwining as they travel through paths in space. These braids are interesting creatures. Some braids look complex, but are just one twist from being untied; other braids can look simple, but are in truth much more obstinate. However, the study takes an unexpected turn upon discovering braids, like numbers, can be multiplied and divided to create new braids: they form an abstract structure called a group. By studying this group, we will realize that by studying braids, we are actually also studying the symmetries of a disk.

3.1 Preliminaries and Notation

In this paper, we will assume a working knowledge of basic set theory and the functions between them. The bare minimum assumptions on sets and functions are listed further below. Additionally, we will assume continuity of multivariable functions — an understanding on the level of a first multivariate calculus course should be enough to follow intuitively. Finally, we assume elementary knowledge of matrices and determinants, though this is purely contained in examples and informal discussions — it is not strictly necessary to understand the core content.

This is technically all the required prerequisites, but the reader without any previous exposure to group theory may find this paper unfairly demanding. Nonetheless, Section 3.2 gives a primer on all the group theory required from the beginning. We hope it is enough to fill in any gaps in knowledge.

Throughout this paper, many of the definitions and proofs will come with informal discussions alongside the full technical details. It is not necessary to fully understand every technical detail, only to glean some intuition. The formal proofs are generally not essential for the story, so feel free to skip them. The subject of this paper is very visual, and the core ideas can nearly always be grasped by intuition alone.

Readers are encouraged to skip any sections they are already familiar with.

Notation

$\varnothing$	The empty set
$s \in S$	s is an element of the set S
$A \subset B$	A is a subset of B, or $a \in A$ implies $a \in B$
$A \cup B$	The union of A and B $\{s : s \in A \text{ or } s \in B\}$
$A \cap B$	The intersection of A and B $\{s : s \in A \text{ and } s \in B\}$
$A \setminus B$	The set difference of A and B $\{s : s \in A \text{ and } s \notin B\}$
$A \times B$	The cartesian product $\{(a, b) : a \in A \text{ and } b \in B\}$
S^n	The cartesian product $\{(s_1, s_2, \dots, s_n) : s_j \in S\}$
$f : A \to B$	A function from the set A to the set B
$a \mapsto b$	(of a function) maps the element a to the element b
$\mathrm{id} : A \to A$	The identity function that takes each point of A to itself.

Sets

$\mathbb{Z}$	The set of integers $\{\dots, -2, -1, 0, 1, 2, \dots\}$
$\mathbb{R}$	The set of real numbers
D^n	The n-dimensional unit disk $\big\{(x_1, \dots, x_n) \in \mathbb{R}^n : \sum_{i=1}^{n} x_j^2 \leq 1\big\}$
S^n	The n-dimensional unit sphere $\big\{(x_1, \dots, x_{n+1}) \subset \mathbb{R}^{n+1} : \sum_{i=1}^{n+1} x_i^2 = 1)\big\}$.

Definition 3.1.1 (Properties of functions)**.** A function $f : A \to B$ between sets is **injective** if for all $a, a' \in A$ such that $a \neq a'$, $f(a) \neq f(a')$.

Likewise, f is **surjective** if for all $b \in B$, there exists $a \in A$ such that $f(a) = b$. If f is both injective and surjective, we say f is **bijective**.

If there exists $g : B \to A$ such that for all $a \in A$ and $b \in B$,

$$(f \circ g)(b) = b \text{ and } (g \circ f)(a) = a,$$

we say f is **invertible**.

Theorem 3.1.2. *Let $f : A \to B$ be a function between sets. Then f is bijective if and only if f is invertible.*

3.2 Introduction to Groups

Definition 3.2.1 (Groups, [DF04, p. 16])**.**

Informal. We can think of groups as a way to generalize adding and multiplying to more abstract settings. For example, in the familiar situation of adding integers together, we can think of addition as a function whose input is an ordered pair (a, b) and whose output is another integer we call $a + b$. We call this a *binary operation*, and the operation $+$ has the following nice properties in the integers:

(i) for all $a, b, c \in \mathbb{Z}$, $(a + b) + c = a + (b + c)$,

(ii) for all $a \in \mathbb{Z}$, $a + 0 = a = 0 + a$,

(iii) for all $a \in \mathbb{Z}$, $a + (-a) = 0 = (-a) + a$,

(iv) for all $a, b \in \mathbb{Z}$, $a + b = b + a$.

A group generalizes this idea to more abstract sets than the integers and more abstract operations than addition. For instance, we will see in Example 3.2.9 later that the symmetries of a triangle satisfy properties (i), (ii), and (iii). Further in the paper, we will see Mapping Class groups (Def. 3.5.6) and Braid groups (Def. 3.4.8). Any set with an operation that satisfies (i), (ii), and (iii), we call a *group* under that operation. If we further have property (iv), that group is called *abelian*.

Formal. A **group** is a pair $(G, *)$ where G is a set and $*$ is a **binary operation**

$$* : G \times G \to G,$$

where we denote $*(g_1, g_2) = g_1 * g_2$, satisfying the **group axioms**:

(i) **Associativity**. For all $g_1, g_2, g_3 \in G$, $(g_1 * g_2) * g_3 = g_1 * (g_2 * g_3)$.

(ii) **Identity**. There exists $e \in G$ such for all $g \in G$, $e * g = g = g * e$. We call e an **identity element** of G, or simply an identity.

(iii) **Inverses**. For all $g \in G$, there exists $h \in G$ such that $h * g = e = g * h$. We call h an **inverse** of g.

A group that further satisfies

(iv) **Commutativity**. For all $g, h \in G$, $g * h = h * g$

is called **abelian**.

Proposition 3.2.2 ([DF04, p. 18])**.** *If $(G, *)$ is a group, then*

(i) The identity of G is unique,

(ii) For each $g \in G$, the inverse of g is uniquely determined.

By Proposition 3.2.2, we can unambiguously denote the identity of a group $(G, *)$ by 1 and the inverse of an element $g \in G$ by g^{-1}. Then for any $n \in \mathbb{Z}$ and $g \in G$, we define

$$g^n = \begin{cases} g^n = 1 & \text{if } n = 0, \\ \underbrace{g * g * \cdots * g}_{n \text{ times}} & \text{if } n > 0, \\ (g^{-1})^{-n} & \text{if } n < 0. \end{cases}$$

If $(G, *)$ is abelian, we often choose the symbols 0, $-g$ and ng in place of 1, g^{-1}, and g^n. Moving forward, we will refer to a group $(G, *)$ simply as G if the operation is clear.

Example 3.2.3. As discussed in the informal section of Definition 3.2.1, the integers under addition form an abelian group. That is, $(\mathbb{Z}, +)$ is an abelian group.

Example 3.2.4. Any set G with just one element, $G = \{e\}$, is a group by the binary operation $e * e = e$. Indeed, this operation is associative, e is the identity, and e is its own inverse. This group is abelian as well.

We call this the **trivial group**.

Example 3.2.5. The integers under multiplication do not form a group because no elements besides 0, -1, and 1 have multiplicative inverses. That is, $(\mathbb{Z}, \cdot)$ is not a group.

Example 3.2.6. The real numbers under addition, $(\mathbb{R}, +)$, forms a group for similar reasons to $(\mathbb{Z}, +)$. However, the real numbers under multiplication, $(\mathbb{R}, \cdot)$, does not form a group because 0 does not have a multiplicative inverse. However, if we define

$$\mathbb{R}^\times = \mathbb{R} \setminus \{0\},$$

then $(\mathbb{R}^\times, \cdot)$ forms an abelian group.

Example 3.2.7. The integers under subtraction, $(\mathbb{Z}, -)$, is not a group. Although $-$ admits an identity and inverses, $-$ is not associative. For example,

$$(1-2)-(-3)=2$$

while

$$1-(2-(-3))=-4.$$

Example 3.2.8. Fix a positive integer n. The space $M(n,\mathbb{R})$ of $n \times n$ matrices is not a group under multiplication. Although the multiplication is associative and the diagonal matrix of 1's is an identity, not every matrix admits an inverse. For example,

$$\begin{bmatrix} 2 & 6 \\ 1 & 3 \end{bmatrix} \in M(2,\mathbb{R})$$

has determinant 0, thus is not invertible.

If we restrict to the invertible matrices, $GL(n,\mathbb{R}) \subset M(n,\mathbb{R})$, then $GL(n,\mathbb{R})$ is a group under multiplication. Note $(GL(n,\mathbb{R}),\cdot)$ is not abelian. For example,

$$\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix},$$

but

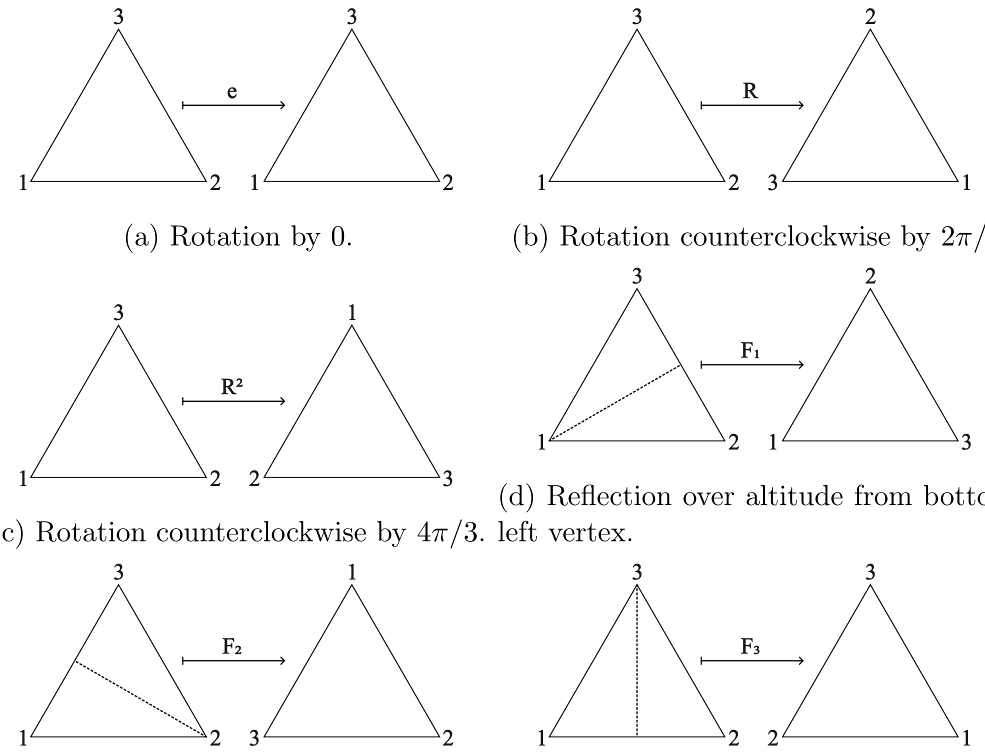
$$\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}.$$

Example 3.2.9 (Dihedral group)**.** Our first unfamiliar example: the symmetries of a triangle, D_3, form a group under composition $\circ$.

By symmetries of a triangle, we mean this. Given an equilateral triangle, counterclockwise rotations by 0, $2\pi/3$, and $4\pi/3$ about the center result in the same triangle with the vertices in different positions. We call these rotations e, R, and R^2 respectively. In Figure 3.1, these rotations correspond to (a), (b), and (c).

Another three symmetries are given by reflections across the three altitudes of the triangle. We call these reflections F_1, F_2, and F_3, pictured in (d), (e), and (f) of Figure 3.1. These symmetries comprise the elements of D_3, making six in total.

The operation is composition — that is, if $A, B \in D_3$, then $A \circ B$ means apply B to the triangle, then apply A to the resulting triangle (see Examples in Figure 3.2). Note that R^2 is simply $R \circ R$, justifying the notation. We can work out by force that for any two elements $A, B \in D_3$, $A \circ B \in D_3$, so $\circ$ is a well-defined binary operation $D_3 \times D_3 \to D_3$.



(a) Rotation by 0.

(b) Rotation counterclockwise by $2\pi/3$

(c) Rotation counterclockwise by $4\pi/3$.

(d) Reflection over altitude from bottom left vertex.

(e) Reflection over altitude from bottom right vertex.

(f) Reflection over altitude from top vertex.

Figure 3.1: Symmetries of the triangle.

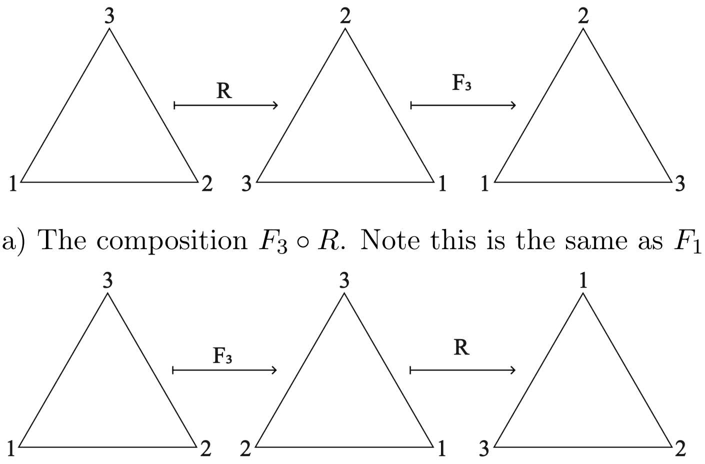
In fact, $\circ$ is associative, e is the identity, and every element has an inverse ($R^{-1} = R^2$, the remaining elements are their own inverses). It follows that $(D_3, \circ)$ is a group. Note that D_3 is not abelian by the Example in Figure 3.2.

In general, the symmetries of an n-gon is denoted D_n, and $(D_n, \circ)$ is a group with $2n$ elements. This is known as the **Dihedral group**.

Example 3.2.10 (Symmetric group)**.** Let $n \in \mathbb{Z}$ be positive. A **permutation** on the set $S = \{1, 2, \ldots, n\}$ is a bijection $S \to S$. We denote by S_n the set of all permutations of S. Note that the identity map $\mathbb{1} \in S_n$ that sends every element to itself is an identity under composition. That is, for every $\sigma \in S_n$,

$$\sigma \circ \mathbb{1} = \mathbb{1} \circ \sigma = \sigma.$$

Moreover, bijections are invertible, so $\sigma \in S_n$ implies there exists $\sigma^{-1} \in S_n$. Finally, function composition is associative, so $(S_n, \circ)$ is a group. We call this the **symmetric group on n indices**.



(a) The composition $F_3 \circ R$. Note this is the same as F_1.

(b) The composition $R \circ F_3$. Note this is the same as F_2.

Figure 3.2: Examples of compositions in D_3. Observe $F_3 \circ R \neq R \circ F_3$.

Definition 3.2.11.

Informal. Note that the Dihedral group D_3 of Example 3.2.9 is kind of similar to the symmtric group S_3 of Example 3.2.10. For example, note that the rotation R by $2\pi/3$ of Figure 3.1(b) maps the indices like

$$1 \mapsto 3,\ 2 \mapsto 1,\ 3 \mapsto 2.$$

But this is a permutation of $\{1, 2, 3\}$, we can define a function $D_3 \to S_3$ that sends R to the permutation above and the other symmetries to their corresonding permutations. This hints at some deep similarity between D_3 and S_3 — in fact, we say they are *isomorphic*, denoted $D_3 \cong S_3$.

There are also weaker relationships between groups. For example, it is true that for any $A, B \in GL(n, \mathbb{R})$ (see Example 3.2.8),

$$\det(AB) = \det(A)\det(B).$$

That is, the determinant turns multiplication in $GL(n, \mathbb{R})$ into multiplication in $\mathbb{R}^\times$ (see Example 3.2.6). So somehow, the determinant is telling us the multiplications on $GL(n, \mathbb{R})$ and $\mathbb{R}^\times$ induce somewhat similar group structures. We say that det is a *homomorphism*. An isomorphism is a homomorphism that is really nice — one that is invertible.

Formal. Let $(G, *)$, $(G', \star)$ be groups. A **group homomorphism** from G to G' is a map

$$\varphi : G \to G'$$

such that for all $a, b \in G$,

$$\varphi(a * b) = \varphi(a) \star \varphi(b).$$

A group homomorphism $\varphi : G \to G'$ is a **group isomorphism** if φ is invertible. Then we denote $G \cong G'$.

Example 3.2.12. Given any group $(G, *)$, the identity map $\mathrm{id} : G \to G$, defined by

$$\mathrm{id}(g) = g$$

for all $g \in G$, is a group homomorphism. To see this, observe that for any $a, b \in G$,

$$\mathrm{id}(a * b) = a * b = \mathrm{id}(a) * \mathrm{id}(b).$$

In fact, id is also its own inverse, thus an isomorphism.

A concrete example of this is the familiar function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x$.

Example 3.2.13. Given any groups $(G, *)$, $(G', \star)$, where we denote the identity of G' by $\mathbb{1}'$, the function $\varphi : G \to G'$ defined by

$$\varphi(g) = \mathbb{1}'$$

for all $g \in G$ is a group homomorphism.

Example 3.2.14. As we discussed in Definition 3.2.11, the determinant

$$\det : GL(n, \mathbb{R}) \to \mathbb{R}^{\times}$$

is a group homomorphism. In fact, one can show det is a surjective but not injective homomorphism.

Example 3.2.15. The function $\varphi : \mathbb{Z} \to \mathbb{Z}$ that maps

$$\varphi(a) = 2a$$

for all $a \in \mathbb{Z}$ is a group homomorphism. To see this, observe that for any $a, b \in \mathbb{Z}$,

$$\varphi(a + b) = 2(a + b) = 2a + 2b = \varphi(a) + \varphi(b).$$

Additionally, if $a \neq b$, then $2a \neq 2b$, so φ is injective. However, φ is not surjective as it does not reach any odd numbers. Note though, that φ is bijective as a function from $\mathbb{Z}$ to $2\mathbb{Z}$. We will see another example of this occuring below.

Example 3.2.16. A very interesting example is the exponential map $\exp : \mathbb{R} \mapsto \mathbb{R}^\times$ given by

$$\exp(x) = e^x$$

for all $x \in \mathbb{R}$. Note that this is a homomorphism from $(\mathbb{R}, +)$ to $(\mathbb{R}^\times, \cdot)$ because for all $x, y \in \mathbb{R}$,

$$\exp(x + y) = e^{x+y} = e^x e^y = \exp(x) \cdot \exp(y).$$

Note that e^x is an invertible function from $\mathbb{R}$ to the positive real numbers $\mathbb{R}^+$ (since the inverse is $\log(x)$), so exp is injective, but only surjective on $\mathbb{R}^+$. Ultimately, exp is not surjective on all of $\mathbb{R}$.

This leads in to our next topic.

Definition 3.2.17 (Subgroups [DF04, p. 22])**.**

Informal. The homomorphisms examined in Examples 3.2.15 and 3.2.16 were almost isomorphisms. For φ in Example 3.2.15, φ was bijective as a function $\mathbb{Z} \to 2\mathbb{Z}$. For exp in 3.2.16, exp was bijective as a function $\mathbb{R} \to \mathbb{R}^+$. In fact, we can show $(2\mathbb{Z}, +)$ and $(\mathbb{R}^+, \times)$ are groups in their own right, so φ and exp give isomorphisms $\mathbb{Z} \cong 2\mathbb{Z}$ and $(\mathbb{R}, +) \cong (\mathbb{R}^+, \cdot)$ respectively. They give isomorphisms to groups that lie inside other groups.

This sort of situation happens quite frequently: within a group are often other groups, which are in fact groups by the same operation. If a group G contains a group H, we call H a *subgroup* of G.

Formal. Let $(G, *)$ be a group, and let H be a nonempty subset of G. If we have the properties:

(i) **Closure under inverses**. For all $h \in H$, $h^{-1} \in H$.

(ii) **Closure under** $*$. For all $h, k \in H$, $h * k \in H$,

then we say $(H, *)$ is a **subgroup** of G. We often denote this by $H \leq G$.

Note these conditions tell us $(H, *)$ is a group in its own right, independent of G.

Example 3.2.18. Let G be any group. Then $\{1\}, G \leq G$ are easy examples of subgroups.

Example 3.2.19. As noted in Definition 3.2.17, $2\mathbb{Z} \leq \mathbb{Z}$ and $\mathbb{R}^+ \leq \mathbb{R}^\times$.

Example 3.2.20. The negative real numbers $\mathbb{R}^-$ not a subgroup of $\mathbb{R}^\times$ because $-1 \in \mathbb{R}^-$, but $(-1)(-1) = 1$ is not in $\mathbb{R}^\times$. Therefore, $\mathbb{R}^-$ is not closed under multiplication.

Example 3.2.21. Recall the definition of $GL(n, \mathbb{R})$ from Example 3.2.8. Let

$$SL(n, \mathbb{R}) = \{A \in GL(n, \mathbb{R}) : \det(A) = 1\}.$$

This is a subgroup by some properties of matrices:

(i) for any $A \in SL(n, \mathbb{R})$, $\det(A^{-1}) = 1$ implies $A^{-1} \in SL(n, \mathbb{R})$,

(ii) for any $A, B \in SL(n, \mathbb{R})$, $\det(AB) = 1$ implies $AB \in SL(n, \mathbb{R})$.

Example 3.2.22. From Example 3.2.9, note that $\{e, R_1, R_2\} \leq D_3$. Indeed, the inverses of rotations are all rotations, and the compositions of rotations are rotations as well.

We also have $\{e, F_i\} \leq D_3$ for all $i = 1, 2, 3$. This is because F_i is always its own inverse.

From the discussion in Definition 3.2.17, we have managed to recover isomorphisms from the homomorphisms in Examples 3.2.15 and 3.2.16 by considering subgroups. But can we do the same for a homomorphism like det in Example 3.2.14, which is surjective but not injective? We would like to “shrink” the domain in such a way that no two elements take on the same value, while still maintaining a group structure. To do something like this, we will need to take a detour to set theory.

Definition 3.2.23 (Partitions [DF04, p. 3])**.**

Informal. If we're working with some set A, it's often useful to separate the elements of A into smaller subsets, maybe based on some property. For example, $\mathbb{Z}$ can be divided up into the subsets

$$\begin{aligned} 3\mathbb{Z} &= \{\ldots, -6, -3, 0, 3, 6, \ldots\}, \\ 3\mathbb{Z} + 1 &= \{\ldots, -5, -2, 1, 4, 7, \ldots\}, \\ 3\mathbb{Z} + 2 &= \{\ldots, -4, -1, 2, 5, 8, \ldots\}. \end{aligned}$$

These subsets divide up $\mathbb{Z}$ really nicely in the sense that none of their elements overlap and their union is all of $\mathbb{Z}$. When we divide up a set A into subsets whose intersections with each other are empty and union is the whole set, we get a *partition* of A.

Formal. A **partition** of a nonempty set A is a collection $\{A_i : i \in I\}$ of nonempty subsets of A (here, I is an **indexing set** that helps us keep track of the sets in the collection. Common ones are the positive integers $\mathbb{Z}^+$ and $\mathbb{R}$) such that

$$A = \bigcup_{i \in I} A_i$$

is the union of all the A_i and

$$A_i \cap A_j = \varnothing$$

for all $i, j \in I$ such that $i \neq j$.

Example 3.2.24. Let A be a nonempty set. The partition consisting of just the subset $A \subset A$ is technically a partition of A.

Example 3.2.25. Let A be a nonempty set. The partition consisting of the one-element set $\{a\} \subset A$ for all $a \in A$ is a partition on A.

Example 3.2.26. From the discussion in Definition 3.2.23, $\mathbb{Z}$ is partitioned by the subsets $3\mathbb{Z}$, $3\mathbb{Z} + 1$, and $3\mathbb{Z} + 2$.

These partitions are intimately related to the next idea.

Definition 3.2.27 (Equivalence relations [DF04, p. 3])**.**

Informal. Seemingly a departure from what we've been talking about, we would now like to generalize the idea of equality. Note that equality, =, has the following properties in a set A:

(i) for all $a \in A$, $a = a$,

(ii) for all $a, b \in A$, $a = b$ implies $b = a$,

(iii) for all $a, b, c \in A$, $a = b$ and $b = c$ implies $a = c$.

We call any relationship of elements in a set satisfying the same types of properties is an *equivalence relation.*

Formal. A **relation** on a nonempty set A is a subset R of $A \times A$, and we write $a \sim b$ if and only if $(a, b) \in R$. An **equivalence relation** on A is a relation on A satisfying:

(i) **Reflexivity**. For all $a \in A$, $a \sim a$.

(ii) **Symmetry**. For all $a, b \in A$, $a \sim b$ implies $b \sim a$.

(iii) **Transitivity**. For all $a, b, c \in A$, $a \sim b$ and $b \sim c$ implies $a \sim c$.

Example 3.2.28. An easy equivalence relation on a nonempty set A is the entire set $R = A \times A$. This amounts to saying that for all $a, b \in A$, $a \sim b$.

Example 3.2.29. The familiar $=$ relation is given by the subset

$$R = \{(a, a) \in A \times A : a \in A\}.$$

That is, every element is related only to itself.

Example 3.2.30. $\cong$ is an equivalence relation on any set of groups. If G is any group, then $\mathrm{id} : G \to G$ (see Example 3.2.12) is an isomorphism, so $G \cong G$. Symmetry comes from the fact inverses of bijective group homomorphisms are themselves group homomorphisms. Transitivity comes from the fact compositions of homomorphisms are still group homomorphisms. We will omit the proofs of these.

Example 3.2.31. $\leq$ is not an equivalence relation on $\mathbb{Z}$. Although $\leq$ is reflexive and transitive, $\leq$ fails symmetry. For example, $4 \leq 5$, but $5 \not\leq 4$.

Example 3.2.32. Define a relation on $\mathbb{Z}$ via $a \sim b$ if and only if a has the same remainder as b when divided by 3. This is an equivalence relation:

(i) For all $a \in \mathbb{Z}$, a has the same remainder mod 3 as itself.

(ii) For all $a, b \in \mathbb{Z}$, if a has the same remainder as b mod 3, then b has the same remainder as a mod 3.

(iii) For all $a, b, c \in \mathbb{Z}$, $a \sim b$ and $b \sim c$ means a has the same remainder as b when divided by 3, but c also has the same remainder as b when divided by 3. Thus, a must have the same remainder as c when divided by 3, hence $a \sim c$.

Definition 3.2.33 ([DF04, p. 3])**.**

Informal. Given an equivalence relation $\sim$ on a set A and an element $a \in A$, we can think about the subset of elements that relate to A. This is called the *equivalence class* of a with respect to $\sim$. We denote this by $[a]_\sim$.

Formal. Let A be a nonempty set, $\sim$ an equivalence relation on A. Then for all $a \in A$, the **equivalence class of** a with respect to $\sim$ is defined as

$$[a]_\sim = \{b \in A : a \sim b\}.$$

Example 3.2.34. In Example 3.2.28, the equivalence class of any element is the whole set. That is, for all $a \in A$,

$$[a]_\sim = A.$$

Note this gives the sets for the partition for 3.2.24.

Example 3.2.35. In Example 3.2.29, the equivalence class of any element is the set containing just itself. That is, for all $a \in A$,

$$[a]_\sim = \{a\}.$$

Note this gives the sets in the partition for Example 3.2.25.

Example 3.2.36. In Example 3.2.32, the equivalence class of any element is one of $3\mathbb{Z}$, $1 + 3\mathbb{Z}$, and $2 + 3\mathbb{Z}$ depending on if the remainder of that element mod 3 is 0, 1, or 2 respectively. Note that the equivalence classes give precisely the subsets for the partition in 3.2.26.

In Examples 3.2.34, 3.2.35, and 3.2.36, we see that equivalence relations induce partitions via equivalence classes. Interestingly, this pattern always holds, and in fact goes both ways.

Theorem 3.2.37 ([DF04, p. 3])**.** *Let A be a nonempty set.*

(i) If $\sim$ is an equivalence relation on A, then the set of equivalence classes $\sim$ forms a partition of A.

(ii) Conversely, if $\{A_i : i \in I\}$ is a partition of A, then there is an equivalence relation on A whose equivalence classes give that partition.

Proof. (i): Let A be a nonempty set, and suppose $\sim$ is an equivalence relation on A. Then we will show set of equivalence classes of A with respect to $\sim$ form a partition of A.

(a) First,

$$A = \bigcup_{a \in A} [a]_\sim$$

because for all $a \in A$, $a \in [a]_\sim$.

(b) Second, if $[a]_\sim = [b]_\sim$ for some $a, b \in A$, then $a \in [a]_\sim$ implies $a \in [b]_\sim$. But this means $a \sim b$. We can then show this means $[a]_\sim$ and $[b]_\sim$ are actually the same set. Therefore, equivalence classes that are different from each other must have empty intersection, lest they actually be equal.

(ii): Let A be a nonempty set, and suppose $\{A_i : i \in I\}$ is a partition of A. Then define the relation $\sim$ on A where for all $a, b \in A$, $a \sim b$ if and only if $a, b \in A_i$ for some $i \in I$. We will show this is an equivalence relation.

(a) For all $a \in A$, by the definition of a partition of A, $a \in A_i$ for some $i \in I$, so $a \sim a$.

(b) For all $a, b \in A$, $a, b \in A_i$ implies $b, a \in A_i$ pretty self-evidently. Thus, $a \sim b$ implies $b \sim a$.

(c) For all $a, b, c \in A$, $a, b \in A_i$ and $b, c \in A_j$ for some $i, j \in I$ implies $b \in A_i \cap A_j$. Thus, $A_i \cap A_j \neq \varnothing$, which means $i = j$, so $a \sim c$.

Now we will show the equivalences classes of $\sim$ give the desired partition. Given any $i \in I$, we have

$$A_i = [a]_\sim$$

for any $a \in A_i$ by definition of $\sim$. Likewise, for any $a \in A$, $a \in A_i$ for some $i \in I$, so

$$[a]_\sim = A_i.$$

We conclude the equivalence classes of $\sim$ give the desired partition $\{A_i : i \in I\}$. □

So partitions and equivalence relations are really equivalent ideas. We will now circle back to groups. The strategy will be to "shrink down" a group by partitioning that group, then turning the set of subsets in that partition into a new group. First, we will need the following more technical definitions.

Definition 3.2.38 (Normal subgroup [DF04, p. 82])**.** Let G be a group, $N \leq G$ a subgroup of G. We call N a **normal subgroup** of G if for all $n \in N, g \in G$, $gng^{-1} \in N$. We denote this $N \trianglelefteq G$.

Example 3.2.39. $3\mathbb{Z} \leq \mathbb{Z}$ is a normal subgroup because given any $a \in 3\mathbb{Z}$, $b \in \mathbb{Z}$,

$$b + a + (-b) = a \in 3\mathbb{Z}.$$

In fact, the same method shows that subgroups of abelian groups are normal.

Example 3.2.40. Recall the subgroup $\{e, F_1\} \leq D_3$ in Example 3.2.22. This is not a normal subgroup because

$$R_2 \circ F_1 \circ R_1 = F_3 \notin \{e, F_1\}.$$

Example 3.2.41. Recall from Example 3.2.21 that $SL(n,\mathbb{R})$ is a subgroup of $GL(n,\mathbb{R})$. In fact, $SL(n,\mathbb{R})$ is a normal subgroup of $GL(n,\mathbb{R})$. To see this, let $A \in SL(n,\mathbb{R}), B \in GL_n(\mathbb{R})$ be arbitrary. Then

$$\det\bigl(BAB^{-1}\bigr) = \det(B)\det(A)\det(B)^{-1} = \det(A) = 1,$$

so $A \in SL(n,\mathbb{R})$. The subset

$$SL(n,\mathbb{R}) = \{A \in GL(n,\mathbb{R}) : \det(A) = 1\}$$

is a normal subgroup.

Definition 3.2.42 (Cosets [DF04, p. 77])**.** Let $(G,*)$ be a group, $H \leq G$ a subgroup of G. For any $g \in G$, define the set

$$gH = \{g * h \in G : h \in H\}.$$

We call these left **cosets** of H in G. Any element of a coset is a called a **representative** for the coset.

If $(G,+)$ is abelian, we often write $g+H$ in place of gH.

Example 3.2.43. Let G be a group. Recall that $\{1\}, G \leq G$ are subgroups. In fact, $\{1\}, G \trianglelefteq G$ are normal subgroups as well.

Example 3.2.44. The sets $3\mathbb{Z}$, $1+3\mathbb{Z}$, and $2+3\mathbb{Z}$ defined in Definition 3.2.23 are cosets of the subgroup $3\mathbb{Z} \leq \mathbb{Z}$. Numbers like $1, -2, 10$ are all representatives of $1+3\mathbb{Z}$. Note the name representative makes sense, since

$$1+3\mathbb{Z} = -2+3\mathbb{Z} = 10+3\mathbb{Z}.$$

Example 3.2.45. The cosets of $SL(n,\mathbb{R})$ (see Example 3.2.21) are precisely the sets of elements with the same determinant. That is, suppose $A \in GL(n,\mathbb{R})$ has determinant u for some $u \in \mathbb{R}$. Then every element of $A(SL(n,\mathbb{R}))$ has determinant u. Conversely, if another matrix $B \in GL(n)$ has determinant n, then

$$B = A(A^{-1}B)$$

where $\det(A^{-1}B) = 1$ implies $B \in A(SL(n,\mathbb{R}))$. Therefore, we can say det is injective on the cosets of $SL(n,\mathbb{R})$. Then to salvage an isomorphism out of det, we need only create a new group whose elements are the cosets of $SL(n,\mathbb{R})$. We will show how to do this.

Proposition 3.2.46. *Let G be a group, $H \leq G$ any subgroup. Then the cosets of H in G partition G.*

Proof. Let $\sim$ be the equivalence relation on G defined by $a \sim b$ if and only if a and b are in the same coset of H for all $a, b \in G$. This is an equivalence relation because

(i) for any $a \in G$, $a \in aH$, so $a \sim a$,

(ii) for any $a, b \in G$, $a, b \in gH$ implies $b, a \in gH$, so $b, a \in G$.

(iii) for any $a, b, c \in G$, $a, b \in gH$ and $b, c \in g'H$ implies there exist $h, h_1', h_2' \in H$ such that

$$b = gh,\ b = g'h_1',\ c = g'h_2'.$$

We deduce using inverses that

$$g' = bh_1'^{-1} = ghh_1',$$

so

$$c = ghh_1'h_2' \in gH.$$

Thus, $a \sim c$.

Since $\sim$ is an equivalence relation, the equivalence classes of $\sim$ partition G. Note these equivalence classes are precisely the cosets of H. □

Proposition 3.2.47. *Let G be a group, $N \trianglelefteq G$ a normal subgroup. Then for any two cosets aN, bN, the coset abN is independent of the choice of representatives. That is, given any other representatives a', b' of aN and bN respectively, $a'b'N = abN$.*

Proof. Let $g \in abN$. Then we can write $g = abn$ for some $n \in N$. If $a' \in aN$, $b' \in bN$, then we can also write $a' = an_1$, $b' = bn_2$ for $n_1, n_2 \in N$. Then

$$g = abn = (a'n_1^{-1})(b'n_2^{-1})n = a'(b'b'^{-1})n_1^{-1}b'n_2^{-1}n = a'b'(b'^{-1}n_1^{-1}b')n_2^{-1}n \in a'b'N$$

because $b'^{-1}n_1b' \in N$ by N being normal. By the same argument the other direction, $abN = a'b'N$. □

Definition 3.2.48 (Quotient groups)**.**

Informal. By putting a partition $\{A_i : i \in I\}$ on a group G, we can form a new group $\overline{G}$ whose elements are the subsets A_i themselves. A concrete example is helpful. Recall the partition of 3.2.36 on $\mathbb{Z}$ by the cosets $3\mathbb{Z}$,

$1 + 3\mathbb{Z}$, and $2 + 3\mathbb{Z}$. The elements of our new group, which we denote $\mathbb{Z}/3\mathbb{Z}$, are as follows:

$$\mathbb{Z}/3\mathbb{Z} = \{3\mathbb{Z}, 1 + 3\mathbb{Z}, 2 + 3\mathbb{Z}\}.$$

We emphasize that the sets themselves have become elements. $\mathbb{Z}/3\mathbb{Z}$ is a group in the following way. If we define the sum of cosets of $3\mathbb{Z}$ by

$$(a + 3\mathbb{Z}) + (b + 3\mathbb{Z}) = (a + b) + 3\mathbb{Z},$$

then

$$3\mathbb{Z} + 3\mathbb{Z} = 3\mathbb{Z},\ 3\mathbb{Z} + (1 + 3\mathbb{Z}) = 1 + 3\mathbb{Z},\ (1 + 3\mathbb{Z}) + (2 + 3\mathbb{Z}) = 3\mathbb{Z},\ \text{etc.}$$

Note the “niceness” here. The sums of the sets in $\mathbb{Z}/3\mathbb{Z}$ always give another element in $\mathbb{Z}/3\mathbb{Z}$. Moreover, the result stays the same regardless of which representative we choose for each coset. All in all, we get a group structure on $\mathbb{Z}/3\mathbb{Z}$, which we call a the *quotient group* of $\mathbb{Z}$ with respect to the subgroup $3\mathbb{Z}$. This only works because $3\mathbb{Z}$ is normal.

Formal. Let $(G, *)$ be a group, $N \trianglelefteq G$ be a normal subgroup of G. Then define G/N to be the set of cosets of N in G. That is,

$$G/N = \{gN : g \in G\}.$$

The operation is as follows: for all $a, b \in G$,

$$(aN) * (bN) = (a * b)N.$$

By Propositions 3.2.46 and 3.2.47, the definition of this operation is unambiguous — every element uniquely represents a coset and the operation is independent of which representative we choose. We can show $(G/N, *)$ forms a group, which we call the **quotient group** of G with respect to N.

Example 3.2.49. We needed N to be normal in Definition 3.2.48 because otherwise, the group operation is not well-defined. For instance, we showed in Example 3.2.40 that $\{e, F_1\} \leq D_3$ is not a normal. Call this subgroup Δ. Then

$$R_2\Delta = F_2\Delta = \{R_2, F_2\},$$

but then

$$R_2\Delta \circ (R_2\Delta) = R_1\Delta = \{1, F_3\},$$

while

$$F_2\Delta \circ F_2\Delta = \Delta = \{1, F_1\}.$$

Therefore, the operation is not independent of the representative we choose for the cosets, and so does not have a well-defined output. This is rectified by our choice of N to be normal.

Remark 3.2.50. In fact, if $(G, *)$ is a group with subgroup $H \leq G$, then $(G/H, *)$ is a group if and only if H is normal in G.

Example 3.2.51. Recall $\{1\}, G \trianglelefteq G$ are normal subgroups. Then G/G has just one element, while $G/\{1\}$ is isomorphic to G. That is, $G/G \cong \{1\}$, $G/\{1\} \cong G$.

Example 3.2.52. By taking the determinant homomorphism (Example 3.2.14) on the quotient $GL(n, \mathbb{R})/SL(n, \mathbb{R})$ (Example 3.2.41), we get $\overline{\det} : GL(n, \mathbb{R})/SL(n, \mathbb{R}) \to \mathbb{R}^\times$ via

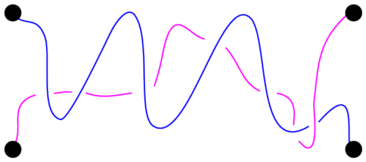
$$\overline{\det}(A(SL_n(\mathbb{R}))) = \det(A)$$

for all $A(SL_n(\mathbb{R})) \in GL(n, \mathbb{R})/SL(n, \mathbb{R})$. We can check that $\overline{\det}$ is independent of the choice of representative, and in fact, is a group isomorphism. Thus, $GL(n, \mathbb{R})/SL(n, \mathbb{R}) \cong \mathbb{R}^\times$.

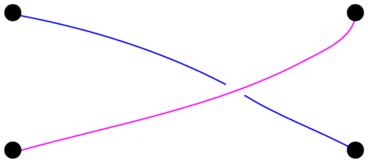
Example 3.2.53. $\mathbb{Z}/3\mathbb{Z}$, as discussed in Definition 3.2.48, is a quotient group. This group represents arithmetic modulo 3, where we only distinguish elements up to their remainder when divided by 3. This idea can be generalized to $\mathbb{Z}/n\mathbb{Z}$ for any $n \in \mathbb{Z}$ a positive integer.

3.3 Homotopy

Before getting to braids, there is one more thing we need — that is the notion of a "continuous deformation." For example, if we are to talk about braids, there must be a bit of wiggle room in what makes a braid. To illustrate, these messy strands



can easily be turned into the braid



just by nudging the blue strand up a bit and nudging the pink strand down a bit. Maybe it would be a bit unfair or unproductive to say these are different braids, simply because their strands do not strictly occupy the same coordinates in space. But to speak about this precisely, we must make precise what we mean by "nudging."

Definition 3.3.1 (Homotopy)**.**

Informal. The most basic form of a "nudge" is a *homotopy*. To illustrate, suppose f and g are functions that parameterize a line segment in $\mathbb{R}^2$. So f, g can be a continuous functions $[0,1] \to \mathbb{R}^2$ as below in Figure 3.3:

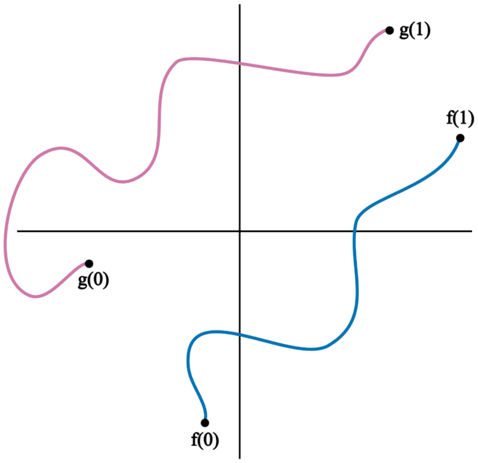


Figure 3.3: Continuous functions $f, g : [0,1] \to \mathbb{R}^2$

We call these **paths** in $\mathbb{R}^2$. Now let's say we want to "nudge" the path f to be the path g. Intuitively, this is something we can do, say if f and g represent strings on a table. To express this mathematically, we must write a function that parameterizes the paths themselves.

At time 0, the function must output the path f, and at time 1, the function must output the path g. All this must be done continuously, without skipping any space or breaking the strings, as below in Figure 3.4:

Perhaps this is easier to do than it sounds. In this case, the function is given by $F : [0,1]^2 \to \mathbb{R}^2$ where for all $(s,t) \in [0,1]^2$,

$$F(s,t) = f(s)(1-t) + g(s)t.$$

Observe that at time $t = 0$, $F(s,0)$ gives the path $f(s)$, and at time $t = 1$, $F(s,1)$ is the desired path $g(s)$. As F is a sum of products of continuous functions, F is itself continuous, so no breaking or skipping occurs.

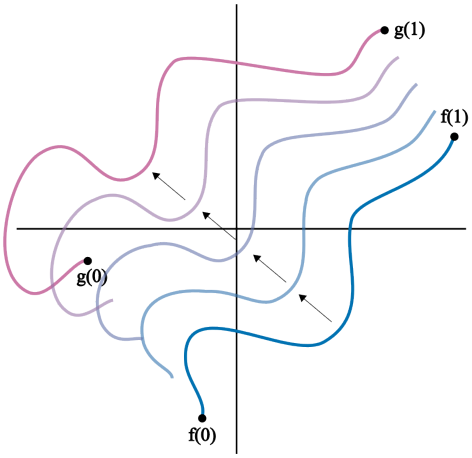


Figure 3.4: A homotopy from f to g.

Formal. Let $U \subset \mathbb{R}^n$, $V \subset \mathbb{R}^m$ be arbitrary subsets, and $f, g : U \to V$ continous functions. A **homotopy** from f to g is a continuous function

$$F : U \times [0,1] \to V$$

such that

$$F(x,0) = f(x),$$
$$F(x,1) = g(x).$$

When such a homotopy exists, we say f and g are **homotopic**.

To simply notation, we often denote $F(x,t) = F_t(x)$ for all $t \in [0,1]$, where F_t is a function $U \to V$.

Moving forward, we assume all functions are continuous.

Proposition 3.3.2. *Let $\sim$ denote the relation on functions $U \to V$ where $f \sim g$ if and only if there exists a homotopy from f to g. Then $\sim$ is an equivalence relation (Def. 3.2.27).*

Proof.

Informal. Homotopy behaves quite like equality, on many many levels. But it all starts with homotopy being an equivalence relation. Every function is homotopic to itself simply by a homotopy that keeps it still for all $t \in [0,1]$. Moreover, if f can be deformed to g by a homotopy, then it makes sense we can deform g back to f by reversing the motion. And, if we can deform f to g, then g to h, we can f deform to h in the same time frame just by performing both deformations twice as fast in succession.

Formal. We must prove $\sim$ is reflexive, symmetric, and transitive.

(i) To see $\sim$ is reflexive, let $f : U \to V$ be arbitrary. Then the function

$$F : U \times [0,1] \to V$$

defined by

$$F(x,t) = f(x)$$

is a homotopy from f to itself. Thus $f \sim f$.

(ii) To see $\sim$ is symmetric, suppose we have $f, g : U \to V$ such that $f \sim g$. That is, there is a homotopy

$$F : U \times [0,1] \to V.$$

Then we define

$$\overline{F} : U \times [0,1] \to V$$

by

$$\overline{F}(x,t) = F(x, 1-t).$$

Then

$$\begin{aligned} \overline{F}(x,0) &= F(x,1) = g(x), \\ \overline{F}(x,1) &= F(x,0) = f(x) \end{aligned}$$

shows $\overline{F}$ is indeed a homotopy from g to f, so $g \sim f$.

(iii) To see $\sim$ is transitive, suppose $f, g, h : U \to V$ satisfy $f \sim g$, $g \sim h$. Then there exist

$$F, G : U \times [0,1] \to V$$

where F is a homotopy from f to g and G is a homotopy from g to f. Then define

$$H : U \times [0,1] \to V$$

via

$$H(x,t) = \begin{cases} F(x,2t) & \text{if } 0 \le t \le \frac{1}{2}, \\ G(x,2t-1) & \text{if } \frac{1}{2} \le t \le 1. \end{cases}$$

One can check H is a homotopy from f to h, so $f \sim h$.

□

Since H is an equivalence relation, it is unambiguous to say f and g are homotopic if $f \sim g$.

Remark 3.3.3. For readers who know topology, the definition of homotopy extends to any maps $f : X \to Y$. The other definitions in this section will also have natural topological generalizations.

Example 3.3.4. As may be evident from the discussion in Definition 3.3.1, given any paths $f, g : [0,1] \to \mathbb{R}^2$, there is is a homotopy from f to g given by $F : [0,1]^2 \to \mathbb{R}^2$,

$$F(s,t) = f(s)(1-t) + g(s)t.$$

Example 3.3.5. The existence of homotopies depends on the choice of codomain. Given the paths $f, g : [0,1] \to \mathbb{R}^2 \setminus \{(0,0)\}$ defined by

$$\begin{aligned} f(t) &= (\sin(2\pi t), \cos(2\pi t)), \\ g(t) &= f(t) + (1,1), \end{aligned}$$

pictured in Figure 3.5 below, there is no function that can move f to g without breaking or skipping at the origin.

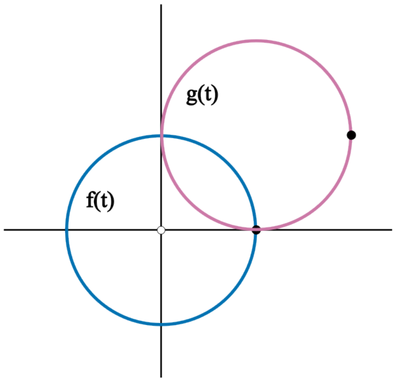


Figure 3.5: f and g are homotopic as paths in $\mathbb{R}^2$, but not as paths in $\mathbb{R}^2 \setminus \{(0,0)\}$. There is no way to move f to g without f at the origin.

Example 3.3.6. Here's a more interesting example. Let C denote the hollow cylinder $S^1 \times [0,1] \subset \mathbb{R}^3$. Let the functions $\mathrm{id} : C \to C$, $r : C \to C$ be defined by

$$\begin{aligned} \mathrm{id}(x,s) &= (x,s), \\ r(x,s) &= (x,0). \end{aligned}$$

They are both continuous. Observe that r is collapsing all of C onto the base circle $S^1 \times \{0\}$. We claim there is a homotopy from id to r.

Let $F : C \times [0,1] \to C$ be defined by

$$F((x,s),t) = (x, s(1-t)).$$

This function is continuous and indeed,

$$\begin{aligned} F((x,s),0) = (x,s) &= \mathrm{id}(x,s), \\ F((x,s),1) &= (x,0). \end{aligned}$$

This gives r a nice physical interpretation. We can think of id as representing the cylinder C in its original state, not changing the positions of any points. Then, to apply r, we squish C down to a circle in accordance to F until we reach r. So r can be interpreted squishing C down. This is pictured below in Figure 3.6.

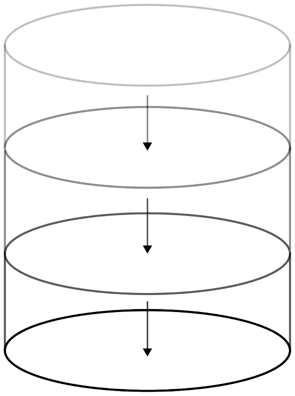


Figure 3.6: Physical interpretation of r from Example 3.3.6 via the homotopy from id to r. As time progresses from $t = 0$ to $t = 1$, we move the points of C in accordance to where they are mapped by f until we get r.

We will set up some more terminology.

Definition 3.3.7. Let $U \subset \mathbb{R}^n, V \subset \mathbb{R}^m$, and let $f : U \to V$ be a continuous function. If f is a continuous function with a continuous inverse, we say f is a **homeomorphism**. If there is a homeomorphism between U and V, we say U and V are **homeomorphic** and write $U \cong V$.

Like homotopy, homeomorphism defines an equivalence relation. Hence, it is unambiguous to say two spaces are homeomorphic.

Example 3.3.8. Let $U \subset \mathbb{R}^n$. The identity map $U \to U$ is continuous and is its own inverse. Hence, it is a homeomorphism. We say $U \cong U$, or U is homeomorphic to itself.

Example 3.3.9. The function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^3$ is a homeomorphism because $f^{-1} : \mathbb{R} \to \mathbb{R}$, $f^{-1}(x) = x^{1/3}$ is continuous.

Example 3.3.10. The function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^2$ is not a homeomorphism because f is not injective, hence has no inverse $\mathbb{R} \to \mathbb{R}$.

Example 3.3.11. It is not true in general that if a function is continuous and invertible, that its inverse is continuous. For example define $f : [0, 1) \to S^1$ via

$$f(x) = (\cos(2\pi x), \sin(2\pi x))$$

(pictured in Figure 3.7).

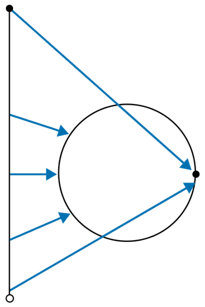


Figure 3.7: f of Example 3.3.11 is something like this. The interval $[0, 1)$ maps bijectively and continuously onto S^1, but reversing this function will break S^1.

Observe f wraps the interval $[0, 1)$ into a circle. This is continuous and bijective, hence invertible. However, the inverse requires the circle to taken to the interval $[0, 1)$. This breaks the circle, hence is not continuous.

Definition 3.3.12. Let $f : U \to V$ be a continuous function, and let $W \subset V$ be the image of f. If the function $f : U \to W$ given by f is a homeomorphism, we call f an **embedding**.

Example 3.3.13. Embeddings are not as restrictive as homeomorphisms, but give more well-behaved functions than with just continuity. For example, the curve in Figure 3.8 is continuous, but not an embedding. Likewise, the curves in Figure 3.5 are not embeddings (though they can be rewritten as embeddings $S^1 \to \mathbb{R}^2$).

Meanwhile, the curves of Figure 3.3 are embeddings.

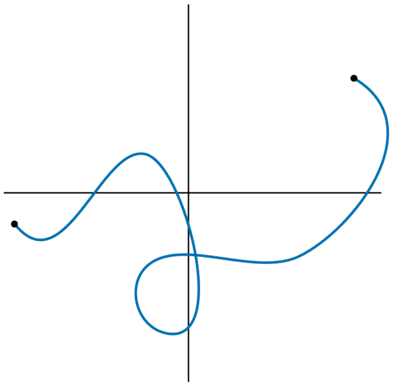


Figure 3.8: Continuous, but not embedded path $[0, 1] \to \mathbb{R}^2$.

This allows us to define a stronger type of homotopy.

Definition 3.3.14. Let $U \subset \mathbb{R}^n$, $V \subset \mathbb{R}^m$, and $f, g : U \to V$ be embeddings. Then a homotopy $F : U \times I \to V$ from f to g is an **isotopy** if $F(s, t)$ is an embedding for all fixed $t \in [0, 1]$. If there is an isotopy between two embeddings, then we say they are **isotopic**.

In short, an isotopy is a homotopy that is particularly nice.

Like homotopy and homeomorphisms, isotopy defines an equivalence relation. Hence it is unambiguous to say two embeddings are isotopic.

3.4 Introduction to Braids

We are finally ready to study braids. We will see quickly that this study puts to good use the accumulated knowledge of the previous sections.

Definition 3.4.1 (Braid [MK99, p. 3], [FM12, p. 240])**.**

Informal. To construct a mathematical braid with n strands, pick n points on a disk in 3-dimensional space. Then pick a parallel disk with the same distinguished points. We obtain a **braid on** n **strands** by drawing non-intersecting lines between the points on the two planes. The lines are not allowed to go backward.

Formal. Fix a positive integer n. Let $p_1, \ldots, p_n$ be points in the 2-disk D^2. A **braid on** n **strands**, or **n**-braid, is a collection of n paths $f_i : [0, 1] \to D^2 \times [0, 1], 1 \leq i \leq n$, called **strands**, and a permutation $\overline{f} \in S_n$ such that each of the following holds:

(i) the strands $f_i([0,1])$ are disjoint,

(ii) $f_i(0) = (p_i, 0)$,

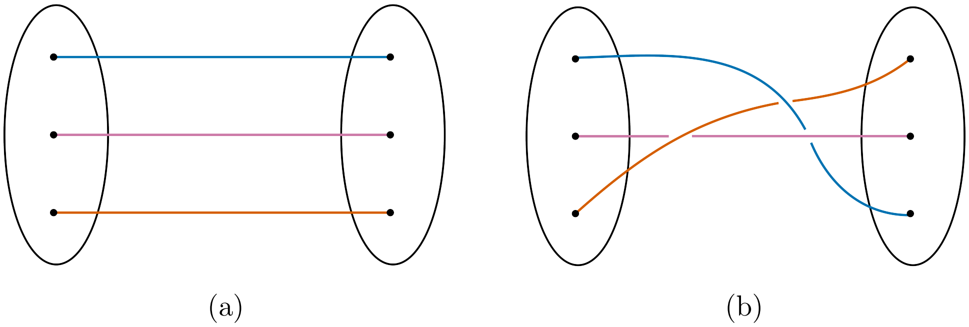
(iii) $f_i(1) = (p_{\overline{f}(i)}, 1)$,

(iv) $f_i(t) \in D^2 \times \{t\}$ for all $t \in [0,1]$.

Moving forward, we fix points $p_1, \dots, p_n$ for all positive integers n. We will assume every n-braid has these starting/ending points. Moreover, when we say $\beta = \{f_i : 1 \leq i \leq n\}$ is an n-braid, we will also use β to mean the subset of strands in $D^2 \times [0,1]$ given by β. Therefore, if we have a function f whose domain is $D^2 \times [0,1]$, the expression $f(\beta)$ makes sense.

To simplify notation, we will denote $\mathbb{D} = D^2 \times [0,1]$.

Example 3.4.2. In Figure 3.9, we have two examples of braids on 3 strands, which we refer to as 3-braids. These braids are a subset of the cylinder formed by the parallel disks, or $\mathbb{D}$. Moving forward, we omit the disks in figures.



(a)

(b)

Figure 3.9: (a): On the left, we have what's technically a braid on 3 strands. We can describe this by the paths $f_i(t) = p_i$ for all $t \in [0,1]$. (b): On the right, a slightly more interesting braid on 3 strands.

Example 3.4.3. In Figure 3.10, we have three non-example of braids.

Definition 3.4.4 (Braid equivalence [MK99, p. 96])**.**

Informal. We want to say two braids are the same if we can pull the strands around so that they are equal. For example, we want to say the three braids in Figure 3.11 are the same. The rules are, strands cannot cross each other and the starting/ending points cannot change. They are like infinitely elastic rubber bands.

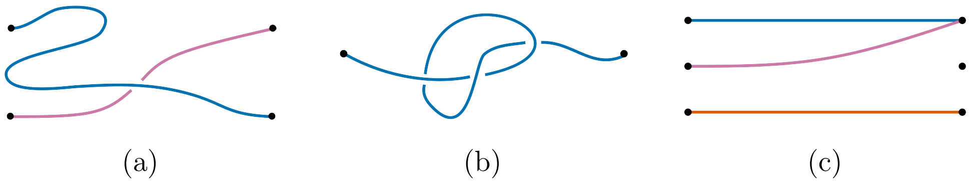


Figure 3.10: (a) and (b) are not braids because they turn back on theselves. (c) is not a braid because the points are not permuted.

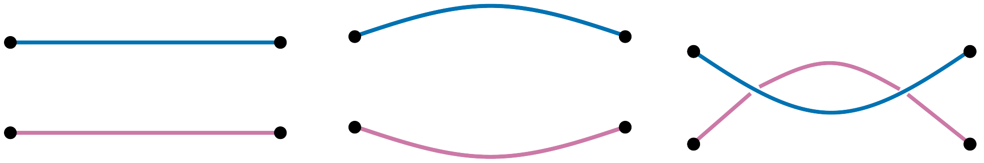


Figure 3.11: These three 2-braids are equivalent.

Formal. Denote $\mathbb{D} = D^2 \times [0,1]$. We declare two braids β and β' **equivalent** if there exists a continuous map

$$h : \mathbb{D} \times [0,1] \to \mathbb{D}$$

such that:

(i) for all $t \in [0,1]$, $h_t : \mathbb{D} \to \mathbb{D}$ is a homeomorphism,

(ii) for all $t \in [0,1]$, h_t is the identity on the cylindrical boundary of $\mathbb{D}$. That is, h_t fixes the boundary pointwise.

(iii) h_0 is the identity and $h_1(\beta) = \beta'$. We call h an **ambient isotopy**.

An ambient isotopy is even stronger than an isotopy. In the previous definition (Def. 3.4.4), we obtain an isotopy from β to β' by hitching a ride on an isotopy of homeomorphisms of the entire space $\mathbb{D}$. It is quite strict, but the ambient isotopy is equivalent to the intuitive notion of pulling and nudging the strands without breaking them or crossing them through each other.

Author comment. In truth, I do not know why an ambient isotopy is required as opposed to simply an isotopy of the map $\beta : \{p_1, \ldots, p_n\} \times I \to \mathbb{D}$ that fixes the endpoints. Although I understand how all knots can be isotoped to the unknot, I have not seen similar arguments for braids. If anyone has an explanation or counterexample, I would very grateful to hear it!

Remark 3.4.5. There is also a much more easy notion of braid equivalence formulated in terms in elementary moves. It takes no fancy math to understand, and the interested reader can find it in [MK99, p. 4].

Example 3.4.6. The braids in Figure 3.12 are equivalent.

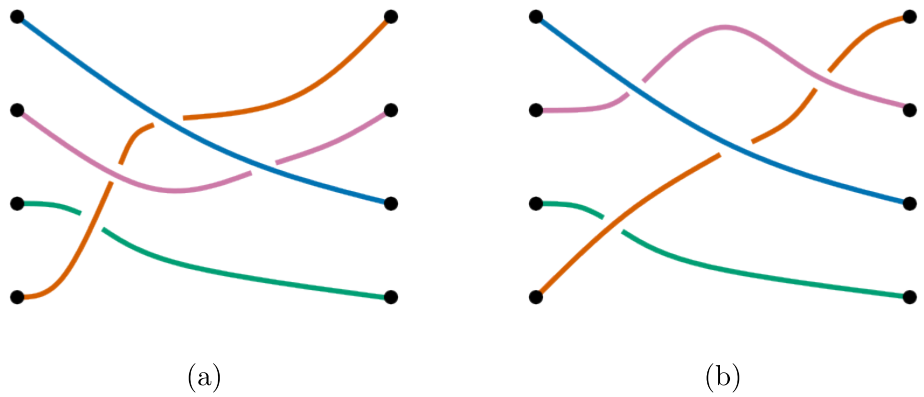


Figure 3.12: These 4-braids are equivalent. From (a), shift the pink line upward and nudge the orange line down to get (b).

Example 3.4.7. The braids in Figure 3.13 are not equivalent.

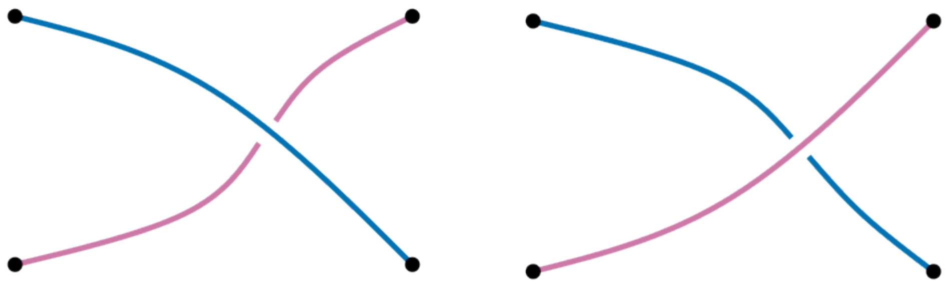


Figure 3.13: These 2-braids are not equivalent

The braid equivalence we have defined turns out to be an equivalence relation (Def. 3.2.27). But arguably this is what we would expect: after all, we *want* to partition big collections of braids into subsets of braids that are equivalent. As desired, the equivalence class of any braid ends up being the set of all braids equivalent to it.

From now on, when we say braid, we will actually be referring to the whole class of braids we have deemed equivalent. When we say, for example, that β is a braid, remember β is a representative of some class of equivalent braids.

Definition 3.4.8.

Informal. Now that we have put our intuition of equivalent braids into mathematics, we can see this structure is truly a very natural one, despite the complicated formalisms. We can now define a product on the set of braids that is nice enough to form a group structure.

The product of braids β and β' with the same number of strands is simply connecting β' to the end of β. Technically, braids must be parameterized by the interval $[0, 1]$, so the new braid will have to be reparameterized to travel the two braids twice as fast.

Therefore, we can now form a *braid group*, whose elements are equivalence classes of braids with the some fixed number of strands, and where the multiplication is concatenating the braids.

Formal. We define the **braid group on n strands B_n** to be the set of equivalence classes of braids where the product of any two braids $\{f_i : 1 \leq i \leq n\}$, $\{g_i : 1 \leq i \leq n\}$ is $\{h_i : 1 \leq i \leq n\}$ where for all $1 \leq i \leq n$,

$$h_i(t) = f_i(t) * g_i(t) = \begin{cases} f_i(2t) & \text{if } 0 \leq t \leq \frac{1}{2}, \\ g_{\overline{f}(i)}(2t-1) & \text{if } \frac{1}{2} \leq t \leq 1. \end{cases}$$

Indeed, we can verify the braid product $*$ is well-defined. That is, if we have β_i and β_i' are equivalent for $i = 1, 2$, then $\beta_1 * \beta_2$ and $\beta_1' * \beta_2'$ are equivalent: there is no ambiguity in choice of representatives. Moreover, we can verify $*$ is associative, that the braid given by $f_i(t) = p_i$ for all $1 \leq i \leq n$ gives an identity element in B_n, and that the mirror image of every braid is its inverse.

Every one of these properties can be seen by just drawing pictures. In the process, we would also realize that the notion of equivalence was essential here.

Example 3.4.9. In Figure 3.14, we see two examples of braid products. Observe in (a) that when we take the product of a braid with the identity, we can pull the braid back to the

None of these identities or inverses of the last example (Example 3.4.9) would have worked without the liberty to pull strands around. Associativity

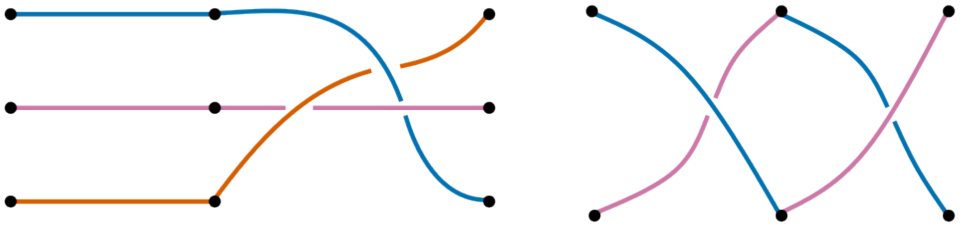


Figure 3.14: (a): On the left is the product of the 2 braids of Figure 3.9, denote it $e * \sigma$. Note that $e * \sigma = \sigma = \sigma * e$, since we can always pull the product strand back to σ. (b): On the right is the product of the braids of Figure 3.13. Note their product can be deformed to the identity, so they are inverses.

similarly relies on this as well, since different parenthesis placement results in the component braids taking up different shares of space in their product. If the author's shady reasoning is to be believed, then we well and truly have a group!

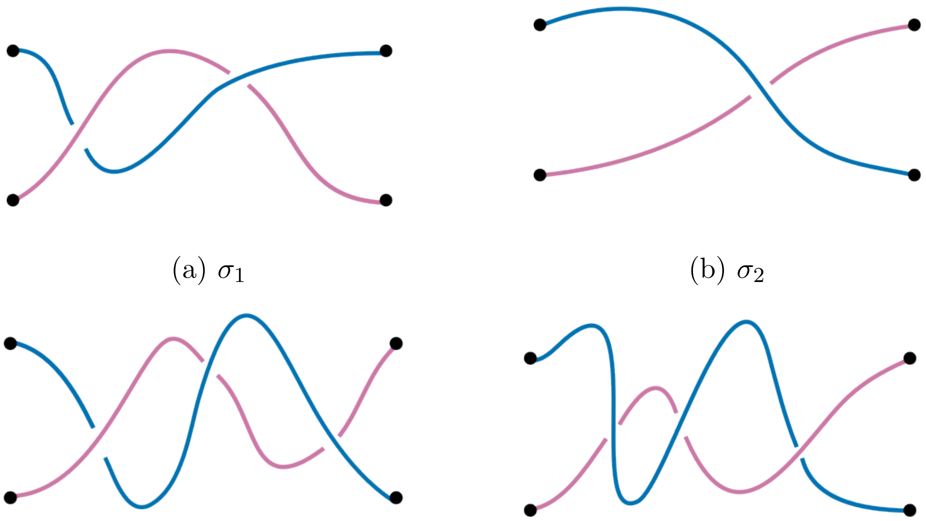
Example 3.4.10. Figure 3.15 showcases more braid multiplication in B_2, just for fun!

Example 3.4.11. The braid group on 1 strand has just one element: the braid given by just a single straight line. Therefore, $B_1 \cong \{e\}$ where $\{e\}$ is the trivial group (Def. 3.2.4).

One can show the braid group on 2 strands is isomorphic to $(\mathbb{Z}, +)$.

Example 3.4.12. One thing we can observe about B_n is that none of the braid groups for $n > 2$ are abelian. The strategy is simple: each braid $\{f_i : 1 \leq i \leq n\}$ has a corresponding permutation $\overline{f} \in S_n$. Then given two braids $\beta, \beta' \in B_n$ with permutations $\sigma, \sigma' \in S_n$, the permutation of $\beta * \beta'$ must be $\sigma' \circ \sigma$. Likewise, $\beta' * \beta$ must have permutation $\sigma \circ \sigma'$.

However, S_n is not abelian for $n > 2$, so B_n cannot be abelian for $n > 2$ as well. One neat way to see this is to recall the Dihedral group D_3 from Example 3.2.9, where from Figure 3.2, $F_3 \circ R \neq R \circ F_3$. But from the discussion in Definiton 3.2.11, elements of D_3 can be regarded as elements in S_3. Thus, take any braids $\beta, \beta' \in B_n$ that permute the first three points p_1, p_2, p_3 via R and F_3 respectively (we can prove these exist simply by drawing an example). Then it cannot possibly be that $\beta * \beta' = \beta' * \beta$.



(a) σ_1

(b) σ_2

(c) $\sigma_1 * \sigma_2 = \sigma_2^{-1}$ by pulling up the pink through and pushing down the blue crest.

(d) $\sigma_2 * \sigma_1 = \sigma_2^{-1}$ by pulling up the blue through and pushing down the pink crest.

Figure 3.15: Braids in B_2.

3.5 Mapping Class Groups

Now we have finally defined braid groups, but how do they relate to the symmetries of a disk? Like how we considered braids to be same up to "nudging," we now want to do the same for functions. Doing so will be a much more straightforward applications of the definitions from Section 3.3.

Definition 3.5.1. Let $U \subset \mathbb{R}^n$ be an arbitrary subset of $\mathbb{R}^n$, $V \subset U$ an arbitrary subset of U. We define

$$\text{Homeo}(U, V) = \{f : U \to U \mid f \text{ a homeomorphism such that } f(v) = v \text{ for all } v \in V\}.$$

In other words, $\text{Homeo}(U, V)$ is the set of all homeomorphisms of U to itself that fix V pointwise. Under composition, $\text{Homeo}(U, V)$ is a group.

Remark 3.5.2. As with homotopies, this definition extends to arbitrary topological spaces.

We will now use this group to define a relation.

Definition 3.5.3. Let $f, g \in \text{Homeo}(U, V)$ be arbitary. We say f and g are **isotopic in** $\text{Homeo}(U, V)$ if there is an isotopy

$$F : U \times [0, 1] \to U$$

from f to g such that for all $t \in [0, 1]$, $F_t \in \text{Homeo}(U, V)$.

Note that the use of isotopy rather than homotopy in the last definition (Def. 3.5.3) is unnecessary since the only allowed F_t are homeomorphisms anyway. Nonetheless, it is standard to say isotopy.

Isotopy in $\text{Homeo}(U, V)$ defines an equivalence relation, which gives a partition of the set $\text{Homeo}(U, V)$ into subsets of functions that are isotopic in $\text{Homeo}(U, V)$. Like with braids, we would like to regard isotopic functions as equivalent and work with the equivalence classes. But do the resulting equivalence classes form another group?

The answer lies in the subset

$$\text{Homeo}_0(U, V) = \{f \in \text{Homeo}(U, V) : f \text{ is isotopic to id in } \text{Homeo}(U, V)\}.$$

Proposition 3.5.4. $\text{Homeo}_0(U, V)$ *is a normal subgroup of* $\text{Homeo}(U, V)$.

Proof. First, we will show $\text{Homeo}_0(U, V)$ is a subgroup of $\text{Homeo}(U, V)$. Suppose $f, g \in \text{Homeo}_0(U, V)$. Then there exist isotopies $F, G : U \times [0, 1] \to U$ from f and g to the identity such that for all $t \in [0, 1]$, F_t, G_t are homeomorphisms that fix V pointwise. It follows that the function defined by

$$\begin{aligned} H : U \times [0, 1] &\to U, \\ H(x, t) &= F(G(x, t), t) \end{aligned}$$

is a homotopy from $f \circ g$ to id such that for all $t \in [0, 1]$, $v \in V$, $H_t(v) = v$. Moreover, the function defined by

$$\begin{aligned} J : U \times [0, 1] &\to U, \\ J(x, t) &= F_t^{-1}(x) \end{aligned}$$

is a homotopy from f^{-1} to id that also fixes V pointwise. It follows that $\text{Homeo}_0(U, V)$ is a subgroup.

Next, we will show $\text{Homeo}_0(U, V) \leq \text{Homeo}(U, V)$ is normal. Let $f \in \text{Homeo}_0(U, V)$, $g \in \text{Homeo}(U, V)$ be arbitrary. Then let $F : U \times [0, 1] \to U$ be an isotopy of f to the identity in $\text{Homeo}(U, V)$. It follows that the function defined by

$$\begin{aligned} H : U \times [0, 1] &\to U, \\ H(x, t) &= (g \circ F)(g^{-1}(x), t) \end{aligned}$$

is a homotopy from $g \circ f \circ g^{-1}$ to id in $\text{Homeo}(U, V)$. Thus, $g \circ f \circ g^{-1} \in \text{Homeo}_0(U, V)$, so $\text{Homeo}_0(U, V) \trianglelefteq \text{Homeo}(U, V)$. □

We will now see why the normality of $\text{Homeo}_0(U, V)$ is essential to creating our desired group.

Proposition 3.5.5. *Let $\sim$ be the relation on* $\text{Homeo}(U, V)$ *where for all $f, g \in$* $\text{Homeo}(U, V)$*, $f \sim g$ if and only if f and g are isotopic in* $\text{Homeo}(U, V)$*. Then the equivalence classes of $\sim$ are exactly the cosets* $\text{Homeo}_0(U, V)$ *in* $\text{Homeo}(U, V)$*.*

Proof. It is sufficient to show that given any $f \in \text{Homeo}(U, V)$,

$$[f]_\sim = f \,\text{Homeo}_0(U, V).$$

Suppose $g \in [f]_\sim$. Then g is isotopic to f in $\text{Homeo}(U, V)$. Let $F : U \times [0, 1] \to U$ be an isotopy from g to f in $\text{Homeo}(U, V)$. Then observe the function

$$H : U \times [0, 1] \to U,$$
$$H(x, t) = (f^{-1} \circ F)(x, t)$$

is an isotopy from $f^{-1} \circ g$ to id in $\text{Homeo}(U, V)$. This means that $f^{-1} \circ g \in \text{Homeo}_0(U, V)$. But then

$$g = f \circ (f^{-1} \circ g) \in f \,\text{Homeo}_0(U, V).$$

Conversely, suppose $g \in f \,\text{Homeo}_0(U, V)$. Then $g = f \circ h$ where $h \in \text{Homeo}_0(U, V)$. Let $F : U \times [0, 1] \to U$ be an isotopy from h to id in $\text{Homeo}_0(U, V)$. It follows that the function

$$H : U \times [0, 1] \to U,$$
$$H(x, t) = (f \circ F)(x, t)$$

is an isotopy from $f \circ h = g$ to f in $\text{Homeo}_0(U, V)$. Thus, $g \in [f]_\sim$. We conclude

$$[f]_\sim = f \,\text{Homeo}_0(U, V).$$

Therefore, the equivalence classes $[f]_\sim$ are the same as the cosets of $\text{Homeo}_0(U, V)$. □

This finally allows us to define our desired group of equivalence classes.

Definition 3.5.6 (Mapping class groups)**.**

Informal. Recall how in the last section (Section 3.4), we grouped together all braids that were reasonably similar and considered them as one object. When we formed the braid groups, we operated on these classes of objects, and in fact this was necessary to reveal the group structure.

In this case, the situation is not quite as difficult. $\mathrm{Homeo}(U, V)$ is already a group. When our functions in $\mathrm{Homeo}(U, V)$ can be reasonably nudged to be the same, we want to regard them as equivalent and operate on those equivalence classes themselves.

In short, we want to turn the set of equivalence classes $[f]_\sim$ into a group. In this case, normal subgroups come to our rescue. The desired equivalence classes are, as luck would have it, precisely the cosets of a normal subgroup. The resulting quotient group has elements $[f]_\sim$ and inherits our desired group structure. We call this the *mapping class group* $MCG(U, V)$ and call the equivalence classes $[f]_\sim$ *mapping classes.*

Formal. The **mapping class group** of U with respect to V is the quotient group

$$MCG(U, V) = \mathrm{Homeo}(U, V)/\,\mathrm{Homeo}_0(U, V).$$

We will denote mapping classes simply as $[f]$, dropping the $\sim$. Very often, even the brackets are dropped, but we will avoid doing so in this paper.

We will work out a relevant example that will soon become very important: the mapping class group of the disk D^2 with respect to its boundary circle $S^1 \subset D^2$.

Lemma 3.5.7 (Alexander trick [FM12, p. 47])**.** *The group $MCG(D^2, S^1)$ is trivial.*

Proof. Let $[f] \in MCG(D^2, S^1)$ be a mapping class. Then function $F : D^2 \times [0, 1] \to D^2$ defined by

$$F(x, t) = \begin{cases} (1-t)f\left(\frac{x}{1-t}\right) & \text{if } 0 \leq |x| < 1-t, \\ x & \text{if } 1-t \leq |x| \leq 1, \end{cases}$$

for all $t \in [0, 1]$ is an isotopy of f to id in $\mathrm{Homeo}(D^2, S^1)$. □

The idea of the Alexander trick is that for every $t \in [0, 1]$, we apply f on a smaller disk of radius $1 - t$. The factor of $1 - t$ inside of f allows us to map every point of D^2 for each t, while the factor of $1 - t$ outside shrinks the image to another disk of radius $1 - t$. While we shrink f to a homeomorphism of smaller and smaller disks, we switch out the remaining space with the identity map. See Figure 3.16 for a visual.

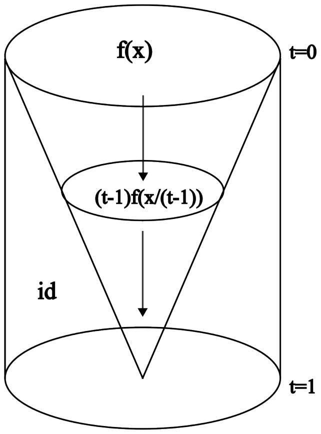


Figure 3.16: A standard visual of the Alexander trick by taking the function $(F(x,t),t)$ where F is the isotopy. In other words, we are taking what $F(x,t)$ is at each time and moving that slightly out of the way of what it was before. No information is lost and we get a nice visualization.

However, the key point is this: every homeomorphism of the disk, so long as it fixes the boundary S^1, is homotopic to the identity. What this means is, like how we could interpret the function r in Example 3.3.6 as squishing the cylinder, we can think of every boundary-fixing homeomorphism of the disk as some continuously pushing the points of the disk around. The pushing here is formally in accordance with the isotopy $F(x,1-t)$ from id to f, where F and f are defined as in Lemma 3.5.7.

Intuitively, this makes such homeomorphisms of a disk rather intuitive: it is as if the disk were made of soft clay, and we are kneading it around while ensuring the whole disk remains covered, and without moving the boundary or breaking it.

This is not true in general for continuous functions or homeomorphisms. For example, even the homeomorphism $f : D^2 \to D^2$ given by $f(x) = -x$ cannot be interepreted this way.

Remark 3.5.8. For those who are interested, there is a lot to learn about mapping class groups. In particular, a lot of work has been done on the mapping class groups of surfaces, where for an orientable surface S, we focus on

$$MCG(S) = \operatorname{Homeo}^+(S, \partial S)/\operatorname{Homeo}_0(S, \partial S),$$

the mapping classes of orientation-preserving homeomorphisms. For example, we know that for every compact orientable surface, possibly with fintely many marked points that must be permuted, $MCG(S)$ is finitely presented [FM12, p. 137], and we can write a presentation down. The same is true for compact non-orientable surfaces with finitely many marked points [Kor02].

For orientable surfaces, [FM12] is great (I think it's amazing!). For non-orientable surfaces, try [Par14].

3.6 Punctured Disk and Braids

The group $MCG(D^2, S^1)$ was the final piece we needed to formally connect braids to disks.

Theorem 3.6.1 ([FM12, p. 243]). *Let D_n denote $D^2 \setminus \{p_1, \ldots, p_n\}$. Assume none of the points p_i are on the boundary circle of D^2. Then*

$$B_n \cong MCG(D_n, S^1).$$

This is not very elementary to prove, but the intuition is much easier. First, we must state the following characterization of homeomorphisms $D_n \to D_n$.

Proposition 3.6.2. *Fix n points $p_1, \ldots, p_n \in D^2$ not on the boundary. Let $\mathrm{Homeo}_p(D^2, S^1)$ denote the subgroup of $\mathrm{Homeo}(D^2, S^1)$ that permutes the p_i's. In other words, for all $f \in \mathrm{Homeo}_p(D^2, S^1)$, there exists $\sigma \in S_n$ such that $f(p_i) = p_{\sigma(i)}$ for all $1 \leq i \leq n$. Then $\mathrm{Homeo}(D_n, S^1) \cong \mathrm{Homeo}_p(D^2, S^1)$.*

Proof.

Informal. The proof of this is a bit technical, but the idea is this. Whenever we have a homeomorphism of the punctured disk D_n, we can actually "fill in" the punctures and send them to each other. For any $f : D_n \to D_n$ a homeomorphism, f being a homeomorphisms ensures each of these punctures has one and only one other puncture they can go to. We find this by checking all the points around each puncture. By continuity, they must all be sent to an area roughly around another puncture. This allows us to extend functions of punctured disks to functions of the whole disk that simply permute the p_i's around.

This corresondence goes both ways. Given any homeomorphism that permutes the p_i's, we can define a new homeomorphism $D_n \to D_n$ just by restricting the domain to D_n.

The proof is on the technical side. A rigorous understanding of it is not required.

Formal. Let $f \in \mathrm{Homeo}(D_n, S^1)$, and pick mutually disjoint punctured neighborhoods U_i for each point p_i. We will define $\overline{f} \in S_n$ as follows. Since f is a homeomorphism, the image of each U_i must be another punctured neighborhood of some p_j, $1 \leq j \leq n$. Let $\overline{f}(i) = j$. Observe that $\overline{f}$ is indeed a bijection. If $i \neq j$, then the assumption $U_i \cap U_j = \varnothing$ implies $\overline{f}(U_i) \cap \overline{f}(U_j) = \varnothing$ because f is a homeomorphism. But all punctured neighborhoods of the same puncture must intersect. Thus, $\overline{f}(i) \neq \overline{f}(j)$. It follows $\overline{f}$ is injective, hence bijective.

Then defining $\widetilde{f} : D^2 \to D^2$ by

$$\widetilde{f}(x) = \begin{cases} f(x) & \text{if } x \in D_n \\ p_{\overline{f}(i)} & \text{if } x = p_i \end{cases},$$

we have a homeomorphism $\widetilde{f} \in \mathrm{Homeo}_p(D^2, S^1)$. This gives us a group homomorphism

$$\Phi : \mathrm{Homeo}(D_n, S^1) \to \mathrm{Homeo}_p(D^2, S^1),$$
$$\Phi(f) = \widetilde{f}.$$

We can define another homomorphism

$$\Psi : \mathrm{Homeo}_p(D^2, S^1) \to \mathrm{Homeo}(D_n, S^1)$$
$$\Psi(f) = f|_{D_n}$$

via restricting to D_n. Φ and Ψ are inverses, so we conclude

$$\mathrm{Homeo}(D_n, S^1) \cong \mathrm{Homeo}_p(D^2, S^1).$$

□

Therefore, whenever we consider a homeomorphism $D_n \to D_n$, we may just as well consider a homeomorphism $D^2 \to D^2$ that permutes the p_i's. This interpretation is quite useful — recall the discussion after the Alexander trick (Lemma 3.5.7). Any homeomorphism $D^2 \to D^2$ that fixes the boundary can be obtained by an isotopy in $\mathrm{Homeo}(D^2, S^1)$ from the identity.

This means that for any $f \in \mathrm{Homeo}(D_n, S^1)$, we can intuitively think of f as corresonding to some "kneading" of D^2. To get f, we continuously move the points of D^2 around, gradually moving each point $x \in D^2$ to where the point $f(x)$ used to be. At the end, remove the p_i's. Since we're thinking of deforming from the identity, this isotopy is the one from the Alexander trick, but backwards:

$$F(x, t) = \begin{cases} xf\left(\frac{x}{t}\right) & \text{if } 0 \leq |x| \leq t, \\ x & \text{if } t \leq |x| \leq 1. \end{cases}$$

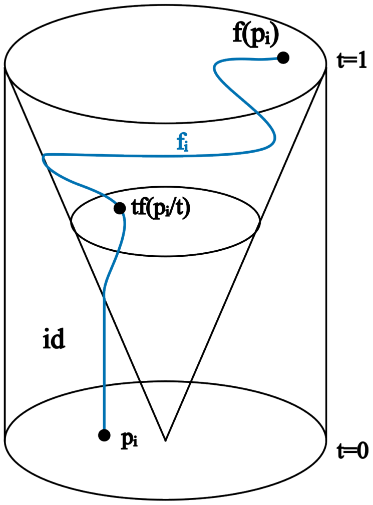


Figure 3.17: Observe how the Alexander lemma gives rise to the strands of a braid.

Really all we are doing is reversing where $t = 0$ and $t = 1$ are on Figure 3.16.

But in this process, pay special attention to how each point p_i traces out a path on D^2 to some $p_{\overline{f}(i)}$ as time moves from $t = 0$ to $t = 1$. There is the connection! For each p_i, we are getting a path

$$\gamma_i : [0, 1] \to D^2$$

defined by how the point p_i moves to $p_{\overline{f}(i)}$. We will now make one modification to the γ_i by defining

$$f_i : [0, 1] \to D^2 \times [0, 1],$$
$$f_i(t) = (\gamma_i(t), t).$$

In effect, the f_i's "raise" the paths outside of D^2, having them move forward as time passes. This is illustrated in Figure 3.17.

Observe that by definition, we have:

(i) the images $f_i[0, 1]$ are disjoint. This is because if for any time t, $f_i(t) = f_j(t)$ for some $i \neq j$, then this implies that p_i and p_j were moved to the same location in the "kneading" process. As we have covered, this is not allowed.

Physically, this means the paths we have defined never intersect each other.

(ii) $f_i(0) = (\gamma_i(0), 0) = (p_i, 0)$.

(iii) $f_i(1) = (\gamma_i(1), 1) = (p_{\overline{f}(i)}, 1)$.

(iv) For any $t \in [0,1]$, $f_i(t) = (\gamma_i(t), t) \in D^2 \times \{t\}$.

But recall from Definition 3.4.1 that these conditions are precisely what defines a braid. Thus, from an arbitrary homeomorphism $f : D_n \to D_n$, we have obtained $\{f_i : 1 \leq i \leq n\}$, a braid on n strands! Let's call $\{f_i : 1 \leq i \leq n\} = \beta_f$. But does this association still make sense for mapping classes?

Let's at least verify first that isotopic functions in $\text{Homeo}(D_n, S^1)$ give equivalent braids.

Proposition 3.6.3. *The function*

$$\Phi : MCG(D_n, S^1) \to B_n$$

defined by

$$\Phi([f]) = \{f_i : 1 \leq i \leq n\} = \beta_f$$

as in the preceding discussion is well-defined.

Proof.

Informal. We wish to show that Φ makes sense. Of course, given an arbitrary homeomorphism in $\text{Homeo}(D_n, S^1)$, it makes sense by our discussion that we just get the braid β_f. But for a quotient group like $MCG(D_n, S^1)$, if we take an arbitrary mapping class $[f]$, do we get β_f regardless of what representative we choose. It is not immediately clear that if $g \in [f]$, β_g is equivalent to β_f.

The way we do it here is by working explicitly with the braid formulas for f and g given by the Alexander trick, then extending the isotopy H from f to g to the entire braid. Visually, we are moving the whole braid from f to g in accordance with H. A bit more technical work is done to make this an ambient isotopy.

Ostensibly, this is the first step of a proof that $MCG(D_n, S^1) \cong B_n$. Using this strategy, we would want to show that Φ is also a homomorphism, and in fact bijective (technically, Φ is not a homomorphism in its current state, but we address this later). However, this is not how this theorem is conventionally proven. The proof is not easy, and we use much more powerful tools to tackle it. The following argument is included just to make the theorem seem more convincing using only elementary arguments. It may also demonstrate why continuing in this manner is not very sustainable.

Formal. Suppose $f \sim g$ in $\mathrm{Homeo}(D_n, S^1)$. Then let $F, G : \mathbb{D} \to D^2$ be the isotopies from id to F and G respectively given by the Alexander trick. We use these to define $\widetilde{F}, \widetilde{G} : \mathbb{D} \to \mathbb{D}$,

$$\begin{aligned}\widetilde{F}(x,s) &= (F(x,s), s),\\ \widetilde{G}(x,s) &= (G(x,s), s).\end{aligned}$$

Since $f \sim g$, let $H : \mathbb{D} \to \mathbb{D}$ be an isotopy from id to $g \circ f^{-1}$ (given by filling in the missing points of the corresponding isotopy in $\mathrm{Homeo}(D_n, S^1)$). Then $h : \mathbb{D} \times [0,1] \to \mathbb{D}$ defined by

$$h((x,s),t) = \begin{cases} \left(sH\left(\frac{x}{s}, t\right), s\right) & \text{if} 0 \leq |x| \leq s, \\ (x,s) & \text{if } s \leq |x| \leq 1. \end{cases}$$

satisfies

(i) for all $t \in [0,1]$, $h_t : \mathbb{D} \to \mathbb{D}$ is a homeomorphism,

(ii) for all $t \in [0,1]$, h_t fixes the boundary of $\mathbb{D}$,

(iii) $h_0(x,s) = (x,s)$ is the identity and

$$\begin{aligned} h_1(f_i(s)) = h_1(F(x,s),s) &= \begin{cases} \left(s(g \circ f^{-1})\left(\frac{F(p_i,s)}{s}\right), s\right) & \text{if} 0 \leq |F(p_i,s)| \leq s, \\ (F(p_i,s),s) & \text{if } s \leq |F(p_i,s)| \leq 1 \end{cases} \\ &= \begin{cases} \left(s(g \circ f^{-1})\left(\frac{sf\left(\frac{p_i}{s}\right)}{s}\right), s\right) & \text{if } 0 \leq |p_i| \leq s, \\ (p_i,s) & \text{if } s \leq |p_i| \leq 1 \end{cases} \\ &= \begin{cases} \left(sg\left(\frac{p_i}{s}\right), s\right) & \text{if } 0 \leq |p_i| \leq s, \\ (p_i,s) & \text{if } s \leq |p_i| \leq 1 \end{cases} \\ &= \widetilde{G}(p_i,s) = g_i(s). \end{aligned}$$

It follows that h is an ambient isotopy taking β_f to β_g, so β_f and β_g are the same braid. Thus, Φ is independent of our choice of mapping class representative and therefore well-defined.

□

In the discussion preceding the proof of Proposition 3.6.3, we briefly mentioned Φ is not technically a homomorphism in its current state. This is because it's backwards: $\Phi([f] \circ [g]) = \beta_g * \beta_f$. The reason is that $f \circ g$ is the function that applies g first, then f; meanwhile $\beta_f * \beta_g$ is first the braid given by f, then the braid given by g.

The fix for this is easy. We just redefine composition in MCG to go backwards. We will say $(f \circ g)(x) = g(f(x))$. This is not too unreasonable: if $\beta * \beta'$ is first β, then β', then why should $f \circ g$ be g first, then f? In some texts, this is the convention, and it ultimately makes no difference to the group structure (they are isomorphic). We will make this adjustment moving forward (I did not introduce it this way because it confuses me a lot).

The proof Φ is a homomorphism, and indeed an isomorphism, is much more involved. We will omit it in lieu of a more intuitive discussion.

Author comment. I am not actually aware if there is an elementary proof in this style. Probably there is a way to at least show Φ is a homomorphism, but I have not found or constructed an elementary argument for it using just ambient isotopy. If anyone has or knows of one, I would love to know it!

Ultimately, what does this isomorphism tell us about the relationship between braids and homeomorphisms of the punctured disk?

(i) First, the very fact Φ is a homomorphism tells us that if we deform the disk in accordance to f, then in accordance to g, that very process is no different (up to isotopy by homeomorphisms) to deforming the disk in accordance to $f \circ g$.

(ii) Second, that Φ is a bijective tells us that every braid arises from a mapping class of homeomorphisms in $\text{Homeo}(D_n, S^1)$. Likewise, every mapping class of homeomorphisms in $\text{Homeo}(D_n, S^1)$ uniquely gives a braid — any other mapping class will give a different braid.

(iii) Third, from the correspondence of mapping classes and braids, we see that kneading around a homeomorphism in $\text{Homeo}(D_n, S^1)$ is essentially no different from nudging around a braid on n strands.

This brings us to our final point: by endowing group structures to our objects — homeomorphisms and braids — we are able to speak precisely about the correspondence of structures between homeomorphisms and braids. Braids and mapping classes in $MCG(D_n, S^1)$ correspond one-to-one, and the natural operations on each correspond perfectly as well. Up to multiplying, these are exactly the same.

Besides just being interesting, this correspondence is quite useful for talking about mapping classes in general. For example, while homeomorphisms can be rather difficult to write down explcitly, in the case of the disk, we can simply specify a braid rather than write down an explicit formula. Writing down and composing functions can all be substituted by drawing braids. Or we can tell if two functions are the same by examining their braids. and when we compose functions, we need only multiply their braids.

This is not limited to the disk. “Surfaces,” such as the sphere, torus, klein bottle, etc. all contain disks. If a torus has some punctures in it, we can cut out a disk, apply braids, then glue that disk back to get a mapping class of the torus (we call this a *half dehn twist*).

This means, remarkably, that braids do not just describe the mapping classes of a punctured disk, but in fact of all shapes that have disks inside of them. All because we realized braids can be multiplied.

3.7 The proof?

The main storyline of this paper is over, but probably there are readers interested in the proof that $MCG(D_n, S^1) \cong B_n$.

The techniques are quite beyond the scope of this paper, but they are worth learning. They are accessible after a typical first course in algebraic topology.

The braid group on n strands is the same as the fundamental group of the unordered configuration space $C(D^2, n)$. Then we use the generalized *Birman exact sequence*, where for any orientable surface S, $S_n = S \backslash \{p_1, \ldots, p_n\}$, there is an exact sequence

$$1 \to \pi_1(C(S,n)) \xrightarrow{\text{Push}} MCG(S_n, \partial S_n) \xrightarrow{\text{Forget}} MCG(S, \partial S) \to 1.$$

Roughly, we define Push by taking *Dehn twists* around the loops in $C(S,n)$, Forget is the natural map where we fill in the punctures, and the exact sequence is obtained by the long exact sequence of homotopy groups of the fiber bundle (the LES of a fiber bundle is covered in Chapter 4 of [Hat02]),

$$\text{Homeo}^+(S_n, \partial S_n) \to \text{Homeo}^+(S, \partial S) \to C(\text{int}(S), n).$$

Of course, since $MCG(D^2, S^1)$ is trivial, we get

$$B_n \cong \pi_1(C(D^2, n)) \cong MCG(D^2, S^1).$$

The full exposition is covered in detail in [FM12].

Bibliography

[DF04] David S. Dummit and Richard M. Foote. *Abstract Algebra.* John Wiley & Sons, Hoboken, NJ, 3 edition, 2004.

[FM12] Benson Farb and Dan Margalit. *A Primer on Mapping Class Groups.* Princeton Mathematical Series. Princeton University Press, 2012.

[Hat02] Allen Hatcher. *Algebraic Topology.* Cambridge University Press, 2002.

[Kor02] Mustafa Korkmaz. Mapping class groups of nonorientable surfaces. *Geometriae Dedicata*, 89(1):107—-131, 2002.

[MK99] Kunio Murasugi and Bohdan I. Kurpita. *A Study of Braids.* Mathematics and Its Applications. Springer, 1 edition, 1999.

[Par14] Luis Paris. Mapping class groups of non-orientable surfaces for beginners, 2014.

4. An Introduction to the Theory of p-Adic Numbers and Local Fields

—*Zachariah Zobair*—

Abstract

First introduced by Kurt Hensel in 1897, for a fixed prime p, the p-adic numbers are a completion of the rational numbers, similar to the real numbers, but with respect to a different sense of "distance" between two rational numbers. They are an example of a nonarchimedean local field. Here we will begin by constructing the p-adic numbers and exploring some of their analytic, algebraic, and topological properties, as well as remarking on how these properties manifest for general local fields. From there, we will shift perspectives and provide a geometric intuition as to how, if one's goal is to solve a Diophantine equation, the p-adic numbers naturally arise as a correct setting in which to do this.

4.1 Introduction

Consider the field of rational numbers, $\mathbb{Q}$. The rational numbers come equipped with a natural notion of "distance", given by the absolute value $|\cdot|$. That is, for two rational numbers x and y, their distance is given by $|x-y|$. More precisely, $|\cdot|$ is an example of a *multiplicative valuation.*

Definition 4.1.1. A *multiplicative valuation* on a field K is a a function $|\cdot|\colon K \to \mathbb{R}$ such that

1. $|x| = 0$ if and only if $x = 0$, and $|x| > 0$ otherwise,
2. $|xy| = |x||y|$, and
3. $|x+y| \leq |x| + |y|$ (the triangle inequality).

The second condition in the definition above shows why these valuations are called multiplicative. One readily checks (and probably knows by intuition) that the absolute value on $\mathbb{Q}$ satisfies all these properties. If the reader has taken a course in real analysis, then they know that $\mathbb{Q}$ is not *complete* with respect to the absolute value. That is, there exist Cauchy sequences (read: sequences in which the entries of the sequence become arbitrarily close as one goes further out in the sequence) that do not converge in $\mathbb{Q}$. As an example, take the sequence

$$(3, 3.1, 3.14, 3.141, 3.1415, 3.14159, \ldots).$$

We can see that the entries here (which are all in $\mathbb{Q}$) are getting arbitrarily close to π, yet since $\pi \notin \mathbb{Q}$, this has no hope of converging to an element in $\mathbb{Q}$. A definition of the real numbers, $\mathbb{R}$, is the *completion* of $\mathbb{Q}$ with respect to the absolute value $|\cdot|$. That is, we essentially fill in all these "gaps" in $\mathbb{Q}$ where we expect a Cauchy sequence to converge.

Perhaps unsurprisingly, the standard absolute value $|\cdot|$ on $\mathbb{Q}$ is not the only function $\mathbb{Q} \to \mathbb{R}$ satisfying the requirements of definition 4.1.1.

Definition 4.1.2. For a fixed prime p, the *p-adic* absolute value on $\mathbb{Q}$, denoted $|\cdot|_p$, is a function $|\cdot|_p \colon \mathbb{Q} \to \mathbb{R}$ defined via

$$|x|_p = \begin{cases} 0 & \text{if } x = 0, \\ \frac{1}{p^{n-m}} & \text{if } x \neq 0, \end{cases}$$

where, writing $x = a/b$ for $a, b \in \mathbb{Z}$, n is the highest power of p dividing a and m is the highest power of p dividing b.

Proposition 4.1.3. The p-adic absolute value is a multiplicative valuation on $\mathbb{Q}$.

We will take as a first working definition of the p-adic numbers, denoted by $\mathbb{Q}_p$, to be the completion of the rational numbers with respect to the absolute value $|\cdot|_p$.

4.2 The Structure of $\mathbb{Q}_p$

The p-adic absolute value satisfies a condition stronger than the triangle inequality. One can check that for any $x, y \in \mathbb{Q}$, we have

$$|x + y|_p \leq \max\{|x|_p, |y|_p\}.$$

This inequality is called the *strong triangle inequality*, or, sometimes, the *ultrametric property.* A multiplicative valuation satisfying the strong triangle inequality is said to be a *nonarchimedean valuation.* Otherwise we say it is *archimedean.*

Given a nonarchimedean multiplicative valuation $|\cdot|\colon K \to \mathbb{R}$, one can obtain an *additive valuation* $v\colon K \to \mathbb{R}\cup\{\infty\}$ by putting $v(x) = -\log|x|$ for $x \neq 0$, and adopting the convention that $v(x) = \infty$ for $x = 0$. Additive valuations share analogous properties to those given in definition 4.1.1:

Definition 4.2.1. An *additive valuation* on a field K is a function $v\colon K \to \mathbb{R}\cup\{\infty\}$ satisfying

1. $v(x) = \infty$ if and only if $x = 0$,
2. $v(xy) = v(x) + v(y)$,
3. $v(x + y) \geq \min\{v(x), v(y)\}$, with equality if $v(x) \neq v(y)$.

Remark 4.2.2. Some authors will take valuation to mean additive valuation and absolute value to mean multiplicative valuation. We will often adopt this terminology.

Of course, one can go the other way: Given an additive valuation v on K, put $|x| = q^{-v(x)}$ for some real number $q > 1$ to obtain a multiplicative valuation. This shift in perspective from multiplicative to additive valuations may seem unmotivated at this point, but later in this paper we will see that the additive framework is often nicer to work with. In fact we will almost exclusively use additive valuations.

One instance in which the multiplicative point of view is more natural is in defining the topology induced on K. An absolute value $|\cdot|$ on K defines a metric d on K. Given $x, y \in K$, their distance is then given by

$$d(x, y) = |x - y|.$$

Thus we can equip K with the metric topology from d. That is, basic open sets are given by the open balls

$$B(x, r) = \{y \in K \mid d(x, y) < r\}.$$

Two absolute values (and so valuations as well) are said to *equivalent* if they induce the same topology on K. We have the following characterizations of equivalence of absolute values.

Proposition 4.2.3. Suppose $|\cdot|_1$ and $|\cdot|_2$ are two absolute values on a field K. The following are equivalent:

1. $|\cdot|_1$ and $|\cdot|_2$ are equivalent.
2. There exists a real number $s > 0$ such that $|x|_1 = {|x|_2}^s$ for all $x \in K$.
3. For any $x \in K$, $|x|_1 < 1$ implies $|x|_2 < 1$.

It is a theorem of Ostrowski that every absolute value on $\mathbb{Q}$ is equivalent to either the standard absolute value $|\cdot|$ (often denoted $|\cdot|_\infty$) or $|\cdot|_p$ for a prime p. An equivalence class of an absolute value on a field is called a *place.*

Consider the formal power series

$$\sum_{n=m}^{\infty} a_n p^n, \quad a_n \in \{0, \dots, p-1\}, m \in \mathbb{Z}. \tag{$*$}$$

Then the sequence (s_k) defined by $s_k = \sum_{n=m}^{m+k} a_n p^n$ is Cauchy with respect to the p-adic absolute value:

$$\begin{aligned}\lim_{k\to\infty} |s_{k+1} - s_k|_p &= \lim_{k\to\infty} |a_{k+1}p^{k+1}|_p \\ &= \lim(\underbrace{|a_{k+1}|_p}_{\leq 1} \cdot |p^{k+1}|_p) \leq \lim_{k\to\infty} \frac{1}{p^{k+1}} = 0.\end{aligned}$$

Thus we see $(s_k) \to x \in \mathbb{Q}_p$. Therefore any power series of the form $(*)$ defines an element of $\mathbb{Q}_p$, and the unicity of limits yields that such a power series defines a unique element. Conversely, we will see *every* element of $\mathbb{Q}_p$ admits such a unique power series representation. Before that we need to develop some further theory on the algebraic structure of the field $\mathbb{Q}_p$.

For the sake of stating results in more generality, we will let (K, v) be a valued field with additive valuation $v \colon K \to \mathbb{R} \cup \{\infty\}$.

Definition 4.2.4. The *valuation group* of K is $v(K^\times) \subseteq \mathbb{R}$. If $v(K^\times) = s\mathbb{Z}$ for some least positive value $s \in v(K^\times)$ then we say v is a *discrete valuation.* If $s = 1$ then v is *normalized.*

Remark 4.2.5. If a discrete valuation v is not normalized, so that $v(K^\times) = s\mathbb{Z}$ for some $s \neq 1$, then we can always normalize the valuation by replacing v with $v' = \frac{1}{s}v$. By proposition 4.2.3, the resulting topology on K induced by v' is no different than that induced by v.

Definition 4.2.6. The *valuation ring* of K is

$$\mathcal{O} = \{x \in K \mid v(x) \geq 0\}.$$

It is an integral domain with $\text{frac}(\mathcal{O}) = K$. The *unit group* of $\mathcal{O}$ is $\mathcal{O}^\times = \{x \in K \mid v(x) = 0\}$.

Proposition 4.2.7. The valuation ring $\mathcal{O}$ is a ring and the subset $\mathfrak{p} = \mathcal{O} \setminus \mathcal{O}^\times = \{x \in K \mid v(x) > 0\}$ is its unique maximal ideal.

Proof. The fact that $\mathcal{O}$ is a ring is immediate by properties of the additive valuation. We show $x \in \mathcal{O}$ is a unit if and only if $v(x) = 0$. Suppose $x \neq 0$ such that $x^{-1} \in \mathcal{O}$. Then, $v(xx^{-1}) = v(x) + v(x^{-1}) = v(1) = 0$ (convince yourself from definitions that $v(1) = 0$). Hence $v(x) = -v(x^{-1})$. Thus we see $x, x^{-1} \in \mathcal{O}$ if and only if $v(x) = v(x^{-1}) = 0$.

It follows then that $\mathfrak{p}$ consists precisely of the noninvertible elements of $\mathcal{O}$. It is a standard result of commutative algebra then that $\mathfrak{p}$ is the unique maximal ideal of $\mathcal{O}$. □

Proposition 4.2.8. There exists $\varpi \in \mathcal{O}$ such that $\mathfrak{p} = (\varpi)$ and all ideals of $\mathcal{O}$ are of the form (ϖ^n) for $n \in \mathbb{Z}_{\geq 0}$. When v is normalized, one may take ϖ to be any element with $v(\varpi) = 1$.

Proof. Normalize v if necessary. Pick $\varpi \in \mathcal{O}$ such that $v(\varpi) = 1$. Let $\mathfrak{a} \subseteq \mathcal{O}$ be any ideal. Put $n = \min_{x \in \mathfrak{a}} v(x)$. Then we have $\varpi^{-n}\mathfrak{a} \subseteq \mathcal{O}$, as all valuations of elements in $\varpi^{-n}\mathfrak{a}$ are nonnegative. However, there exist elements of valuation 0 (which hence are units) in $\varpi^{-n}\mathfrak{a}$ and thus $\varpi^{-n}\mathfrak{a} = \mathcal{O}$. Then we have $\mathfrak{a} = \varpi^n\mathcal{O} = (\varpi^n)$. □

An element ϖ as in proposition 4.2.8 is said to be a *uniformizing element*, *prime element*, or a *uniformizer*. In particular, this shows that $\mathcal{O}$ is a principal ideal domain.

Proposition 4.2.9. Any nonzero element x of $\mathcal{O}$ may be uniquely written as $x = \varpi^n u$ for $u \in \mathcal{O}^\times$ and $n \in \mathbb{Z}_{\geq 0}$. A nonzero element x of K may be uniquely written $x = \varpi^n u$ for $n \in \mathbb{Z}$.

Definition 4.2.10. The field $k = \mathcal{O}/\mathfrak{p}$ is called the *residue field* of K.

Example 4.2.11. Take $K = \mathbb{Q}$ equipped with the p-adic valuation v_p. Then $v_p(\mathbb{Q}^\times) = \mathbb{Z}$ (so v_p is a normalized discrete valuation). We have

$$\mathcal{O} = \mathbb{Z}_{(p)} = \{r/s \in \mathbb{Q} \mid (r, s) = 1, p \nmid s\}.$$

The maximal ideal is then $\mathfrak{p} = p\mathbb{Z}_{(p)}$ and it has residue field $k = \mathbb{Z}_{(p)}/p\mathbb{Z}_{(p)} \cong \mathbb{Z}/p\mathbb{Z}$. For a uniformizing element we can take $\varpi = p$.

Proposition 4.2.12. For $n \geq 0$, there is an isomorphism

$$\mathfrak{p}^n/\mathfrak{p}^{n+1} \cong \mathcal{O}/\mathfrak{p} = k.$$

Proof. The isomorphism is given by $a\varpi^n \mapsto a \pmod{\mathfrak{p}}$. □

For a field K with valuation v, by $\widehat{K}$ we understand the completion of K with respect to the valuation v. The valuation v extends to one on $\widehat{K}$, which we will denote $\hat{v}$. We obtain this extended valuation as follows: For $a \in \widehat{K}$, choose a sequence (a_n) in K converging to a. Then put

$$\hat{v}(a) = \lim_{n\to\infty} v(a_n).$$

For why this limit exists, see ([NS13], Ch. II §4).

Proposition 4.2.13. Let $\mathcal{O} \subseteq K$ (resp. $\widehat{\mathcal{O}} \subseteq \widehat{K}$) be the valuation ring of discrete valuation v (resp. $\widehat{v}$), and let $\mathfrak{p}$ (resp. $\widehat{\mathfrak{p}}$) be its maximal ideal. Then, for $n \geq 1$, one has

$$\widehat{\mathcal{O}}/\hat{\mathfrak{p}}^n \cong \mathcal{O}/\mathfrak{p}^n.$$

Proof. Consider the map $\mathcal{O} \to \widehat{\mathcal{O}}/\widehat{\mathfrak{p}}^n$ via $a \mapsto a \pmod{\hat{\mathfrak{p}}^n}$. The kernel of this map is $\mathfrak{p}^n$, so it induces an injection of $\mathcal{O}/\mathfrak{p}^n$ into $\widehat{\mathcal{O}}/\hat{\mathfrak{p}}^n$. For surjectivity, observe that for any $x \in \widehat{\mathcal{O}}$, there exists $a \in \mathcal{O}$ such that $\hat{v}(x-a) \geq n$. In other words, $a \equiv x \pmod{\widehat{\mathfrak{p}}^n}$. □

Proposition 4.2.14. Let $S \subseteq \mathcal{O}$ be a system of representatives for $k = \mathcal{O}/\mathfrak{p}$ such that $0 \in S$ and let $\varpi \in \mathcal{O}$ be a uniformizer. Then every nonzero $x \in \widehat{K}$ admits a unique representation as a convergent series

$$x = \varpi^m(a_0 + a_1\varpi + a_2\varpi^2 + \cdots)$$

where $a_i \in S$, $a_0 \neq 0$, and $m \in \mathbb{Z}$.

Proof. By proposition 4.2.9, we may write $x = \varpi^m u$ for some $u \in \widehat{\mathcal{O}}^\times$, $m \in \mathbb{Z}$. Then from proposition 4.2.13, the residue class of $u \pmod{\widehat{\mathfrak{p}}}$ has a unique representative $a_0 \in S$, with $a_0 \neq 0$. Thus we may write

$$u = a_0 + \varpi b_1$$

for some $b_1 \in \widehat{\mathcal{O}}$. Repeating this process yields

$$u = a_0 + a_1\varpi + \cdots + \varpi^{n+1}b_{n+1},$$

with $b_{n+1} \in \widehat{\mathcal{O}}$. Then we get

$$u = \sum_{n=0}^{\infty} a_n \varpi^n,$$

since, if s_n denotes the nth partial sum of the above series, we have

$$\lim_{n\to\infty} |u - s_n| \leq \lim_{n\to\infty} |\varpi^n b_{n+1}| = 0.$$

□

Since $\mathbb{Q}_p$ is the completion of $\mathbb{Q}$ with respect to v_p, by way of proposition 4.2.14 we see that we can write

$$\mathbb{Q}_p = \left\{ \sum_{n=m}^{\infty} a_n p^n \mid a_n \in \mathbb{F}_p, m \in \mathbb{Z} \right\}.$$

We denote the valuation ring of $\mathbb{Q}_p$ as $\mathbb{Z}_p$. These are the *p-adic integers*. By proposition 4.2.13, the residue field of $\mathbb{Z}_p$ is $\mathbb{F}_p$. A characterization of $\mathbb{Z}_p$ can be given as

$$\mathbb{Z}_p = \left\{ \sum_{n=0}^{\infty} a_n p^n \mid a_n \in \mathbb{F}_p \right\}.$$

4.3 Some Topology & a Shift in Perspective

The ultrametric quality of a nonarchimedean valuation yields an unfamiliar topology on a nonarchimdean field K. Some of the peculiarities are collected in the following proposition:

Proposition 4.3.1. Let K be a complete nonarchimedean field.

1. Any point of an open ball is a center of that open ball.

2. Two open balls are either disjoint, or one is contained in the other.

3. Every ball is both open and closed, and all balls are homeomorphic to each other.

This amounts to saying that K is *totally disconnected*, topologically[1]. Note that, by definition, $\mathcal{O} = B(0, 1)$, the open (closed) ball centered at 0 of

[1] Henceforth, unless stated otherwise, K is assumed to be complete with respect to its valuation, so we drop the $\widehat{K}$ notation.

radius 1. Similarly, the ideals $\mathfrak{p}^n$ are the balls $B(0, \frac{1}{q^{n-1}})$. For $n \geq 1$, there are canonical homomorphisms $\mathcal{O} \to \mathcal{O}/\mathfrak{p}^n$, given by reduction modulo $\mathfrak{p}^n$. Similarly we have reduction maps

$$\mathcal{O}/\mathfrak{p} \xleftarrow{\lambda_1} \mathcal{O}/\mathfrak{p}^2 \xleftarrow{\lambda_2} \mathcal{O}/\mathfrak{p}^3 \xleftarrow{\lambda_3} \cdots \tag{$**$}$$

Thus by universal properties we get a canonical map

$$\mathcal{O} \to \varprojlim \mathcal{O}/\mathfrak{p}^n \subseteq \prod_{n \in \mathbb{Z}_{\geq 0}} \mathcal{O}/\mathfrak{p}^n$$

to the *projective limit* $\varprojlim \mathcal{O}/\mathfrak{p}^n$ of the projective system $(**)$. For a review of projective limits, see ([Lan11], Ch. I §10). Endowing $\mathcal{O}/\mathfrak{p}^n$ with the discrete topology, then $\prod_{n \in \mathbb{Z}_{\geq 0}} \mathcal{O}/\mathfrak{p}^n$ is given the product topology, from which $\varprojlim \mathcal{O}/\mathfrak{p}^n$ inherits its topology.

Proposition 4.3.2. The canonical map $\mathcal{O} \to \varprojlim \mathcal{O}/\mathfrak{p}^n$ is an isomorphism of rings and of topological spaces.

Proof. See ([NS13], Ch. 2 §4). □

In particular, we have $\mathbb{Z}_p = \varprojlim \mathbb{Z}/p^n\mathbb{Z}$. This shift to viewing $\mathcal{O}$ as a projective limit may initially come off as abstract, but it is a very natural way to understand these objects. In fact, we more often than not take the projective limit as the *definition* of $\mathbb{Z}_p$.

When $\mathcal{O}/\mathfrak{p} = k$ is finite, then so too is $\mathcal{O}/\mathfrak{p}^n$ for all n. A standard result of topology (Tychonoff's theorem) then yields that $\varprojlim O/\mathfrak{p}^n = \mathcal{O}$ is compact. Then, by the earlier discussion on the fact that $\mathcal{O}$ and the $\mathfrak{p}^n$ are open balls of arbitrarily small radii in K, we conclude that if K has a finite residue field it is *locally compact*. What's more is that the converse is true: For a (complete) field K with discrete valuation v to be locally compact is is necessary and sufficient that its residue field k be a finite field (cf. [SG13], Ch. 2 §1 prop. 1).

Definition 4.3.3. A *nonarchimedean local field* is a field complete with respect to a discrete valuation with a finite residue field.

We say that a local field K is of *mixed characteristic* if $\text{char}(K) \neq \text{char}(k)$. Otherwise it is of *equal characteristic*.

4.4 Local to Global Principles

One of the primary reasons we study local fields is because of so-called "local to global" principles. The philosophy is that if we want to answer some global problem, then we can solve it "locally everywhere" and hopefully the data of local solutions can be stitched together in some way to recover a global solution. This being a useful technique is of course contingent on the fact that solving the problem locally should be somehow simpler than solving it globally. We should have techniques to get our hands on solutions in the local setting.

Perhaps unsurprisingly, if our problem is number theoretic (solving a Diophantine equation) in nature, then local fields are the correct "local" setting to work[2]. One of the most essential techniques we have in solving equations over local fields is the following.

Proposition 4.4.1 (Hensel's Lemma)**.** Let $f \in \mathcal{O}[x]$ be a monic polynomial and suppose there exists $a \in \mathcal{O}$ such that $f(a) \equiv 0 \pmod{\varpi}$ and $f'(a) \not\equiv 0 \pmod{\varpi}$ (the derivative being taken formally). Then there exists a unique lift $\alpha \in \mathcal{O}$ such that $f(\alpha) = 0$ and $\alpha \equiv a \pmod{\varpi}$.

Proof. See ([Gui18], Ch. 2, thm. 21). In fact, a stronger version of Hensel's lemma is proven here. □

Hensel's lemma says that given monic integral polynomial over a local field K, if we can find a simple root modulo ϖ, then we can lift that to a root of the polynomial. Furthermore, the proof of Hensel's lemma is constructive and so gives a way to actually compute the roots of polynomials.

Example 4.4.2 (Roots of Unity)**.** Consider $f = x^{p-1} - 1 \in \mathbb{Z}_p[x]$. The nonzero elements of $\mathbb{F}_p$ are precisely the roots of f when reduced mod p, and so f splits completely in $\mathbb{F}_p[x]$. Thus we may repeatedly apply Hensel's lemma to conclude that f has $p-1$ many distinct roots in $\mathbb{Z}_p$. In other words, $\mathbb{Q}_p$ contains the $(p-1)$st roots of unity.

A *quadratic form in n variables* over $\mathbb{Q}$ is (loosely) a polynomial of the form $\sum_{i,j} a_{ij}x_ix_j$, where $i, j \in \{1, \ldots, n\}$ and $a_{ij} \in \mathbb{Q}$. We say a quadratic form Q *represents 0* if there is a nontrivial solution to $Q(x_1, \ldots, x_n) = 0$. One of the most famous local-global principles in number theory concerns when quadratic forms over $\mathbb{Q}$ represent 0.

[2] We will justify this statement more later

Theorem 4.4.3 (Hasse-Minkowski)**.** *A quadratic form Q over $\mathbb{Q}$ represents 0 if and only if Q represents 0 over $\mathbb{Q}_p$ for all p (including $p = \infty$).*

Proof. See ([Ser73], Ch. IV §3). □

Remark 4.4.4. The Hasse-Minkowski theorem really *only* concerns quadratic forms. It immediately fails for ternary forms. Consider the equation

$$3x^3 + 4y^3 + 5z^3 = 0.$$

One can check that the equation has nonzero solutions in $\mathbb{Q}_p$ for every prime p (using Hensel lifting) and in $\mathbb{R}$. However, it has no nonzero solutions in $\mathbb{Q}$.

Earlier we said that "local fields are the correct 'local' setting to work". The sequel is dedicated to expounding this claim. This will be more informal than what has preceded, and is intended to help build an intuition as to why the p-adic numbers naturally arise. The discussion will require the basics of the theory of schemes. We acknowledge that the following viewpoint is largely inspired by the exposition [You17].

As we stated at the beginning of the section, there is a near ubiquitous philosophy of problem solving in mathematics that a "problem on X" should be equivalent to "problems locally on X" and "gluing data", with the understanding that find the local solutions and the data of how to glue them together into a global one is more approachable than just solving the problem globally. As an example, say we have some map $f\colon Y \to X$ (we are being quite vague here; X and Y are understood to be some geometric objects and f a morphism in the relevant category) which we want to find a section $s\colon X \to Y$ of. Then we can find local sections and try to glue them to find a global section.

Now suppose that our goal is to solve a Diophantine equation. Then there is a way to put this into the setting of finding a section. Associated to a ring A, we have a locally ringed space (a topological space equipped with a structure sheaf $\mathscr{O}$) $\operatorname{Spec} A$. An element $x \in A$ then defines a function

$$x\colon \operatorname{Spec} A \to \bigsqcup_{\mathfrak{p} \in \operatorname{Spec} A} \operatorname{frac}(A/\mathfrak{p})$$

given by $\mathfrak{p} \mapsto x \pmod{\mathfrak{p}}$. If we now fix an m-tuple of polynomials over A,

$$f = (f_1, \dots, f_m),$$

with $f_i \in A[x_1, \ldots, x_n]$, we can consider the *solution set functor*

$$X_f\colon \{A\text{-algebras}\} \to \text{Set}.$$

This functor associates to an A-algebra S the set of n-tuples in S which are solutions to all f_i:

$$X_f(S) = \{(y_1, \ldots, y_n) \in S^n \mid f_i(y_1, \ldots, y_n) = 0 \text{ for all } i = 1, \ldots, m\}.$$

On the other hand, a map

$$\alpha\colon \frac{S[x_1, \ldots, x_n]}{(f_1, \ldots, f_m)} \to S$$

corresponds uniquely to an element of $X_f(S)$ by assigning to α the element $(\alpha(x_1), \ldots, \alpha(x_n))$. Thus there is a natural bijection between elements of $X_f(S)$ (solutions to $f = 0$) and maps $\alpha\colon S[x_1, \ldots, x_n]/(f_1, \ldots, f_m) \to S$. It is well known that a map of rings corresponds to a map in the opposite direction on their spectra:

$$\operatorname{Spec} S \to \operatorname{Spec}(S[x_i]/(f_j)).$$

Thus we have a bijection

$$X_f(S) \leftrightarrow \{\text{Sections of } \operatorname{Spec}(S[x_i]/(f_j)) \to \operatorname{Spec} S\}.$$

With this we have placed the task of solving a polynomial equation in the setting of finding a section of a map between geometric objects. If our goal is specifically Diophantine equation solving, then we can take $S = A = \mathbb{Z}$ in the setup above.

As remarked before, we now look to work "locally". In a more familiar *milieu*, say in $\mathbb{C}$, we could just consider an open neighborhood of a point. The issue, however, is that the Zariski topology is not well suited for this. It is far too coarse. Recall for $X = \operatorname{Spec} A$, the basic open sets of X are distinguished opens $D(f)$ for $f \in A$, which are the complements of the vanishing loci of elements in A. Towards the end of obtaining a "more local" neighborhood of a given point $\mathfrak{p} \in \operatorname{Spec} A$, we may consider intersecting all opens which contain $\mathfrak{p}$:

$$\bigcap_{D(f) \ni \mathfrak{p}} D(f) = \varprojlim_{D(f) \ni \mathfrak{p}} D(f).$$

Recall that the ring of functions on a distinguished open is

$$\mathscr{O}(D(f)) = A\left[\frac{1}{f}\right].$$

Since a sheaf is a contravariant functor (inclusion reversing), we have

$$\mathscr{O}\left(\varprojlim_{D(f)\ni\mathfrak{p}} D(f)\right) = \varinjlim_{f\notin\mathfrak{p}} A\left[\frac{1}{f}\right].$$

Taking $A = \mathbb{Z}$, this becomes

$$\varinjlim_{f\notin(p)} \mathbb{Z}\left[\frac{1}{f}\right] = \mathbb{Z}_{(p)},$$

the localization of $\mathbb{Z}$ at the prime ideal (p). This suggests then that maybe $\operatorname{Spec}\mathbb{Z}_{(p)}$ is a sufficiently small neighborhood. Unfortunately, it is still not quite adequate. The obstruction now is a topological one. One of the reasons why we like to work with an open disk centered at a point in $\mathbb{C}$ is that it is contractible. Hoping for a similar level of niceness in our framework, we would want the map

$$\{(p)\} = \operatorname{Spec}\mathbb{F}_p \to \operatorname{Spec}\mathbb{Z}_p$$

to be some sort of "(weak) homotopy equivalence"[3], but it is not. Without going into full detail, the way to obtain this "homotopy equivalence" is to take the *Henselization* of $\mathbb{Z}_{(p)}$, $\mathbb{Z}_{(p)}^h$. Morally, we can think of $\mathbb{Z}_{(p)}^h$ being obtained by adding to $\mathbb{Z}_{(p)}$ minimally many points so as to make Hensel's lemma work. Returning to the complex analog, the parallel object is the ring

$$\mathbb{C}\langle x-p\rangle = \varinjlim_{U\ni p}\{\text{Holomorphic } f\colon U\to\mathbb{C}\}.$$

This is a Henselian local ring with maximal ideal $(x-p)$. Essentially this is just the ring of power series $\sum_{i\geq 0} a_i(x-p)^i$ which converge in a neighborhood of p. In practice, when working with such power series it is nice to disregard convergence properties initially to find a solution, then show the solution suitably converges. Thus we work in the ring $\mathbb{C}[[x-p]]$ of formal power series at $(x-p)$ and argue that our result lives in $\mathbb{C}\langle x-p\rangle$. By definition, one has

$$\mathbb{C}[[x-p]] = \varprojlim \mathbb{C}\langle x-p\rangle/(x-p)^n.$$

Putting this into the context of $\mathbb{Z}$, taking the projective limit of quotients of our Henselian local ring by powers of its maximal ideal yields

$$\varprojlim \mathbb{Z}_{(p)}^h/p^n\mathbb{Z}_{(p)}^h.$$

[3] This is made rigorous using the theory of *Grothendieck topologies.*

After noting that $\mathbb{Z}_{(p)}^h/p^n\mathbb{Z}_{(p)}^h \cong \mathbb{Z}_{(p)}/p^n\mathbb{Z}_{(p)} \cong \mathbb{Z}/p^n\mathbb{Z}$, we see that

$$\varprojlim \mathbb{Z}_{(p)}^h/p^n\mathbb{Z}_{(p)}^h \cong \varprojlim \mathbb{Z}/p^n\mathbb{Z} = \mathbb{Z}_p,$$

and thus the p-adic integers have appeared.

In this way we observe that, steadfast in our number theoretic goals of solving Diophantine equations, by following the classical mathematical philosophy of "solving global problems locally", we inevitably arrive at the p-adic numbers.

Bibliography

[Gui18] P. Guillot. *A Gentle Course in Local Class Field Theory: Local Number Fields, Brauer Groups, Galois Cohomology.* Cambridge University Press, 2018.

[Lan11] S. Lang. *Algebra.* Graduate Texts in Mathematics. Springer New York, 2011.

[NS13] J. Neukirch and N. Schappacher. *Algebraic Number Theory.* Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2013.

[Ser73] J.P. Serre. *A Course in Arithmetic.* Graduate Texts in Mathematics. Springer, 1973.

[SG13] J.P. Serre and M.J. Greenberg. *Local Fields.* Graduate Texts in Mathematics. Springer New York, 2013.

[You17] A. Youcis, 2017.

5. Root Systems and Lie Groups

—Zoe Siegelnickel—

Abstract

To motivate this paper, I shall simply state two theorems, which we will endeavor to prove;
Existence: For any irreducible root system ¶, there exists a simple Lie algebra over $\mathbb{C}$ which has a root system equivalent to ¶.
Uniqueness: It is also the case that any two Lie algebras over $\mathbb{C}$ with equivalent root systems are isomorphic.

5.1 Root systems

Definition: *A Euclidean vector space is a real vector space V with a positive definite symmetric bilinear form which we will call the dot product, i.e a bilinear form B such that $B(v,w) = B(w,v)$ for all $v, w \in V$ and $B(v,v) > 0 \ \forall v \neq 0$.*
Definition: *Let Φ be a subset of a finite dimensional real vector space V which is equipped with the dot product. Φ is a root system if:*

- *Φ is a finite set of non-zero vectors*
- *Φ spans V.*
- $\alpha, \beta \in \Phi \implies \beta - \frac{2\langle\alpha,\beta\rangle}{\langle\alpha,\alpha\rangle}\alpha \in \Phi$

If the root system is crystalline, then we have a fourth condition:

- $\alpha, \beta \in \Phi \implies \frac{2\langle\alpha,\beta\rangle}{\langle\alpha,\alpha\rangle} \in \mathbb{Z}$

Definition: *A subset $\Delta \subset \Phi$ is a base if the following conditions are satisfied:*

- *Δ is a basis for V as s vector space, where $\Phi \subseteq V$*

- *Each root $\alpha \in \Phi$ can be expressed as a linear combination of elements in Δ with linear coefficients such that the coefficients are either all positive or all negative.*

A root in Δ is called a simple root.

Definition: *Let $\langle \beta, \alpha \rangle = \frac{2(\alpha,\beta)}{(\alpha,\alpha)}$. Two root system (V_1, Φ_1) and (V_2, Φ_2) are isomorphic if there is an invertible linear map between V_1 and V_2 that preserves $\langle \alpha, \beta \rangle$.*

Definition: *For $\alpha \in V$, H_α denotes the hyperplane perpendicular to α, i.e $\beta \in V : \langle \alpha, \beta \rangle = 0$*

In any root system Φ the hyperplanes H_α for some α divide V into connected components, which are the Weyl chambers of V.

Definition: *Let Φ be a root system in a Euclidian space V. For each root $\alpha \in \Phi$, define $s_\alpha(\beta)$ as $\beta - 2\frac{(\beta,\alpha)}{(\alpha,\alpha)}\alpha$ where $(,)$ is the inner product on V. The Weyl group of Φ is the subgroup generated by the s_α*

It is a fact that every root is conjugate to a simple root under the Weyl group.

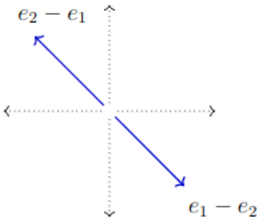
Definition: *A root system Φ which is non empty is said to be irreducible if it is not the direct sum of two nonempty root systems*

Definition: *A nonempty root system Φ is said to be reducible if it can be written as a disjoint union of nonempty root system Φ_1, Φ_2, i.e $\Phi = \Phi_1 \bigsqcup \Phi_2$*

Each root system can be written as the direct sum of irreducible root systems, and this summation is unique up to the ordering of the terms. Therefore, it suffices to only consider the irreducible root systems in our classification.

5.1.1 Examples

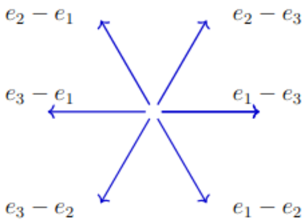
Take $V = \mathbb{R}^2$ with the standard basis $\{e_1, e_2\}$. The A_1 root system $\Phi = \{e_1 - e_2, e_2 - e_1\}$ is pictured below:



We can check the integrality condition:

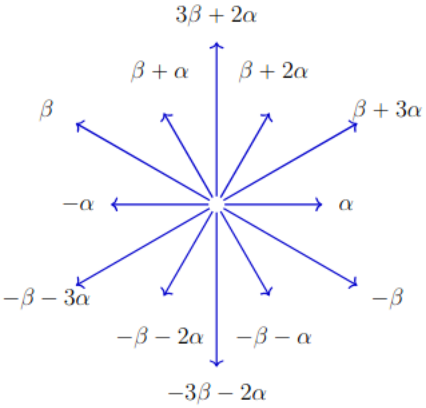
$$\frac{2(e_1 - e_2, e_2 - e_1)}{(e_2 - e_1)} = \frac{2(-1-1)}{(1+1)} = -2$$

Let e_1, e_2, e_3 be the standard basis of $\mathbb{R}^3$. The A_2 root system $\Phi = \{e_1-e_2, e_2-e_1, e_1-e_3, e_3-e_1, e_2-e_3, e_3-e_2\}$ is a root system in the subspace $V = Span(\Phi)$, which is the plane with normal vector $e_1+e_2+e_3$. This root system is the A_2 root system, and fulfills the last integrality condition, and has base $\Delta = \{e_1-e_2, e_3-e_1\}$



In general, we can define the A_l root system as $\Phi = \{\pm(e_i, e_j) : 1 \leq i|j \leq l+1\}$ where $e_1, e_2, ..., e_{l+1}$ is the standard basis of $\mathbb{R}^{l+1}$, and $V = Span(\Phi) \subset \mathbb{R}^{l+1}$ equipped with the dot product.

We now consider the more complex G_2 root system. Let e_1, e_2, e_3 and V be as before. Then, the G_2 root system is the set of vectors $\{\pm(e_1-e_2), \pm(e_1-e_3), \pm(e_2-e_3), \pm(2e_1-e_2-e_3), \pm(2e_2-e_1-e_3), \pm(2e_3-e_1-e_2)\} = A_2 \cup \{\pm(2e_1-e_3-e_3), \pm(2e_2-e_1-e_3), \pm(2e_3-e_1-e_2)\}$. Let $\alpha = e_1-e_2$ and $\beta = 2e_2-e_1-e_3$. The base for G_2 is $\Delta = \{\alpha, \beta\}$.



5.1.2 Classification

It is an interesting consequence that the integrality condition yields some constraints on the possible angles between two roots. Consider the following:

$$\langle\beta,\alpha\rangle\langle\alpha,\beta\rangle = 2\frac{(\alpha,\beta)}{(\alpha,\alpha)}\frac{2(\alpha,\beta)}{(\beta,\beta)}$$

$$= 4\frac{(\alpha,\beta)^2}{|\alpha|^2|\beta|^2} = 4\cos^2(\theta) = (2\cos\theta)^2 \in \mathbb{Z}$$

Since $2\cos\theta \in [-2, 2]$, we see that the only possible values for $\cos\theta$ are $0, \pm\frac{1}{2}, \pm\frac{\sqrt{2}}{2}, \pm\frac{\sqrt{3}}{2}, \pm 1$.The corresponding angles are $60°, 120°, 90°, 45°, 135°, 30°, 150°, 0°, 180°$. Recall that if α is a root, the only multiples of the α in the root system are α and $-\alpha$. Therefore, 0° and 180° are not possible angles, since they correspond to 2α and -2α. We note that roots at an angle of 60° or 120° are of equal length, roots at an angle of 45° or 135° have a ratio of $\sqrt{2}$, and roots at an angle of 30° or 150° correspond to a length ratio of $\sqrt{3}$.

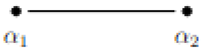
Dynkin Diagrams

Let Φ be a root system with base Δ. We can construct the associated Dynkin diagram by drawing a vertex for each root in Δ and drawing edges between these vertices according to the following rules:

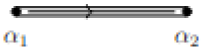
- If the roots associated with two vertices is orthogonal, then there is no edge.
- If the two roots form an angle of 120°, then there is an undirected single edge.
- If the vectors form an angle of 135°, then there is a directed double edge.
- If the vectors form an angle of 150°, there is a directed triple edge.

5.1.3 Examples

Recall the A_2 root system. The Dynkin diagram has two vertices α_1, α_2, with one undirected edge:

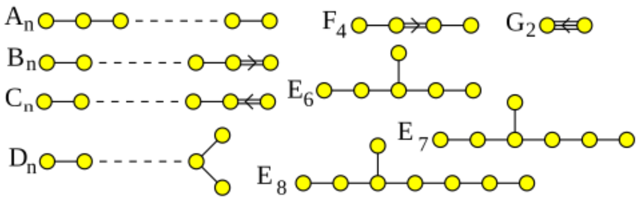


Let α_1, α_2 be vertices representing the two elements in the base of G_2. We see that they form an angle of 150°, and so the Dynkin diagram is



Connected Dynkin diagrams can all be classified as one of 8 pictures:

$$A_n, B_n, C_n, D_n, G_2, F_4, E_6, E_7, E_8$$



5.2 Classification of Lie Algebras

Definition. *A Lie Algebra is a vector space $\mathfrak{g}$ over a field with a Lie bracket, which satisfies the following:*

- $[ax + by, z] = a[x, z] + b[y, z]$
- $[z, ax + by] = a[z, x] + b[z, y]$
- *Jacobi Identity:* $[x, [y, z]] + [y, [z, x]] + [z, [x, y]] = 0 \forall x, y, z \in \mathfrak{g}$

A Lie algebra is semisimple if it is a direct sum of non-abelian Lie algebras with no non-zero proper ideals (simple Lie algebras).

Definition: *A Cartan Subalgebra $\mathfrak{h} \subseteq \mathfrak{g}$ is an abelian, diagonalizable subalgebra which is maximal under set inclusion, with dimension equal to the rank of $\mathfrak{g}$.*

Cartan subalgebras always exist for finite dimensional complex Lie algebras, and are all conjugate to each other under automorphisms of the Lie algebra, meaning that they all have the same dimension. It is possible to classify semisimple Lie algebras defined over a algebraically closed field of characteristic zero by finding the root systems associated with their Cartan Subalgebras, which as we have discussed above, are classified according to their Dynkin diagrams. Let $\{H_1, ..., H_2\}$ be a basis for $\mathfrak{h}$. Extending this basis to a basis of $\mathfrak{g}$ will yield a basis with very nice commutator relations, since any Cartan subalgebra is abelian and so $[H_i, H_j] = 0$.

Definition: *The adoint operator of x for $x \in \mathfrak{g}$, denoted $ad_x : \mathfrak{g} \to \mathfrak{g}$ takes $x \mapsto [x, y]$.*

The adoint operators determine the linear mapping $ad : \mathfrak{g} \to gl(\mathfrak{g})$, the Lie algebra of all linear endomorphisms of $\mathfrak{g}$. Since we consider only finite dimensional Lie Algebras, $gl(\mathfrak{g})$ is the Lie algebra of square matrices under matrix

multiplication. We see that ad is a representation of $\mathfrak{g}$ called the adoint representation.
We will now note some nice facts about linear operators:

- Pairwise commuting, diagonalizable linear operators share a common set of eigenvectors.
  Proof. Since we are working with matrices, we shall do a nice matrix proof. Note that if $Ax = \lambda x$. Then $ABx = BAx = B\lambda x = \lambda Bx$ since we have assumed that A, B are pairwise commuting. Then, x, Bx are eigenvectors of A,

- For $H_1, H_2 \in \mathfrak{h}$, ad_{H_1}, ad_{H_2} commute and are diagonalizable. By the first fact, they then share a common set of eigenvectors.
  Proof. First, we show that they commute; by the Jacobi identity, we have that

  $$[H_1,[H_2,X]] = -[H_2,[X,H_1]]-[X,[H_1,H_2]] = -[H_2,[X,H_1]]-[X,0] = [H_2,[H_1,X]]$$

  Recall from linear algebra that if two linear transformations have the same eigenvectors, then they can be simultaneously diagonalized. Therefore, we have the desired result.

By the spectral theorem, we can decompose $\mathfrak{g}$ into shared eigenspaces g_α of the adoint operators:

$\mathfrak{g} = \mathfrak{h} \oplus \bigoplus_{\alpha \in \Phi} \mathfrak{g}_\alpha$ where the $\alpha's$ are the eigenvalues of ad_{H_i} on the eigenspace $\mathfrak{g}_\alpha$

Therefore, for each eigenvector $E \in \mathfrak{g}_\alpha$, $[H_i, E] = \alpha_i E$. Each such α_i is called a root of $\mathfrak{g}$. Let Φ denote the set of roots. Φ forms a root system in $\mathbb{R}^r$, where r is the rank of $\mathfrak{g}$. In particular, each eigenspace $\mathfrak{g}_\alpha$ for $\alpha \in \Phi$ is one-dimensional. We can now direct our attention to proving the two theorems stated in the beginning.

5.2.1 The Nice Stuff

Serre's Theorem: Given a root system Φ in a Euclidean space with inner product $(,)$, $\langle \beta, \alpha \rangle$ defined as before and base $\{\alpha_1, \alpha_2, ...\alpha_n\}$, the Lie algebra $\mathfrak{g}$ defined by $3n$ generators e_i, f_i, h_i and the relations

$$[h_i, h_j] = 0$$

$$[e_i, f_i] = h_i, \ [e_i, f_j] = 0, i \neq j$$

$$[h_i, e_j] = \langle \alpha_i, \alpha_j \rangle e_j,\ [h_i, f_j] = -\langle \alpha_i, \alpha_j \rangle f_j$$
$$ad(e_i)^{-\langle \alpha_i, \alpha_j \rangle + 1}(e_j) = 0,\ i \neq j$$
$$ad(f_i)^{-\langle \alpha_i, \alpha_j \rangle + 1}(f_j) = 0,\ i \neq j$$

is a finite-dimensional semisimple Lie algebra with the Cartan subalgebra generated by the $h_i's$ and with the root system Φ.

Sketch: $L_0 = \bar{L}/\bar{K}$ where $\bar{K}$ is the ideal in $\bar{L}$ where $\bar{L}$ is a free Lie algebra generated on 3n elements by the following generators: $\{e_i, f_i, h_i | 1 \leq i \leq l\}$. Let $\bar{K}$ be generated by $[h_i, h_j], [e_i, f_i] - \delta_{ij} h_i, [h_i, e_i] - c_{ji}, [h_i, f_i] + c_{ji} f_i$ where c_{ij} is the Cartan integer $\langle \alpha_i, \alpha_j \rangle$. Let L_0 be decomposed into $E + F + H$ where E is generated by the e_i and F is generated by the f_i.

Now, let $L = L_0/K$ where K is the ideal generated by all e_{ij}, f_{ij} $i \neq j$.

We will first consider elements of L_0. Let I be the ideal of E generated by all the e_{ij} and J be the ideal of F generated by the f_{ij}. Note that this means that K includes I and J. We shall proceed from here in steps, to avoid any further confusion than that caused by these definitions.

1. **I and J are ideals of L_0.** The argument for I and J will be roughly the same, so we consider only J. First, we see that y_{ij} is an eigenvector for $ad\ h_k$ (this is discussed above) with eigenvalue $-c_{jk} + (c_{ji} - 1)c_{ik}$. Since $ad\ h_k(F) \subset F$, we have that $ad\ h_k(J) \subset J$ by the Jacobi identity. However, it is also the case that $ad\ e_k(f_{ij}) = 0$. Then, e_k maps F into $F + H$, and so since $ad\ h_k(J) \subset J$, we have that $ad\ e_k(J) \subset J$ again by the Jacobi identity. Then we have also $ad\ L_0(J) \subset J$.

2. $K = I + J$. Recall that $I + J \subset K$. But by 1), we have that $I + J$ is an ideal of L_0 which contains all e_{ij}, f_{ij}, and K is the smallest such ideal. Therefore, we have that $I + J = K$.

3. **Let $N^- = E/F$, $N = E/I$. Then, $L = N^- + H + N$ where $+$ denotes the direct sum of subspaces.** Let H be identified with its image under the canonical map $L_0 \to L$. This follows fairly directly from 2) and the direct sum decomposition $L_0 = E + F + H$.

4. **$E \oplus F \oplus H$ is isomorphic to L.** We won't thoroughly prove this, but it follows loosely from the relations detailed above, since we have already shown that H maps isomorphically into L by 3). As a consequence, we can identify e_i, f_i, h_i with elements of L, and in fact these generate L.

5. **If $\lambda \in H^*$, then $L_\lambda = \{x \in L | [hx] = \lambda(h)(x)\ \forall h \in H\}$. Then, $H = L_0$ and $N = \sum_{\lambda>0} L_\lambda$, $N^- = \sum_{\lambda<0} L_\lambda$, and each L_λ is finite dimensional**. This remark follows from 3) and 4).

6. **For $1 \leq i \leq n$, we have that $ad\ e_i$ and $ad\ f_i$ are locally nilpotent endomorphisms of L.** Again, we have that the arguments for the e_i is roughly the same as the argument for the f_i, so we consider only the e_i. Let M be the subspace of all elements of L that are killed by some power of ade_i. If $e \in M$ is killed by $(ade_i)^r$, and $f \in M$ is killed by $(ade_i)^r$, then $[e, f]$ is killed by $(ade_i)^{r+s}$. Then M is a subalgebra of L, but all $e_k \in M$ and all $f_k \in M$, and these elements generated L, so $M = L$.

7. **Let $\tau_i = \exp(ade_i)\exp(ad(-f_i))\exp(ade_i)$ for $1 \leq i \leq n$. Then, τ_i is a well defined automorphism of L.** We also won't prove this fact rigorously, but it follows from 6).

8. **If $\lambda, \mu \in H^*$, and $\sigma\lambda = \mu$ for σ in the Weyl group of Φ, then $\dim L_\lambda = \dim L_\mu$.** It suffices to consider only the generators of the Weyl group. The automorphism τ_i of L from 7) coincides on the finite dimensional space $L_\lambda + L_\mu$, and we see that τ_i interchanges L_λ and L_μ. In particular, we see that $\dim L_\lambda = \dim L_\mu$.

9. **For $1 \leq i \leq n$, $\dim L_\alpha = 1$, while $L_{k\alpha_i} = 0$ for $k \neq -1, 0, 1$.** It follows from 4) that this holds for L_0, and consequently must hold for L.

10. **If $\alpha \in \Phi$, then $\dim L_\alpha = 1$ and $L_{k\alpha} = 0$ for $k \neq -1, 0, 1$.** Recall that each root is conjugate to a simple root under the action of the Weyl group. Therefore, this follows 8), 9).

11. **If $L_\lambda \neq 0$, then either $\lambda \in \Phi$ or $\lambda = 0$.** If this were not the case, then λ would be an integral combination of simple roots with coefficients that were either all positive or all negative. We see that by 10), λ is not a multiple of a root. Let $\sigma\lambda$ be a conjugate of λ under the Weyl group action. By various properties of this action, we see that $L_{\sigma\lambda} = 0$, which contradicts 8).

12. **$\dim L = n + |\Phi| < \infty$.** Since by 5 we see that each L_λ is finite dimensional, this follows by 10) and 11).

13. **L is semisimple.** Let A be an abelian ideal of L. We show that $A = 0$. Note that $ad\ H$ stabilizes A, and so $A = A \cap H + \sum_{\alpha\in\Phi}(A \cap L_\alpha)$

since $L = H + \sum_{\alpha \in \Phi} L_\alpha$. If $L_\alpha \in A$, then $[L_{-\alpha}, L_\alpha] \subset A$ where $L_{-\alpha} \subset A$ and $\mathfrak{sl}_2(F) \subset A$ where L is an algebra over F. This cannot be the case, and so $A = A \cap H \subset H$ where $[L_\alpha, A] = 0$ for $\alpha \in \Phi$ and $A \subset \bigcap_{\alpha \in \Phi} \text{Ker}\alpha = 0$.

14. H **is a Cartan subalgebra of** L **and** Φ **is the root system.** H is abelian, and therefore nilpotent and, due to the direct sum decomposition self-normalizing. This is precisely the definition of a Cartan subalgebra, and it is immediete that Φ is the corresponding set of roots. □

This theorem implies existence. Let us restate the uniqueness theorem as follows:
Let L, L' be semisimple Lie algebras, with respective Cartan sub-algebras H, H' and root system Φ, Φ'. let an isomorphism $\Phi \to \Phi'$ be given, sending a given base Δ to a base Δ', and inducing the isomorphism $\pi : H \to H$. For each $\alpha \in \Delta$ (respectively ($\alpha' \in \Delta'$)), select an arbitrary nonzero $x_\alpha \in L_\alpha$ (respectively ($x'_\alpha \in L'_\alpha$)). Then, there exists a unique isomorphism $\pi : L \to L'$ extending $\pi : H \to H'$ and sending x_α to $x_{\alpha'}$ for $\alpha \in \Delta$.
Proof. It suffices to show the case where L is the lie algebra constructed according to Serre's theorem. Take e_α, f_α and $h_\alpha = [e_\alpha, f_\alpha]$ to be the specified generators with $\alpha \in \Delta$. Set $h'_\alpha = \pi(h_\alpha)$ and choose $f'_{\alpha'}$ uniquely satisfying $[x'_\alpha, y'_\alpha] = h'_{\alpha'}$ for each $\alpha' \in \Delta'$. Since $\Phi \cong \Phi/$, the chosen elements in L' satisfy the relations in Serre's theorem. Therefore, Serre's theorem provides a unique homomorphism $\pi : L \to L'$ sending $e_\alpha, f_\alpha, h_\alpha (\alpha \in \Delta)$ to $e'_\alpha, f'_\alpha, h'_\alpha$ respectively, extending the given isomorphism $\pi : H \to H'$. Since $\dim L = \dim H + |\Phi| = \dim H' + |\Phi'| = \dim L'$, we see that π is indeed an isomorphism. □

5.2.2 Examples

Consider the special linear Lie algebra $\mathfrak{sl}_n(\mathbb{C})$, and let $\mathfrak{h}$ be the subalgebra of diagonal matrices with trace 0. Then, the root vectors are matrices $E_{i,j}$ where $i \neq j$, with a 1 in i, j spot and zeroes everywhere else. Then, $[H, E_{i,j}] = (\lambda_i - \lambda_j)E_{i,j}$ where H is the diagonal matrix with entries $\lambda_1, ..., \lambda_n$. Therefore, we can represent the roots as the linear functionals $\alpha_{i,j}(H) = \lambda_i - \lambda_j$. However, we can identify $\mathfrak{h}$ with its dual $\mathfrak{h}^*$, and so we can rewrite the roots as the vectors $\alpha_{i,j} = e_i - e_j$ in the subspace of $\mathbb{R}^n$ consisting of n-tuples that sum to 0. This can be identified as the A_{n-1} root system. For example, we see that the associated root system of $\mathfrak{sl}_2(\mathbb{C})$ is $\{e_1 - e_2, e_2 - e_1\}$ which is the A_1 root system.

Bibliography

[Emo] Melissa Emory. Paws 2024: Symmetries of root systems.

[Hum] James Humphreys. *Introduction to Lie Algebras and Representation Theory.* Springer, rev. and compl. ed. edition.

[Meh] David Mehrle. Lie algebras and their root systems.

[Mor24] J Morgan. Lie groups: Fall, 2024 lecture ii lie algebras, the adjoint action, and the exponential mapping. 2024.

[Ver] Root System Basics.

6. Elliptic Curve Cryptography

—*Skyler Marks*—

Abstract

The digital world is kept secure by cryptography. The idea behind most modern cryptographic systems is that if I have a public number (usually called a public key) and a secret number (usually called a private key) I can perform some operation with them to generate a third number (which is also public). This number is easy to generate by combining a public and private key, but hard to generate any other way; this allows us to verify the authenticity of a private key very easily. One such operation involves finding collinear points on an elliptic curve, giving rise to elliptic curve cryptography. This talk introduces elliptic curves over finite fields, explains how we can use their algebraic geometry to define an appropriate operation, and touches on why this operation is appropriate for cryptography. Although this talk is considerably mathematical, no background beyond calculus and basic (high school) algebra is necessary (although some linear algebra and a basic grasp of set theory will help).

Warning: Throughout these notes, I've been somewhat loose with notation, using conventions that are standard in algebra but may seem confusing to the layperson. This is because these notes are meant to accompany a lecture, and I have the opportunity to explain as I go in that setting. If you have any questions, please consult a textbook ([**DF**] is an excellent resource for algebra), or email me at `skyler@bu.edu`

6.1 Intuition

Elliptic curve cryptography is fairly simple on the intuitive level; elliptic curves have the interesting property that each (non-vertical) line intersects that curve in three points. This allows us to define a operation, kind of like multiplication, by finding successive intersections.

Definition 6.1.1. An **elliptic curve** is the set of pairs of 'numbers' (for an appropriate definition of 'numbers', as we will describe) (x, y) satisfying the equation:

$$y^2 = x^3 + ax + b$$

Generally, in math, this definition is a little more complicated, because mathematicians deal with more exotic spaces than just the complex plane. However, for the sake of simplicity, we'll just work in within the complex numbers. An abstract equation is a little tricky to work with, so let's graph a prototypical elliptic curve:

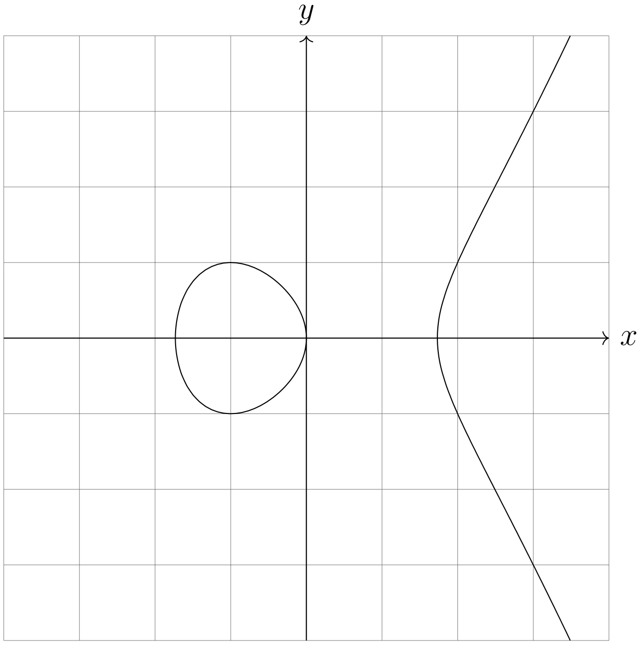


Figure 6.1: The graph of the equation $y^2 = x^3 - 3x$.

For my illustrations, I'll work with this curve - mostly because it looks nice. Some properties we can observe immediately include the symmetry over the x axis - this will be extremely useful for our algorithm. First, we pick a base point B on the curve - this will function like a seed for our algorithm. We'll also pick another point K. The main operation of the algorithm is to take a point K and connect it by a line with B. This line will then intersect the elliptic curve in a third point C:

Now that we've got the gist of this operation down, we'll define our algorithm. The idea is we pick a private key n, which is just an integer (we'll see later on that there are some interesting ways to make the points B, C,

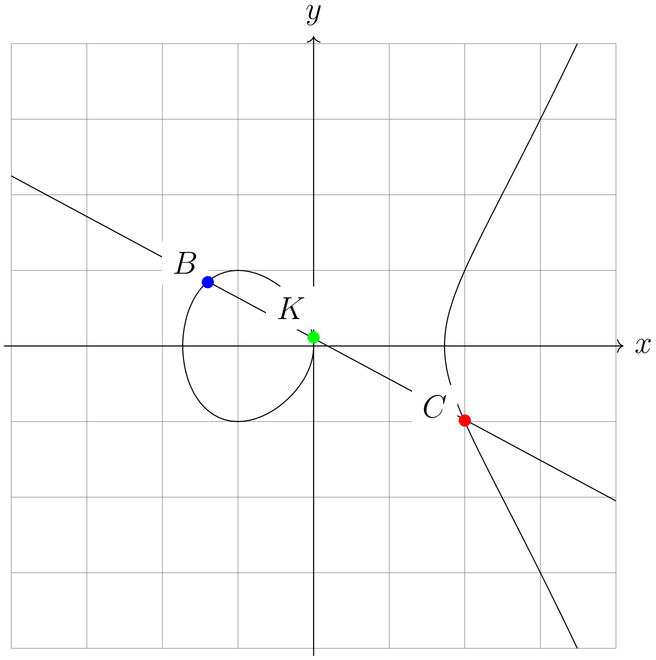


Figure 6.2: The first iteration of our process.

and K integers as well). Leveraging the symmetry of the elliptic curve, we now reflect the point C over the X axis to obtain a new point on the elliptic curve K_1, and run the whole game again with K_1 to obtain a new point C_1:

We continue this n times, reflecting C_{i-1} to K_i and finding a point C_i collinear with both K_i and B, to obtain a final point C_n, which will be our public key. It should be intuitively clear that, while it's easy to perform this process a given number of times, it's 'quite difficult' to work out how many times this process was performed in order to attain a particular result. Solving this problem (computing how many times this operation is performed) is called computing the **elliptic curve discrete logarithm** (a special case of the **discrete logarithm problem**). An active field of mathematical research is working out how difficult this problem is for elliptic curves; so far, no subexponential time has been found (see [**ecdlt**]) and the time complexity is greater than that of modular exponentiation.

This intuitive explanation is fairly complete, but lacks some nuance. In particular, this talk will address two main follow up questions to this naive treatment; firstly, we'll discuss a geometric formalism that allow us to deal with difficult edge cases regarding this operation, and secondly we'll use some abstract algebra to make computations with these curves feasible.

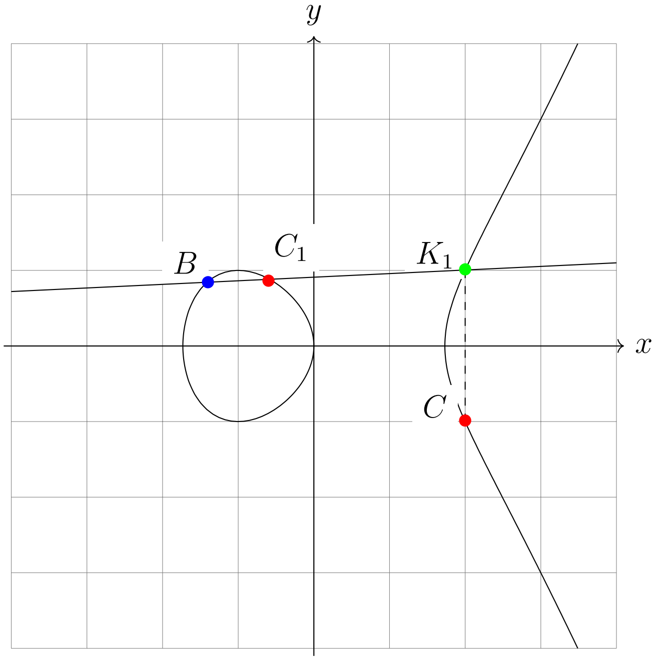


Figure 6.3: The second iteration of our process.

6.2 A whirlwind tour of Abstract Algebra

6.2.1 Set Theory

Set theory is the language of mathematics; we'll need some of it to formulate most of the notions we'll be discussing.

Definition 6.2.1. For the purposes of this talk, a **set** will be a collection of objects. For more information, consider looking up the Wikipedia page for ZFC (Zermelo - Frankel - Choice set theory, the foundations for most modern math). An element of a set is one of the objects in the set. The notation $a \in A$ means that the object a is in the set A.

Definition 6.2.2. Let A and B be sets. The union of A and B, denoted $A \cup B$, is the set of all elements which are in A or in B (inclusive or). The intersection of A and B, denoted $A \cap B$ is the set of all elements which are in A and in B. The difference $A - B$ or $A \setminus B$ is the set of all elements in A which are not in B.

Definition 6.2.3. A **function** from a set A to a set B is a rule that associates with every element of A an element of B. A function is **injective** if $f(a) =$

$f(b)$ implies $a = b$, and **surjective** if for each b in B there is an a in A with $f(a) = b$.

Definition 6.2.4. The **Cartesian product** of two sets A and B, written $A \times B$, is the set of all pairs (a, b) with $a \in A$ and $b \in B$.

Definition 6.2.5. An **equivalence relation** on a set S is a subset R of $S \times S$ satisfying the following properties:

- Reflexivity: For every a in S, (a, a) is in R.
- Symmetry: If (a, b) in R, then (b, a) is in R.
- Transitivity: If (a, b) is in R and (b, c) is in R, then (b, a) is in R.

We write $a \sim b$ to indicate that (a, b) is in the set R.

Lemma 6.2.6. *Let S be a set and $\sim$ be an equivalence relation on S. Let $[a]$ denote the set of all $b \in S$ satisfying $a \sim b$. Every $[a]$ is either equal or disjoint to every other $[b]$, and every element of S is in some $[a]$.*

Proof. By reflexivity, $a \in [a]$, so every element $a \in S$ is in $[a]$. If there is some c in both $[a]$ and $[b]$, then $a \sim c$ and $c \sim b$. But then because $d \sim b$ for each $d \in [b]$, we know that $a \sim d$ for each $d \in [b]$ by transitivity, so $[a]$ contains $[b]$. Similarly, $[b]$ contains $[a]$. □

6.2.2 Groups

Somewhat central to elliptic curve cryptography (and most discrete logarithm style problems) is the notion of a group:

Definition 6.2.7. A **group** is a set G equipped with a binary operation, that is, a function $f : G \times G \to G$. We'll often write the group operation using infix notation using an operator like $\bullet$; $(a \bullet b)$, for example, denotes $f(a, b)$. This binary operation satisfies the following properties:

- Associativity: $a \bullet (b \bullet c) = (a \bullet b) \bullet c$.
- Identity: There is an element e of the set G such that for each g in the set g, $e \bullet g = g \bullet e = g$.
- Inverses: For every g in the set G, there is an element g^{-1} in G satisfying $gg^{-1} = g^{-1}g = e$.

Definition 6.2.8. A group is **abelian** or **commutative** if $a \bullet b = b \bullet a$ for each $a, b \in G$.

Example 6.2.9. The symmetries of a triangle are a group.

Example 6.2.10. The integers (under addition) form a group

Example 6.2.11. The integers modulo n form a group under addition.

Example 6.2.12. The integers modulo a *prime* p, if you take away 0, form a group under multiplication.

Proof. Associativity and identity follow from the fact that $a = b \implies a \equiv b \mod p$ (mod is a function); it suffices to show there are inverses. Let g be an element of G, the set of integers modulo a prime p. Recall that the GCD of two integers is a sum with integer coefficients of those integers (this can be proved using the Euclidean division algorithm). The GCD of g and p is 1, so there are integers a, b such that $ag + bp = 1$. Then $ag \equiv 1 \mod p$, so a is an inverse for g mod p (because modulo is multiplicative). □

Exercise 6.2.13. Write (in pseudocode) an algorithm to find the GCD of two integers a and b using repeated division / finding the remainder. Prove that this algorithm terminates, and use it to show that the GCD of two integers can be written in the form $a'a + b'b$ where a', b' are also integers.

Exercise 6.2.14. If n is an non-prime integer, what elements in the set of integers mod n have multiplicative inverses?

Lemma 6.2.15 (Shoe-Socks Lemma). *In any group,* $(ab)^{-1} = b^{-1}a^{-1}$.

Proof. Note that $b^{-1}a^{-1}ab = b^{-1}eb = e$. □

Especially important to us are cyclic subgroups of a group.

Definition 6.2.16. A **subgroup** H of a group G is a subset of G that satisfies the group axioms for the same operation as G.

Definition 6.2.17. A **cyclic subgroup generated by** g for some g in a group G is the set of all 'powers' of g, that is the set of all elements of the form $g \cdot g \cdot ...$ or $g^{-1} \cdot g^{-1} \cdot ...$, together with the identity.

Exercise 6.2.18. Check that a cyclic subgroup is a subgroup!

6.2.3 Rings and Fields

Definition 6.2.19. A **ring** is a set R with two binary operations called multiplication and addition, satisfying the following properties:

- Both operations are associative
- The set R is a group under multiplication with identity 0
- There is a multiplicative identity 1
- Multiplication distributes over addition; i.e., for every $a, b, c \in R$

$$a(b + c) = ab + ac \text{ and } (b + c)a = ba + ca$$

A ring is called **commutative** if $ab = ba$ for all a and b in the ring.

Lemma 6.2.20 (Nifty Facts About Rings)**.** *For any a in R, we have that $0a = 0$; furthermore, the additive group of a ring R is abelian.*

Proof. Suppose $a \in R$. Then $a(0+1) = a$. Distributing and subtracting a on both sides yields $a0 = 0$. If $0 = 1$, this implies $a = 0$ for every a. If not, then for $a, b \in R$ we have $-1(a + b) = (-1a + -1b) = -a - b$. But by Shoe-Socks, $-1(a + b) = -b - a$, so $-a - b = -b - a$. Multiplying both sides by -1 and distributing gives $a + b = b + a$. □

Example 6.2.21. The integers are a ring.

Example 6.2.22. The integers mod n are a ring.

Example 6.2.23. Polynomials in n variables with real, integer, or complex coefficients (actually, in any ring) form a ring under the multiplication and addition formulas we're familiar with.

Definition 6.2.24. A **subring** of a ring R is a subgroup of the additive group of R that is closed under multiplication. Subrings do not need to include 1.

Definition 6.2.25. An **ideal** of a ring R is a subring of R that is closed under multiplication by *every element* in the ring. If $\{a_k\}_{k \in K}$ is a collection of elements of R, the **ideal generated by** $\{a_k\}_{k \in K}$ or $(\{a_k\}_{k \in K})$ is the set of all sums of R-multiples of elements of the set $\{a_k\}_{k \in K}$.

Definition 6.2.26. An ideal I of R is **prime** if, for any elements a and b in R, $ab \in I$ implies either a or b (or both) in I. Equivalently, $R - I$ is multiplicatively closed.

Definition 6.2.27. An ideal $\mathfrak{m}$ of R is maximal if there is no ideal I satisfying $\mathfrak{m} \subsetneq I \subsetneq R$.

Example 6.2.28. The ideals of $\mathbb{Z}$ are of the form $n\mathbb{Z}$. The primes are of the form $p\mathbb{Z}$ for p prime.

The ideal structure of polynomials is extremely complex and the study of them comprises a large part of the mathematical discipline known as *commutative algebra.*

Definition 6.2.29. A **ascending chain** of ideals is a set of ideals $\mathfrak{a}_1, ..., \mathfrak{a}_n$ of a ring R satisfying $\mathfrak{a}_1 \subsetneq \mathfrak{a}_2 \subsetneq \mathfrak{a}_3 \subsetneq ... \subsetneq R$.

Definition 6.2.30. A ring is **Noetherian** if every ascending chain of ideals terminates.

Exercise 6.2.31 (Hilbert's Basis Theorem)**.** Show that if R is a Noetherian ring, then $R[x]$ is likewise a Noetherian ring.

If R is a Noetherian ring, then $R[x_1, ..., x_n]$ is Noetherian.

Exercise 6.2.32. The ideals of every Noetherian ring are finitely generated.

Definition 6.2.33. The **codimension** of a Noetherian ring is the length of the longest ascending chain of ideals.

Definition 6.2.34. A **field** is a commutative ring F where the set $F - \{0\}$ is a group under the ring multiplication.

Example 6.2.35. The rational numbers $\mathbb{Q}$, the set of ratios $\frac{p}{q}$ for p, q integers, form a field under the standard 'fraction multiplication'.

Example 6.2.36. The real numbers $\mathbb{R}$ and the complex numbers $\mathbb{Q}$ are fields.

Example 6.2.37. The integers modulo a prime p are a field.

Proof. Examples 6.2.11 and 6.2.2. □

Definition 6.2.38. A **vector space** over a field k is an abelian group V together with an operation $\cdot \colon k \times V \to V$ called scalar multiplication that is distributive and satisfies $0 \cdot v = \vec{0}$ (where $\vec{0}$ is the identity of the group) and $1 \cdot v = v$.

Example 6.2.39. The Cartesian product $k \times k \times k ... \times k$ is a vector space under componentwise addition and the scalar multiplication law:

$$x \cdot (a_1, a_2, ..., a_n) = (xa_1, xa_2, ..., xa_n)$$

We call this construction **affine** n**-space over** k

6.3 Zero Sets and Projective Space

We want to discuss the zero set of a particular polynomial, in a very general sense. To do this, we introduce some terminology:

Definition 6.3.1. Fix a field k. The **zero set** of a polynomial $P(x_1, ..., x_n)$ is the set of all $(x_1, ..., x_n)$ in affine n-space such that $P(x_1, ..., x_n) = 0$.

Definition 6.3.2. Projective n-space over a field k is the set of equivalence classes of the set $k \times k \times .... \times k - (0, ..., 0)$ (multiplied $n+1$ times) by the equivalence relation $a \sim b$ if and only if $(a_1, ..., a_{n+1}) = \lambda(b_1, ..., b_{n+1})$

Projective space is called 'projective' because it can be characterized as the set of all lines through the origin (or the additive identity of the vector space), and the notion of identifying a point with the line through it and the origin yields a formalization of the notion of projection.

Exercise 6.3.3. For those of you with a linear algebra or multivariate calculus background, convince yourself of the above statement.

The naming convention for projective n-space (as a quotient of affine $n+1$ space) may seem strange, but it comes from the fact that the equivalence relation 'cuts down' the dimension by 1. In particular, projective n-space 'looks locally' like affine n space; we can write down a function f the set of all points (equivalence classes) in projective n-space where the kth variable is nonzero to affine n space by the rule:

$$[(x_1, ..., x_n)] \mapsto \left(\frac{x_1}{x_k}, ..., \hat{x}_k, ..., \frac{x_n}{x_k}\right)$$

This map is 'continuous' in a way that we don't have time to make precise, but effectively it identifies affine space with projective space in a way that preserves geometric properties. This also motivates the idea that projective space 'adds points at infinity' - we can think of the points as infinity as the limits as that one coordinate goes to zero (after fixing a preferred coordinate at infinity).

We ultimately want to work with polynomial equations (specifically elliptic curves) defined over projective space. Naively, we could try to just plug the variables for projective space into a polynomial; however, that doesn't yield a well defined function. If P is a polynomial in n variables, in general we have:

$$\lambda P(x_1, ..., x_n) \neq P(\lambda x_1, ..., \lambda x_n)$$

In fact, equality only holds when P is linear. To fix this, we introduce:

Definition 6.3.4. The **degree** of a term in a polynomial is the sum of the powers of the indeterminate variables in that term. A **homogeneous polynomial** in n variables is a polynomial who's terms all have the same degree.

Example 6.3.5. The polynomials $f(x) = x^2$, $g(x, y) = x^3 + 3xy^2 + 6y^3$, and $h(x, y, z) = x^3 + y^2x + x^2y + y^3 + z^3$ are all homogeneous polynomials.

The main reason for this construction is that

Lemma 6.3.6. *The 'zero' of a homogeneous polynomial is a well defined notion in projective space.*

Proof. Let $[x] = [(x_1, ..., x_{n+1})]$ be an equivalence class in projective n-space over k and let P be a homogeneous polynomial of degree r. Then if $(y_1, ..., y_{n+1})$ is in $[x]$, we must have that $(y_1, ..., y_n) = (\lambda x_1, ..., \lambda x_n)$, and so $P(y_1, ..., y_n) = \lambda^r P(x_1, ..., x_n)$, which is zero if and only if $P(x_1, ..., x_n)$ is zero. □

We can thus talk meaningfully about the zero set of a homogeneous polynomial in projective space. The use for this is that projective space effectively adds 'points at infinity' to our space; where, in the past, we would have some lines going off to infinity, now we can keep track of what happens at those points.

We can take a curve defined by a polynomial $P(x, y)$ with maximum degree n in affine 2-space (such as our chosen elliptic curve) and embed it into projective space by considering the zero set of the homogeneous polynomial $z^n P(x/z, y/z)$. A quick check demonstrates that all the zs in the denominator cancel, and the result is a homogeneous polynomial of degree n. Moreover, letting $z = 1$ recovers our original polynomial; as such, the 'affine zeros' will remain the same (we'll just add some zeros 'at infinity', or when $z = 0$). This notion becomes slightly more complex in more dimensions.

Definition 6.3.7. A **projective variety** is the zero set of some collection of homogeneous polynomials in projective space.

Remark 6.3.8. A far better way of proving that the group law we've defined is actually a group law is to demonstrate that it agrees with something called the *divisor class group* of the elliptic curve. For more information, see [**Hart**].

6.4 Putting it All Together

Now that we've developed some machinery, we get to play! Here's the gist of what we do.

1. Take an **elliptic curve** defined by a polynomial equation $P(x, y)$ over a **finite field** k (for computability).

2. Pick a **base point** for our elliptic curve.

3. Embed this curve into **projective space** using the homogeneous polynomials associated to $P(x, y)$.

4. We now have a **group** based on our geometric operation outlined in the introduction: Find collinear points, reflect across the y-axis. The identity of this group is the point at infinity, and the inverse of a point $[(x, y, z)]$ is $[(x, -y, z)]$ (this should be clear from geometry). Associativity for this group law, as well as closure (that adding two points on the curve gives you another point on the curve) are complicated and technical results that I don't completely understand at this point, so I'll spare you the details for now - maybe next time!

5. Pick a **private key** (some integer n)

6. Now add the base point *to itself* n times - where n is your private key. The first of these additions is a little strange - adding a point to itself. The idea here is that (if we think in the real numbers for a moment) you take the *tangent* to this point (the limit as another point gets closer and closer to the base point). This can be made precise for a finite field using algebraic geometry, but the formalism requires far too much machinery for the scope of this talk. Suffice it to say, we obtain a point which is not the base point, and we continue with the operation, finding collinear points. There are explicit ways to compute this; see [**Ly**]. We omit these for brevity. This generates your **public key**.

7. If one was to simply add a point to itself over and over again, the time complexity would be prohibitive; we can, instead, repeatedly double the point and add these doubles to obtain the required number of repeated additions.

In order to have a large enough keyspace to make this system worthwhile, people selecting these curves and base points try to chose a base point which maximizes the size of the **cyclic group** generated by that base point. If the cyclic group is too small, it would be relatively easy to determine how many times you multiplied the base point together to get your public key. However, by picking a base point (and elliptic curve) which together lead to a large enough cyclic group, the system is made secure.

In addition to giving us our identity, it's my understanding that embedding into projective space somehow eliminates an inversion operation of a finite field element in the group law, thereby making the point addition far faster (finding inverses in a finite field is slow).

6.5 Gaps in this Picture

Although our image is much more rigorous now, we're still missing three main ideas, namely

- How do we compute the group law? (The answer is some ugly computation).
- Does this operation I've outlined *actually* define a group? (The answer is yes, but proving this is complicated. In particular, associativity is annoying, and there are some *very* subtle aspects to proving closure.).
- Computing $2P$ - how do we add a point to itself. The answer is using the tangent line to that point; showing that the tangent line is a well defined notion (and constructing the correct notion of a tangent line) is also rather subtle and interesting in it's own right - we can't take derivatives in the standard way, because we are not working in anything that can be related to the real numbers.

For more thoughts on these issues take a look at [**Ly**] and [**Hart**].

6.6 Computations with TinyEC

I wrote a quick script that I’ll try to run for you all in python. Credit to [**Nak**] for some of this code. The key here is recommended in [**FIPS**], and I believe still extremely secure (I chose the highest level of security).

```
#!/usr/bin/python

from tinyec import registry
import random

curve = registry.get_curve(’secp521r1’) # This is the largest prime field key
                                        # recommended by the NSA as of
                                        # recentlyish.
print("There are", curve.field.h, "cyclic groups associated with this field.")
print("""The order of the cyclic group generated on this curve by this
         base point is:,""", curve.field.n)
privKey = random.randint(0, curve.field.n) # random isn’t secure but it’s fine.
pubKey = privKey * curve.g
print("My public key is:", pubKey)
print("I check that my private key works, and obtain:", privKey * curve.g)
print("I can’t read that. Is it equal:", privKey * curve.g == pubKey)
print("My private key is (sshhh, don’t tell):", privKey)
```

As a bonus... What does an elliptic curve over a finite field look like? For this we go back to our friend $y^2 = x^3 - 3x$, who looked so nice over the reals. Turns out when you plot this curve over a finite field, it looks a little disconnected (see Figure 6.4 for the final product, and below for the code).

```
import numpy
import matplotlib.pyplot as plt

p = 257 # To irritate the programmers
grid = numpy.zeros((p, p))
for i in range(0, p):
    for j in range(0, p):
        if (i**2)%p == (j**3-3*j)%p:
            grid[i][j] = 1
for j in range(0, p//2):
    grid[p, j] = 1
```

```
plt.imshow(grid, interpolation="nearest")
plt.show()
```

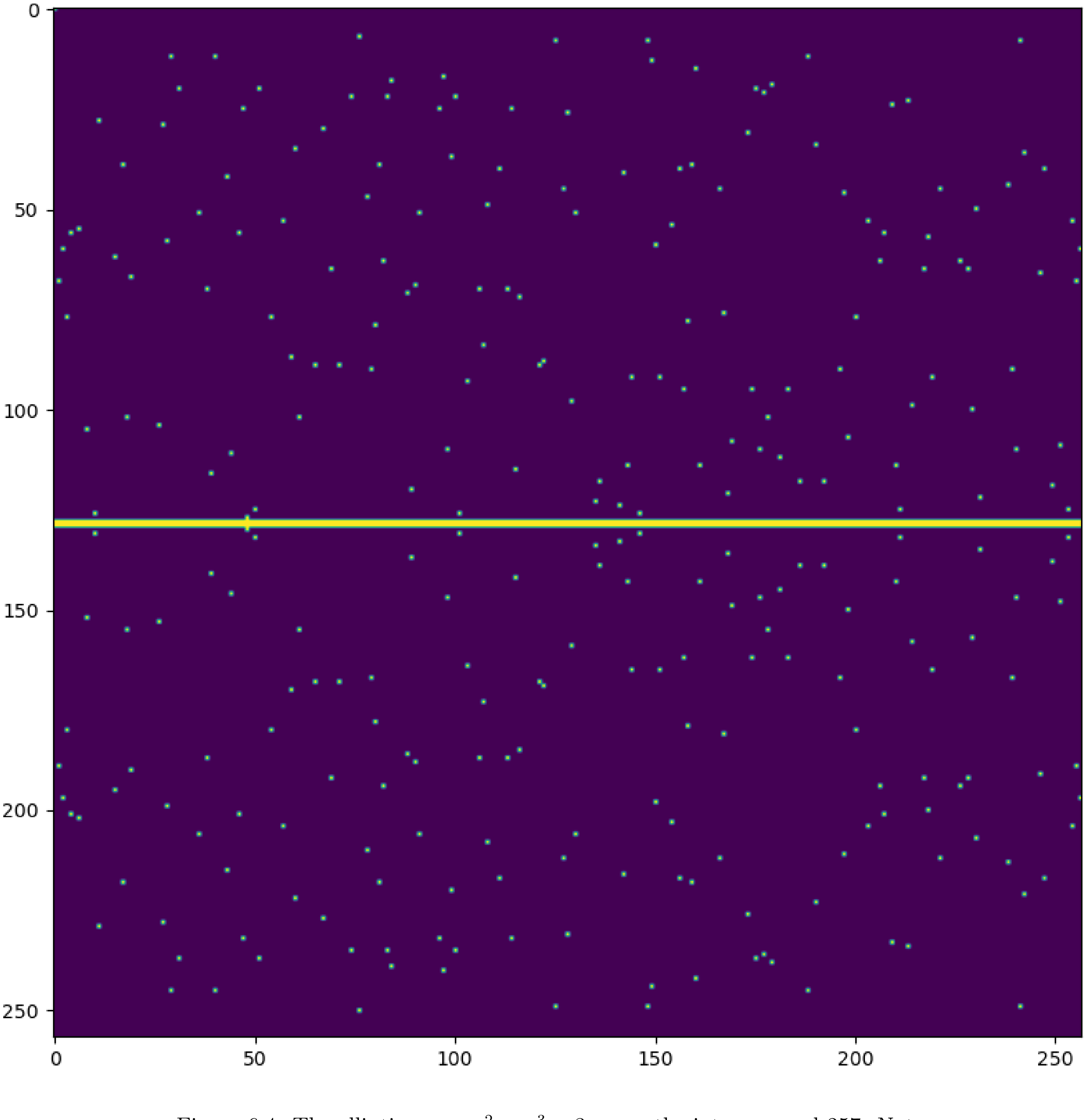


Figure 6.4: The elliptic curve $y^2 = x^3 - 3x$ over the integers mod 257. Note the symmetry across the line in the middle (which is actually the 0 axis), and think back to our group operation, which involved reflecting across the zero-axis...

Bibliography

[DF04] David Steven Dummit and Richard M Foote. *Abstract algebra.* John Wiley & Sons, Danvers, 2004.

[GG15] Steven D. Galbraith and Pierrick Gaudry. Recent progress on the elliptic curve discrete logarithm problem. *Designs, Codes and Cryptography*, 78:51–72, 11 2015.

[Har11] Robin Hartshorne. *Algebraic geometry.* Springer, New York ; London, 2011.

[Lyn] Ben Lynn. Elliptic curves - elliptic curves.

[Nak18] Svetlin Nakov. Github - nakov/practical-cryptography-for-developers-book: Practical cryptography for developers: Hashes, mac, key derivation, dhke, symmetric and asymmetric ciphers, public key cryptosystems, rsa, elliptic curves, ecc, secp256k1, ecdh, ecies, digital signatures, ecdsa, eddsa, 2018.

[oST24] National Institute of Standards and Technology. Security requirements for cryptographic modules. Technical Report Federal Information Processing Standards (FIPS) Publication 186-4, U.S. Department of Commerce, Washington, D.C., 2024.