Lecture Notes on

Undergraduate Math

Kevin Zhou

kzhou7@gmail.com

These notes are a review of the basic undergraduate math curriculum, focusing on the content most relevant for physics. The primary sources were:

- Oxford's Mathematics lecture notes, particularly notes on M2 Analysis, M1 Groups, A2 Metric Spaces, A3 Rings and Modules, A5 Topology, and ASO Groups. The notes by Richard Earl are particularly clear and written in a modular form.
- Rudin, *Principles of Mathematical Analysis*. The canonical introduction to real analysis; terse but complete. Presents many results in the general setting of metric spaces rather than $\mathbb{R}$.
- Ablowitz and Fokas, *Complex Variables*. Quickly covers the core material of complex analysis, then introduces many practical tools; indispensable for an applied mathematician.
- Artin, *Algebra*. A good general algebra textbook that interweaves linear algebra and focuses on nontrivial, concrete examples such as crystallography and quadratic number fields.
- David Skinner's lecture notes on Methods. Provides a general undergraduate introduction to mathematical methods in physics, a bit more careful with mathematical details than typical.
- Munkres, *Topology*. A clear, if somewhat dry introduction to point-set topology. Also includes a bit of algebraic topology, focusing on the fundamental group.
- Renteln, *Manifolds, Tensors, and Forms*. A textbook on differential geometry and algebraic topology for physicists. Very clean and terse, with many good exercises.

Some sections are quite brief, and are intended as a telegraphic review of results rather than a full exposition. The most recent version is here; please report any errors found to kzhou7@gmail.com.

Contents

1 Metric Spaces

1.1 Definitions

We begin with some basic definitions. Throughout, we let E be a subset of a fixed set X.

- A set X is a metric space if it is has a distance function $d(p, q)$ which is positive definite (except for $d(p, p) = 0$), symmetric, and satisfies the triangle inequality.
- A neighborhood of p is the set $N_r(p)$ of all q with $d(p, q) < r$ for some radius $r > 0$.

  Others define a neighborhood as any set that contains one of these neighborhoods, which are instead called "the open ball of radius r about p". This is equivalent for proofs; the important part is that neighborhoods always contain points "arbitrarily close" to p.
- A point p is a limit point of E if every neighborhood of p contains a point $q \neq p$ in E. If p is not a limit point but is in E, then p is an isolated point.
- E is closed if every limit point of E is in E. Intuitively, this means E "contains all its edges". The closure $\overline{E}$ of E is the union of E and the set of its limit points.
- A point p is an interior point of E if there is a neighborhood N of p such that $N \subset E$. Note that interior points must be in E itself, while limit points need not be.
- E is open if every point of E is an interior point of E. Intuitively, E "doesn't have edges".
- E is bounded if there exists M and q so that $d(p, q) < M$ for all $p \in E$.
- E is dense in X if every point of X is a limit point of E or a point of E, or both.
- The interior E^0 of E is the set of all interior points of E, or equivalently the union of all open sets contained in E.

Example. We give some simple examples in $\mathbb{R}$ with the usual metric.

- Finite subsets of $\mathbb{R}$ cannot have any limit points or interior points, so they are trivially closed and not open.
- The set $(0, 1]$. The limit points are $[0, 1]$, so the set is not closed. The interior points are $(0, 1)$, so the set is not open.
- The set of points $1/n$ for $n \in \mathbb{Z}$. The single limit point is 0, so the set is not closed.
- All points. This set is trivially open and closed.
- The interval $[1, 2]$ in the restricted space $[1, 2] \cup [3, 4]$. This is both open and closed. Generally, this happens when a set contains "all of a connected component".

As seen from the last example above, whether a set is closed or open depends on the space, so if we wanted to be precise, we would say "closed in X" rather than just "closed".

Example. There are many examples of metrics besides the usual one.

- For any set S, we may define the discrete metric

$$d(x,y) = \begin{cases} 0 & x = y, \\ 1 & x \neq y. \end{cases}$$

  Note that in this case, the closed ball of radius 1 about p is not the closure of the open ball of radius 1 about p.

- A metric on a vector space can be defined from an inner product, which can in turn be defined from a norm. (However, a norm does not necessarily give a valid inner product.) For example, for continuous functions $f : [a,b] \to \mathbb{R}$ we have the inner product

$$\langle f, g \rangle = \int_a^b f(t)g(t)\,dt$$

  which gives the norm $\|f\| = \sqrt{\langle f, f \rangle}$ and the metric

$$d_2(f,g) = \|f - g\| = \sqrt{\int_a^b (f(t) - g(t))^2\,dt}.$$

- Alternatively, we could use the metric

$$d_\infty(f,g) = \sup_{x \in [a,b]} |f(x) - g(x)|.$$

  These are both special cases of a range of metrics.

We now consider some fundamental properties of open and closed sets.

- E is open if and only if its complement E^c is closed.

  Heuristically, this proof works because open and closed are 'for all' and 'there exists' properties, and taking the complement swaps them. Specifically, if q is an interior point of E, then E contains all points arbitrarily close to q. But if q is a limit point of E^c, there exist points arbitrarily close to q that are in E^c. Only one of these can be true, giving the result.

- Arbitrary unions of open sets are open, because interior points stay interior points when we add more points. By taking the complement, arbitrary intersections of closed sets are closed.

- Finite intersections of open sets are open, because we can take intersections of the relevant neighborhoods. This breaks down for infinite intersections because the neighborhoods can shrink down to nothing, e.g. let $E_n = (-1/n, 1/n)$). By taking the complement, finite unions of closed sets are closed. Infinite unions don't work because they can create new limit points.

Prop. The closure $\overline{E}$ is the smallest closed set containing E.

Proof. The idea behind the proof of closure is that all limit points of $\overline{E}$ must be limit points of E. Formally, let p be a limit point of $\overline{E}$. Then any neighborhood $N \supset p$ contains some $q \in \overline{E}$. Since neighborhoods are open, N must contain a neighborhood N' of q, which then must contain some element of q. Thus p is a limit point of E.

To see that $\overline{E}$ is the smallest possibility, note that adding more points never subtracts limit points. Therefore any closed $F \supset E$ must contain all the limit points of E.

Prop. For $Y \subset X$, the open sets E of Y are precisely $Y \cap G$ for open sets G of X.

Proof. If G is open, then moving to the smaller space Y will keep it open. Now consider the converse. Starting with $E \subset Y$, we construct G by taking the union of all neighborhoods (in X) of points in E. Then G is an open set of X because it is the union of open sets. Moreover, $E = Y \cap G$ because E is open.

Note. Topological spaces further abstract by throwing away the metric but retaining the structure of the open sets. A topological space is a set X along with a set T of subsets of X, called the open sets of X, such that T is closed under all unions and finite intersections, and contains both X itself and the null set. The closed sets are defined as the complements of open sets. The rest of our definitions hold as before, if we think of a neighborhood of a point x as any open set containing x.

For a subspace $Y \subset X$, we use the above proposition in reverse, defining the open sets in Y by those in X. The resulting topology is called the subspace topology.

Note. An isometry between two metric spaces X and Y is a bijection that preserves the metric. However, topological properties only depend on the open set structure, so we define a homeomorphism to be a bijection that is continuous with a continuous inverse; this ensures that it induces a bijection between the topologies of X and Y. As we'll see below, many important properties such as continuity depend only on the topology, so we are motivated to find topological invariants, properties preserved by homeomorphisms, to classify spaces.

1.2 Compactness

Compactness is a property that generalizes "finiteness" or "smallness". Though its definition is somewhat unintuitive, it turns out to be quite useful.

- An open cover of a set E in a metric space X is a set of open sets G_i of X so that their union contains E. For example, one open cover of E could be the set of all neighborhoods of radius r of every point in E.
- K is compact if every open cover of K contains a finite subcover. For example, all finite sets are compact. Since we only made reference to the open sets, not the metric, compactness is a topological invariant.
- Let $K \subset Y \subset X$. Then K is compact in X iff it is compact in Y, so we can refer to compactness as an absolute property, independent of the containing space.

  Proof: essentially, this is because we can transfer open covers of K in Y and of K in X back and forth, using the above theorem. Thus if we can pick a finite subcover in one, we can pick the analogous subcover in the other.
- All compact sets are closed. Intuitively, consider the interval $(0, 1/2)$ in $\mathbb{R}$. Then the open cover $(1/n, 1)$ has no finite subcover; we can get 'closer and closer' to the open boundary.

  Proof: let $K \subset X$ be compact; we will show K^c is open. Fixing $p \in K^c$, define the open cover consisting of the balls with radius $d(p, q)/2$ for all $q \in K$. Consider a finite subcover and let $d_{\min}$ be the minimum radius of any ball in it. Then there is a neighborhood of radius $d_{\min}/2$ of p containing no points of K.

- All compact subsets of a metric space are bounded. This follows by taking an open cover consisting of larger and larger balls.
- Closed subsets of compact sets are compact.

  Proof: let $F \subset K \subset X$ with F closed and K compact. Take an open cover of F, and add F^c to get an open cover of K. Then a finite subcover of K yields a finite subcover of F.
- Intersections of compact sets are compact. This follows from the previous two results.

Note. The overall intuition found above is that compactness is a notion of 'smallness'. An open boundary is not 'small' because it is essentially the same as a boundary at infinity, from the standpoint of open covers. We see that compactness is useful for proofs because the finiteness of a subcover allows us to take least or greatest elements; we show some more examples of this below.

Example. Let K be a compact metric space. Then for any $\epsilon > 0$, there exists an N so that every set of N distinct points in K includes at least two points with distance less than ϵ between them. To show this, consider the open cover consisting of all neighborhoods of radius $\epsilon/2$. Then there's a finite open subcover, with M elements, centered at points p_i. For $N > M$, we are done by the pigeonhole principle.

Example. Let K be a compact metric space. Then K has a subset that is dense and at most countable. To prove this, consider the open cover of all neighborhoods of radius 1. Take a finite subcover centered at a set of points P_1. Then points in P_1 are within a distance of 1 from any point in K. Next construct P_2 using radius $1/2$, and so on. Then $P = \bigcup_n P_n$ is dense and at most countable.

Lemma. All k-cells in $\mathbb{R}^k$ are compact.

Proof. This is the key lemma that uses special properties of $\mathbb{R}$. For simplicity, we consider the case $k = 1$, showing that all intervals $[a, b]$ are compact. Let $\mathcal{U}$ be an open cover of $[a, b]$ and define

$$W = \{x \in [a, b] : \text{finite subcover of } \mathcal{U} \text{ exists for } [a, x]\}, \quad c = \sup(W).$$

First we show $c \in W$. Let $c \in U \in \mathcal{U}$. Since U is open, it includes $(c - \delta, c + \delta)$ for some $\delta > 0$. On the other hand, by the definition of the supremum there must be some element $w \in W$ inside this range. Then we have an finite subcover of $[a, c]$ by taking U along with the finite subcover for $[a, w]$.

Next, by a similar argument, if $x \in W$ and $x < b$, there must be $\delta > 0$ so that $x + \delta \in W$. Hence we have a contradiction unless $c = b$, giving the result. The generalization to arbitrary k is similar. Note that we used the least upper bound property of $\mathbb{R}$ by assuming $c \in \mathbb{R}$.

Theorem (Heine-Borel)**.** For any $E \subset \mathbb{R}^k$, E is closed and bounded if and only if it is compact.

Proof. We have already shown the reverse direction above. For the forward direction, note that if E is bounded it is a subset of a k-cell, and closed subsets of compact spaces are compact.

1.3 Sequences

We begin by defining convergence of a sequence.

- A sequence (p_n) in a metric space X converges to a point $p \in X$ if, for every $\epsilon > 0$, there is an integer N so that if $n \geq N$, then $d(p_n, p) < \epsilon$. This may also be written

$$\lim_{n\to\infty} p_n = p.$$

  If (p_n) doesn't converge, it diverges. Note that convergence depends on X. If X is "missing" the right point, then an otherwise convergent sequence may diverge.

- More generally, in the context of a topological space, a sequence (p_n) converges to $p \in X$ iff every neighborhood of p contains all but finitely many of the p_n.

- Sequences can only converge to one point; this is proven by considering neighborhoods of radius $\epsilon/2$ and using the triangle inequality.

- If a sequence converges, it must be bounded. This is because only finitely many points lie outside any given neighborhood of the limit point p, and finite sets are bounded.

- If $E \subset X$ and p is a limit point of E, then there is a sequence (p_n) in E that converges to p. Conversely, a convergent sequence with range in E converges to a point in $\overline{E}$.

- A topological space is sequentially compact if every infinite sequence has a convergent subsequence, and compactness implies sequential compactness.

  To see this, let (x_k) be a sequence and let

$$X_n = \overline{\{x_k : k > n\}}, \quad U_n = M \setminus X_n.$$

  Assuming the space is not sequentially compact, the intersection of all the X_n is empty, so the U_n are an open cover with no finite subcover, so the space is not compact.

- It can be shown that sequential compactness in a metric space implies compactness, though this does not hold for a general topological space.

Example. Consider the set of bounded real sequences ℓ^∞. The unit cube

$$C = \{(x_k) : |x_k| \leq 1\}$$

is closed and bounded, but it is not compact, because the sequence

$$e_1 = (1, 0, \ldots), \quad e_2 = (0, 1, 0, \ldots), \quad e_3 = (0, 0, 1, 0, \ldots)$$

has no convergent subsequence.

Next, we specialize to Euclidean space, recovering some familiar results.

- Bounded monotonic sequences of real numbers converge to their least upper/greatest lower bounds, essentially by definition.

- If (s_n) converges to s and (t_n) converges to t, then

  $$(s_n + t_n) \to s + t \quad (cs_n) \to cs \quad (c + s_n) \to c + s \quad (s_n t_n) \to st \quad (1/s_n) \to 1/s \text{ (if } s_n \neq 0).$$

  The proofs are easy except for the last two, where we must work to bound the error. For the fourth, we can factor

  $$s_n t_n - st = (s_n - s)(t_n - t) + s(t_n - t) + t(s_n - s)$$

  To get an $O(\epsilon)$ error on the left, we must use a $\sqrt{\epsilon}$ error for the first term.

- If $s_n \leq t_n$ for all n, then $s \leq t$. To prove it, consider $(t_n - s_n)$. The range is bounded below by 0, so the closure of the range can't contain any negative numbers.

- All of the above works similarly for vectors in $\mathbb{R}^k$, and limits can be taken componentwise; the proof is to just use $\epsilon/\sqrt{k}$. In particular, $x_n \cdot y_n \to x \cdot y$, since the components are just multiplications.

- Since compactness implies sequential compactness, we have the Bolzano-Weierstrass theorem, which states that every bounded sequence in $\mathbb{R}^k$ has a convergent subsequence.

Next, we introduce Cauchy sequences. They are useful because they allow us to say some things about convergence without specifying a limit.

- A sequence (p_n) in a metric space X is Cauchy if, for every $\epsilon > 0$ there is an integer N such that $d(p_n, p_m) < \epsilon$ if $n, m \geq N$. Note that unlike regular convergence, this definition depends on the metric structure.

- The diameter of E is the supremum of the set of distances $d(p, q)$ with $p, q \in E$. Then (p_n) is Cauchy iff the limit of the diameters d_n of the sequences $p_n, p_{n+1}, \ldots$ is zero.

- All convergent sequences are Cauchy, because if we get within $\epsilon/2$ of the limit, then the points themselves are within ϵ of each other.

- A metric space in which every Cauchy sequence converges is complete; intuitively, these spaces have no 'missing limit points'. Moreover, every closed subset E of a complete metric space X is complete, since Cauchy sequences in E are also Cauchy sequences in X.

- Compact metric spaces are complete. This is because compactness implies sequential compactness, and a convergent subsequence of a Cauchy sequence is sufficient to guaranteed the Cauchy sequence is convergent.

- The space $\mathbb{R}^k$ is complete, because all Cauchy sequences are bounded, and hence inside a k-cell. Since k-cells are compact, we can apply the previous fact. The completeness of $\mathbb{R}$ is one of its most important properties, and it is what suits it for doing calculus better than $\mathbb{Q}$.

- Completeness is not a topological invariant, because $(0, 1)$ is not complete while $\mathbb{R}$ is; it depends on the details of the metric. However, the property of "complete metrizability" is a topological invariant; a topological space is completely metrizable if there exists a metric while yields the topology under which the space is complete.

Finally, we introduce some convenient notation for limits.

- For a real sequence, we write $s_n \to \infty$ if, for every real M, there is an integer N so that every term after s_n is at least M. We'll now count $\pm\infty$ as a possible subsequential limit.
- Denote E as the set of subsequential limits of a real sequence (s_n), and write

$$s^* = \limsup_{n\to\infty} s_n = \sup E, \quad s_* = \liminf_{n\to\infty} s_n = \inf E.$$

  It can be shown that E is closed, so it contains s^* and s_*. The sequence converges iff $s^* = s_*$.

Example. For a sequence containing all rationals in arbitrary order, every real number is a subsequential limit, so $s^* = \infty$ and $s_* = -\infty$. For the sequence $a_k = (-1)^k(k+1)/k$, we have $s^* = 1$ and $s_* = -1$.

The notation we've defined above will be useful for analyzing series. For example, a series might contain several geometric subsequences; for convergence, we care about the one with the largest ratio, which can be extracted with lim sup.

1.4 Series

Given a sequence (a_n), we say the sum of the series $\sum_n a_n$, if it exists, is

$$\lim_{n\to\infty} s_n, \quad s_n = a_1 + \ldots + a_n.$$

That is, the sum of a series is the limit of its partial sums. We quickly review convergence tests.

- Cauchy convergence test: since $\mathbb{R}$ is complete, we can replace convergence with the Cauchy property. Then $\sum_n a_n$ converges iff, for every $\epsilon > 0$, there is an integer N so that for all $m \geq n \geq N$, $|a_n + \ldots a_m| \leq \epsilon$.
- Limit test: taking $m = n$, the above becomes $|a_n| \leq \epsilon$, which means that if $\sum_n a_n$ converges, than $a_n \to 0$. This is a much weaker version of the above.
- A monotonic sequence converges iff it's bounded. Then a series of nonnegative terms converges iff the partial sums form a bounded sequence.
- Comparison test: if $|a_n| < c_n$ for $n > N_0$ for a fixed N_0, and $\sum_n c_n$ converges, then $\sum_n a_n$ converges. We prove this by plugging directly into the Cauchy criterion.
- Divergence test: taking the contrapositive, we can prove a series diverges if we can bound it from below by a divergent series.
- Geometric series: for $0 \leq x < 1$, $\sum_n x^n = 1/(1-x)$, so the series $a_n = x^n$ converges. To prove this, write the partial sums using the geometric series formula, then take the limit explicitly.
- Cauchy condensation test: let $a_1 \geq a_2 \geq \ldots \geq 0$. Then $\sum_n a_n$ converges iff $\sum 2^n a_{2^n}$ does. This surprisingly implies that we only need a small number of the terms to determine convergence.

  Proof: since the terms are all nonnegative, convergence is equivalent to the partial sums being bounded. Now group the terms a_n two different ways:

$$a_1 + (a_2 + a_3) + \ldots + (a_{2^k} + \ldots + a_{2^{k+1}-1}) \leq a_1 + 2a_2 + \ldots + 2^k a_{2^k} = \sum_1^k 2^n a_{2^k},$$

$$a_1 + a_2 + (a_3 + a_4) + \ldots + (a_{2^{k-1}+1} + \ldots + a_{2^k}) \geq \frac{1}{2}a_1 + a_2 + 2a_4 + \ldots = \frac{1}{2}\sum_1^k 2^n a_{2^k}.$$

Then the sequences of partial sums of a_n and $2^n a_{2^n}$ are within a constant multiple of each other, so each converges iff the other does. As an application, $\sum_n 1/n$ diverges.

Next, we apply our basic tests to more specific situations.

- p-series: $\sum_n 1/n^p$ converges iff $p > 1$.

  Proof: for $p \leq 0$, the terms don't go to zero. Otherwise, apply Cauchy condensation, giving the series $\sum_k 2^k/2^{kp} = \sum_k 2^{(1-p)^k}$, and use the geometric series test.

- Ratio test: let $\sum a_n$ have $a_n \neq 0$. Then the series converges if $\alpha = \limsup_{n\to\infty} |a_{n+1}/a_n| < 1$. The proof is simply by comparison to a geometric series.

- Root test: let $\alpha = \limsup_{n\to\infty} \sqrt[n]{|a_n|}$. Then $\sum_n a_n$ converges if $\alpha < 1$ and diverges if $\alpha > 1$. The proof is similar to the ratio test: for sufficiently large n, we can bound the terms by a geometric series.

- Dirichlet's theorem: let $A_n = \sum_{k=0}^n a_k$ and $B_n = \sum_{k=0}^n b_k$. Then if (A_n) is bounded and (b_n) is monotonically decreasing with $\lim_{n\to\infty} b_n = 0$, then $\sum a_n b_n$ converges.

  Proof: we use 'summation by parts',

  $$\sum_{n=p}^{q} a_n b_n = \sum_{n=p}^{q-1} A_n(b_n - b_{n+1}) + A_q b_q - A_{p-1} b_p.$$

  The result follows immediately by the comparison test.

- Alternating series test: if $|c_i| \geq |c_{i+1}|$ and $\lim_{n\to\infty} c_n = 0$ and the c_i alternate in sign, then $\sum_n c_n$ converges.

  Proof: this is a special case of Dirichlet's theorem; it also follows from the Cauchy criterion.

Example. The series $\sum_{n>0} 1/(n(\log n)^p)$ converges iff $p > 1$, by the Cauchy condensation test. The general principle is that Cauchy condensation can be used to remove a layer of logarithms, or convert a p-series to a geometric series.

Example. Tricking the ratio test. Consider the series that alternates between 3^{-n} and 2^{-n}. Then half of the ratios are large, so the ratio test is inconclusive. However, the root test works, giving $\alpha = 1/2 < 1$. Essentially, the two tests do the same thing, but the root test is more powerful because it doesn't just look at 'local' information.

Note. The ratio and root test come from the geometric series test, which in turn comes from the limit test. That is, fundamentally, they aren't doing anything deeper than seeing if the terms blow up. The only stronger tools we have are the Cauchy condensation test, which gives us the p-series test, and Dirichlet's theorem.

Example. The Fourier p-series is defined as

$$\sum_{k=1}^{\infty} \frac{\cos(kx)}{k^p}.$$

By comparison with p-series, it converges for $p > 1$. For $0 \leq p \leq 1$, use the Dirichlet theorem with $a_n = \cos nx$ and $b_n = 1/n^p$. Using geometric series, we can show that (A_n) is bounded as long as x is not a multiple of 2π, giving convergence.

Next, we extend to the complex numbers and consider power series; note that our previous results continue to work when the absolute value is replaced by the complex norm. Given a sequence (c_n) of complex numbers, the series $\sum_n c_n z^n$ is called a power series.

Theorem. Let $\alpha = \limsup_{n\to\infty} \sqrt[n]{|c_n|}$. Then the power series $\sum_n c_n z^n$ converges when $|z| < R$ and diverges when $|z| > R$, where $R = 1/\alpha$ is called the radius of convergence.

Proof. Immediate by the root test.

Example. We now give some example applications of the theorem.

- The series $\sum_n z^n$ has $R = 1$. If $|z| = 1$, the series diverges by the limit test.
- The series $\sum_n z^n/n$ has $R = 1$. We've already shown it diverges if $z = 1$. However, it converges for all other z on the boundary, as this is just a variant of the Fourier p-series.
- The series $\sum_n z^n/n^2$ has $R = 1$ and converges for all z on the boundary by the p-series test.
- The series $\sum_n z^n/n!$ has $R = \infty$ by the ratio test.

As stated earlier, divergence of power series is not subtle; the terms become unbounded.

We say the series $\sum_n a_n$ converges absolutely if $\sum_n |a_n|$ converges.

- Many properties that intuitively hold for convergent series really require absolute convergence; often the absolute values appear from a triangle-inequality argument.
- All absolutely convergent series are convergent because $|\sum a_i| \leq \sum |a_i|$ by triangle inequality.
- Power series are absolutely convergent within their radius of convergence, because the root test only considers absolute values.

Prop. Let $\sum_n a_n = A$ and $\sum_n b_n = B$ with $\sum_n a_n$ converging absolutely. Then the product series $\sum_n c_n$ defined by

$$c_n = \sum_{k=0}^{n} a_k b_{n-k}$$

converges to AB. This definition is motivated by multiplication of power series.

Proof. Let $\beta_n = B_n - B$. We'll pull out the terms we want from C_n, plus an error term,

$$C_n = a_0 b_0 + \ldots + (a_0 b_n + \ldots + a_n b_0) = a_0 B_n + \ldots + a_n B_0.$$

Pulling out $A_n B$ gives

$$C_n = A_n B + a_0 \beta_n + \ldots + a_n \beta_0 \equiv A_n B + \gamma_n.$$

We want to show that $\gamma_n \to 0$. Let $\alpha = \sum_n |a_n|$. For some $\epsilon > 0$, choose N so that $|\beta_n| \leq \epsilon$ for all $n \geq N$. Then separate the error term into

$$\gamma_n \leq |\beta_0 a_n + \ldots \beta_N a_{n-N}| + |\beta_{N+1} a_{n-N+1} + \ldots + \beta_n a_0|$$

The first term goes to zero as $n \to \infty$, and the second is bounded by $\epsilon\alpha$. Since ϵ was arbitrary, we're done.

Note. Series that converge but not absolutely are conditionally convergent. The Riemann rearrangement theorem states that for such series, the terms can always be reordered to approach any desired limit; the idea is to take just enough positive terms to get over it, then enough negative terms to get under it, and alternate.

2 Real Analysis

2.1 Continuity

We begin by defining limits in the metric spaces X and Y.

- Let f map $E \subset X$ into Y, and let p be a limit point of E. Then we write

$$\lim_{x \to p} f(x) = q$$

  if, for every $\epsilon > 0$ there is a $\delta > 0$ such that for all $x \in E$, with $0 < d_X(x, p) < \delta$, we have $d_Y(f(x), q) < \epsilon$. We also write $f(x) \to q$ as $x \to p$.

- This definition is completely indifferent to $f(p)$ itself, which could even be undefined.

- In terms of sequences, an equivalent definition of limits is that

$$\lim_{n \to \infty} f(p_n) = q$$

  for every sequence $(p_n) \in E$ so that $p_n \neq p$ and $\lim_{n \to \infty} p_n = p$.

- By the same proofs as for sequences, limits are unique, and in $\mathbb{R}$ they add/multiply/divide as expected.

We now use this limit definition to define continuity.

- We say that f is continuous at p if

$$\lim_{x \to p} f(x) = f(p).$$

  In the case where p is not a limit point of the domain E, we say f is continuous at p. If f is continuous at all points of E, then we say f is continuous on E.

- None of our definitions care about E^c, so we'll implicitly restrict X to the domain E for all future statements.

- If f maps X into Y, and g maps range $F \subset Y$ into Z, and f is continuous at p and g is continuous at $f(p)$, then $g \circ f$ is continuous at p. We prove by using the definition twice.

- Continuity for functions $f : \mathbb{R} \to \mathbb{R}$ is preserved under arithmetic operations the way we expect, by the results above. The function $f(x) = x$ is continuous, as we can choose $\delta = \epsilon$. Hence polynomials and rational functions are continuous. The absolute value function is also continuous; we can choose $\delta = \epsilon$ by the triangle inequality. This can be generalized to functions from $\mathbb{R}$ to $\mathbb{R}^k$, which are continuous iff all the components are.

Now we connect continuity to topology. Note that if we were dealing with a topological space rather than a metric space, the following condition would be used to define continuity.

Theorem. A map $f : X \to Y$ is continuous on X iff $f^{-1}(V)$ is open in X for all open sets V in Y.

Proof. The key idea is that every point of an open set is an interior point. Assume f is continuous on X, and let $p \in f^{-1}(V)$ and $q = f(p)$. The continuity condition states that

$$f(N_\delta(p)) \subset N_\epsilon(q)$$

for some δ, given any ϵ. Choosing ϵ so that $N_\epsilon(q) \subset V$, this shows that p is an interior point of $f^{-1}(V)$, giving the result. The converse is similar.

Corollary. If f is continuous, then f^{-1} takes closed sets to closed sets; this follows from taking the complement of the previous theorem.

Corollary. A function f is continuous if, for every subset $S \subset X$, we have $f(\overline{S}) \subset \overline{f(S)}$. This follows from the previous corollary, and exhibits the intuitive notion that continuous functions keep nearby points together.

Example. Using the definition of continuity, it is easy to show that the circle $x^2 + y^2 = 1$ is closed, because this is the inverse image of the closed set $\{1\}$ under the continuous function $f(x, y) = x^2 + y^2$. Similarly, the region $x^2 + xy + y^2 < 1$ is open, and so on. In general continuity is one of the most practical ways to show that a set is open or closed.

We now relate continuity to compactness.

- Let $f : X \to Y$ be continuous on X. Then if X is compact, $f(X)$ is compact.

  Proof: take an open cover $\{V_\alpha\}$ of $f(X)$. Then $\{f^{-1}(V_\alpha)\}$ is an open cover of X. Picking a finite subcover and applying f gives a finite subcover of $f(X)$.

- EVT: let f be a continuous real function on a compact metric space X, and let

  $$M = \sup_{p \in X} f(p), \quad m = \inf_{p \in X} f(p).$$

  Then there exist points $p, q \in X$ so that $f(p) = M$ and $f(q) = m$.

  Proof: let $E = f(X)$. Then E is compact, so closed and bounded. By the definition of sup and inf, we know that M and m are limit points of E. Since E is closed, E must contain them.

- Compactness is required for the EVT because it rules out asymptotes (e.g. $1/x$ on $(0, \infty)$). This is another realization of the 'smallness' compactness guarantees.

Next, we relate continuity to connectedness, another topological property.

- A metric space X is disconnected if it may be written as $X = A \cup B$ where A and B are disjoint, nonempty, open subsets of X. We say X is connected if it is not disconnected. Since it depends only on the open set structure, connectedness is a topological invariant.

- The interval $[a, b]$ is connected. To show this, note that disconnectedness is equivalent to the existence of a closed and open, nonempty proper subset. Let C be such a subset and let $a \in C$ without loss of generality. Define

  $$W = \{x \in [a, b] : [a, x] \subset C\}, \quad c = \sup W.$$

  Then $c \in [a, b]$, which is the crucial step that does not work for $\mathbb{Q}$. We know for any $\epsilon > 0$ there exists $x \in W$ so that $x \in (c - \epsilon, c]$, which implies $[a, c - \epsilon] \subset C$. Since C is closed, this implies $c \in W$. On the other hand, if $x \in C$ and $x < b$, then since C is open, there exists an $\epsilon > 0$ so that $x + \epsilon \in C$. Hence if $c < b$, we have a contradiction, so we must have $c = b$ and $[a, b] = C$.

- More generally, the connected subsets of $\mathbb{R}$ are the intervals, while almost every subset of $\mathbb{Q}$ is disconnected.
- Let $f : X \to Y$ be continuous and one-to-one on a compact metric space X. Then f^{-1} is continuous on Y.

  Proof: let V be open in X. Then V^C is compact, so $f(V^C)$ is compact and hence closed in Y. Since f is one-to-one, $f(V^C) = f(V)^C$, so $f(V)$ is open, giving the result.
- Let $f : X \to Y$ be continuous on X. Then if $E \subset X$ is connected, so is $f(E)$. This is proved directly from the definition of connectedness.
- IVT: let f be a continuous real function defined on $[a, b]$. Then if $f(a) < f(b)$ and $c \in [f(a), f(b)]$, then there exists a point $x \in (a, b)$ such that $f(x) = c$. This follows immediately from the above fact, because intervals are connected.
- A set $S \subset \mathbb{R}^n$ is path-connected if, given any $a, b \in S$ there is a continuous map $\gamma : [0, 1] \to S$ such that $\gamma(0) = a$ and $\gamma(1) = b$.
- Path connectedness implies connectedness. To see this, note that connectedness of S is equivalent to all continuous functions $f : S \to \mathbb{Z}$ being constant. Now consider the map $f \circ \gamma : [0, 1] \to \mathbb{Z}$ for any continuous f. It is continuous, and its domain is connected, so its value is constant and $f(\gamma(0)) = f(\gamma(1))$. Then $f(a) = f(b)$ for all $a, b \in S$.
- All open connected subsets of $\mathbb{R}^n$ are path connected. However, in general connected sets are not necessarily path connected. The standard example is the Topologist's sine curve

  $$X = A \cup B, \quad A = \{(x, \sin(1/x)) : x > 0\}, \quad B = \{(0, y) : y \in \mathbb{R}\}.$$

  The two path components are A and B.

Now we define a stronger form of continuity that'll come in handy later.

- We say $f : X \to Y$ is uniformly continuous on X if, for every $\epsilon > 0$, there exists $\delta > 0$ so that

  $$d_X(p, q) < \delta \text{ implies } d_Y(f(p), f(q)) < \epsilon$$

  for all $p, q \in X$. That is, we can use the same δ for every point. For example, $1/x$ is continuous but not uniformly continuous on $(0, \infty)$ because it gets arbitrarily steep.
- A function $f : X \to Y$ is Lipschitz continuous if there exists a constant $K > 0$ so that

  $$d_Y(f(p), f(q)) \leq K d_X(p, q).$$

  Lipschitz continuity implies uniform continuity, by choosing $\delta = \epsilon/2K$, and can be an easy way to establish uniform continuity.
- Let $f : X \to Y$ be continuous on X. Then if X is compact, f is uniformly continuous on X.

  Proof: for a given ϵ, let δ_p be a corresponding δ to show continuity at the point p. The set of neighborhoods $N_{\delta_p}(p)$ form an open cover of X. Take a finite subcover and let $\delta_{\min}$ be the minimum δ_p used. Then a multiple of $\delta_{\min}$ works for uniform continuity.

Example. The metric spaces $[0,1]$ and $[0,1)$ are not homeomorphic. Suppose that $h : [0,1] \to [0,1)$ is such a homeomorphism. Then the map

$$\frac{1}{1-h(x)}$$

is a continuous, unbounded function on $[0,1]$, which contradicts the IVT.

2.2 Differentiation

In this section we define derivatives for functions on the real line; the situation is more complicated in higher dimensions.

- Let f be defined on $[a,b]$. Then for $x \in [a,b]$, define the derivative

  $$f'(x) = \lim_{t \to x} \frac{f(t)-f(x)}{t-x}$$

  If f' is defined at a point/set, we say f is differentiable at that point/set.

- Note that our definition defines differentiability at all x that are limit points of the domain of f, and hence includes the endpoints a and b. In more general applications, though, we'll prefer to talk about differentiability only on open sets, where we can 'approach from all directions'.

- Differentiability implies continuity, because

  $$f(t) - f(x) = \frac{f(t)-f(x)}{t-x} \cdot (t-x)$$

  and taking the limit $x \to t$ gives zero.

- The linearity of the derivative and the product rule can be derived by manipulating the difference quotient. For example, if $h = fg$, then

  $$\frac{h(t)-h(x)}{t-x} = \frac{f(t)(g(t)-g(x)) + g(x)(f(t)-f(x))}{t-x}$$

  which gives the product rule.

- By the definition, the derivative of 1 is 0 and the derivative of x is 1. Using the above rules gives the power rule, $(d/dx)(x^n) = nx^{n-1}$.

- Chain Rule: suppose f is continuous on $[a,b]$, $f'(x)$ exists at some point $x \in [a,b]$, g is defined on an interval I that contains the range of f, and g is differentiable at $f(x)$. Then if $h(t) = g(f(t))$, then

  $$h'(x) = g'(f(x))f'(x)$$

  To prove this, we isolate the error terms,

  $$f(t) - f(x) = (t-x)(f'(x) + u(t)), \quad g(s) - g(y) = (s-y)(g'(y) + v(s)).$$

  By definition, $u(t) \to 0$ as $t \to x$ and $v(s) \to 0$ as $s \to f(x)$. Now the total error is

  $$h(t) - h(x) = g(f(t)) - g(f(x)) = (t-x)\,(f'(x) + u(t))\,(g'(f(t))) + v(f(x))).$$

  Thus by appropriate choices of ϵ we have the result; note that we need continuity of f to ensure that $f(t) \to f(x)$.

- Inverse Rule: if f has a differentiable inverse f^{-1}, then

$$\frac{d}{dx}f^{-1}(x) = \frac{1}{f'(f^{-1}(x))}$$

  This can be derived by applying the chain rule to $f \circ f^{-1}$.

We now introduce the generalized mean value theorem, which is extremely useful in proofs.

- We say a function $f : X \to \mathbb{R}$ has a local maximum at p if there exists $\delta > 0$ so that $f(q) \leq f(p)$ for all $q \in X$ with $d(p, q) \in \delta$.
- Given a function $f : [a, b] \to \mathbb{R}$, if f has a local maximum at $x \in (a, b)$ and $f'(x)$ exists, then $f'(x) = 0$.

  Proof: sequences approaching from the right give $f'(x) \leq 0$, because the difference quotient is nonnegative once we get within δ of x. Similarly, sequences from the left give $f'(x) \geq 0$.
- Some sources define a "critical point" as a point x where $f'(x) = 0$, $f'(x)$ doesn't exist, or x is an endpoint of the domain. The point of this definition is that these critical points are all the points that could have local extrema.
- Rolle: if f is continuous on $[a, b]$ and differentiable on (a, b), and $f(a) = f(b)$, then there is a point $x \in (a, b)$ so that $f'(x) = 0$.

  Proof: if f is constant, we're done. Otherwise, suppose $f(t) > f(a)$ for some $t \in (a, b)$. Then by the EVT, there is an $x \in (a, b)$ that achieves the maximum, which means $f'(x) = 0$. If $f(a)$ is the maximum, we do the same reasoning with the minimum.
- Generalized MVT: if f and g are continuous real functions on $[a, b]$ which are differentiable in (a, b), then there is a point $x \in (a, b)$ such that

$$[f(b) - f(a)]\, g'(x) = [g(b) - g(a)]\, f'(x)$$

  Proof: apply Rolle's theorem to

$$h(t) = [f(b) - f(a)]\, g'(t) - [g(b) - g(a)]\, f'(t).$$

- Intuitively, if we consider the curve parametrized by $(f(t), g(t))$, the generalized MVT states that some tangent line to the curve is parallel to the line connecting the endpoints.
- MVT: setting $g(x) = x$ in the generalized MVT, there is a point $x \in (a, b)$ so that

$$f(b) - f(a) = (b - a)f'(x).$$

- One use of the MVT is that it allows us to connect the derivative at a point, which is local, with function values on a finite interval. For example, we can use it to show that if $f'(x) \geq 0$, then f is monotonically increasing.
- The MVT doesn't apply for vector valued functions, as there's too much 'freedom in direction'. The closest thing we have is the bound

$$|\mathbf{f}(b) - \mathbf{f}(a)| \leq (b - a)|\mathbf{f}'(x)|$$

  for all $x \in (a, b)$.

Theorem (L'Hospital)**.** Let f and g be real and differentiable in (a, b) with $g'(x) \neq 0$ for all $x \in (a, b)$. Suppose $f'(x)/g'(x) \to A$ as $x \to a$. Then if $f(x) \to 0$ and $g(x) \to 0$ as $x \to a$, or $g(x) \to \infty$ as $x \to a$, then $f(x)/g(x) \to A$ as $x \to a$.

Theorem (Taylor)**.** Suppose f is a real function on $[a, b]$, $f^{(n-1)}$ is continuous on $[a, b]$, and $f^{(n)}(t)$ exists for all $t \in (a, b)$. Let α and β be distinct points in $[a, b]$, and let

$$P(t) = \sum_{k=0}^{n-1} \frac{f^{(k)}(\alpha)}{k!}(t - \alpha)^k$$

Then there exists a point $x \in (\alpha, \beta)$ so that

$$f(\beta) = P(\beta) + \frac{f^{(n)}(x)}{n!}(\beta - \alpha)^n$$

This bounds the error of a polynomial approximation in terms of the maximum value of $f^{(n)}(x)$.

Proof. Applying the MVT, let M be the number such that

$$f(\beta) = P(\beta) + M(\beta - \alpha)^n$$

and define the function

$$g(t) = f(t) - P(t) - M(t - \alpha)^n.$$

By construction, g satisfies the properties

$$g(\alpha) = g'(\alpha) = \ldots = g^{(n-1)}(\alpha) = 0, \quad g(\beta) = 0, \quad g^{(n)}(t) = f^{(n)}(t) - n!M.$$

Then we wish to show that $g^{(n)}(t) = 0$ for some $t \in (\alpha, \beta)$. Applying Rolle's theorem gives a point $x_1 \in (\alpha, \beta)$ where $g'(x_1) = 0$. Repeating this for g' on the interval (x_1, β) gives a point x_2 where $g''(x_2) = 0$, and so on, giving the result.

Corollary. Under the same conditions as above, we have

$$f(x) = P(x) + \epsilon(x)(x - \alpha)^{n-1}$$

where $\epsilon(x) \to 0$ as $x \to \alpha$.

2.3 Integration

In this section, we define integration over intervals on the real line.

- A partition P of the interval $[a, b]$ is a finite set of points $x_0, \ldots, x_n$ with

  $$a = x_0 \leq x_1 \leq \ldots \leq x_{n-1} \leq x_n = b.$$

  We write $\Delta x_i = x_i - x_{i-1}$.

- Let f be a bounded real function defined on $[a, b]$. Then for a partition P, define

  $$M_i = \sup_{[x_{i-1}, x_i]} f(x), \quad m_i = \inf_{[x_{i-1}, x_i]} f(x)$$

  and

  $$U(P, f) = \sum M_i \Delta x_i, \quad L(P, f) = \sum m_i \Delta x_i.$$

- Define the upper and lower Riemann integrals as

$$\overline{\int_a^b} f\,dx = \inf U(P,f), \quad \underline{\int_a^b} f\,dx = \sup L(P,f)$$

  where the inf and sup are taken over all partitions P. These quantities are always defined if f is bounded, because this implies that M_i and m_i are bounded, which implies the upper and lower integrals are. Conversely, our notion of integration doesn't make sense if f isn't bounded, though we'll find a way to accommodate this later.

- If the upper and lower integrals are equal, we say f is Riemann-integrable on $[a,b]$, write $f \in \mathcal{R}$, and denote their common value as $\int_a^b f\,dx$.

- Given a monotonically increasing function α on $[a,b]$, define

$$\Delta\alpha_i = \alpha(x_i) - \alpha(x_{i-1}), \quad U(P,f,\alpha) = \sum_i M_i \Delta\alpha_i, \quad L(P,f,\alpha) = \sum_i m_i \Delta\alpha_i$$

  and the upper and lower integrals analogously. If they are the same, we say f is integrable with respect to α, write $f \in \mathcal{R}(\alpha)$, and denote their common value as $\int_a^b f d\alpha$. This is the Riemann-Stieltjes integral, with the Riemann integral as the special case $\alpha(x) = x$.

Next, we find the conditions for integrability. Below, we always let f be real and bounded, and α be monotonically increasing, on the interval $[a,b]$.

- P^* is a refinement of P if $P^* \supset P$ (i.e. we only split existing intervals into smaller ones). Given two partitions P_1 and P_2, their common refinement is $P^* = P_1 \cup P_2$.

- Refining a partition increases L and decreases U. This is clear by considering a refinement that adds exactly one extra interval.

- The lower integral is not greater than the upper integral. For any two partitions P_1 and P_2,

$$L(P_1,f,\alpha) \le L(P^*,f,\alpha) \le U(P^*,f,\alpha) \le U(P_2,f,\alpha).$$

  Taking sup over P_1 and inf over P_2 on both sides of this inequality gives the result.

- Therefore, $f \in \mathcal{R}(\alpha)$ on $[a,b]$ iff, for every $\epsilon > 0$, there exists a partition so that

$$U(P,f,\alpha) - L(P,f,\alpha) < \epsilon.$$

  This follows immediately from the previous point, and will serve as a useful criterion for integrability: we seek to construct partitions that give us an arbitrarily small 'error' ϵ.

- If $U(P,f,\alpha) - L(P,f,\alpha) < \epsilon$, then we have

$$\sum_i |f(s_i) - f(t_i)| \Delta\alpha_i < \epsilon$$

  where s_i and t_i are arbitrary points in $[x_{i-1}, x_i]$. Moreover, if the integral exists,

$$\left| \sum_i f(t_i)\Delta\alpha_i - \int_a^b f\,d\alpha \right| < \epsilon.$$

We can use these basic results to prove integrability theorems. We write $\Delta\alpha = \alpha(b) - \alpha(a)$.

- If f is continuous on $[a, b]$, then $f \in \mathcal{R}(\alpha)$ on $[a, b]$.

  Proof: since $[a, b]$ is compact, f is uniformly continuous. Then for any $\epsilon > 0$, there is a $\delta > 0$ so that $|x - t| < \delta$ implies $|f(x) - f(t)| < \epsilon$. Choosing a partition with $\Delta x_i < \delta$, the difference between the upper and lower integrals is at most $\epsilon\Delta\alpha$, and taking ϵ to zero gives the result.

- If f is monotonic on $[a, b]$ and α is continuous on $[a, b]$, then $f \in \mathcal{R}(\alpha)$.

  Proof: by the IVT, we can choose a partition so that $\Delta\alpha_i = \Delta\alpha/n$. By telescoping the sum, the error is bounded by $(\Delta\alpha/n)(f(b) - f(a))$. Taking n to infinity gives the result.

- If f is bounded on $[a, b]$ and has only finitely many points of discontinuity, none of which are also points of discontinuity of α, then $f \in \mathcal{R}(\alpha)$.

  Proof: choose a partition so that each point of discontinuity is in the interior of a segment $[u_i, v_i]$, where these segments' $\Delta\alpha_i$ values add up to ϵ. Then f is continuous on the compact set $[a, b] \setminus [u_i, v_i]$, so applying the previous theorem gives an $O(\epsilon)$ error.

  The segments with discontinuities contribute at most $2M\epsilon$, where $M = \sup |f(x)|$ is finite. Then the overall error is $O(\epsilon)$ as desired.

- Suppose $f \in \mathcal{R}(\alpha)$ on $[a, b]$, $m \le f(x) \le M$ on $[a, b]$, ϕ is continuous on $[m, M]$, and $h(x) = \phi(f(x))$ on $[a, b]$. Then $h \in \mathcal{R}(\alpha)$ on $[a, b]$. That is, continuous functions preserve integrability.

Example. The function

$$f(x) = \begin{cases} 0 & \text{if } x \in \mathbb{Q} \\ 1 & \text{otherwise} \end{cases}$$

is not Riemann integrable, because the upper integral is $(b - a)$ and the lower integral is zero.

2.4 Properties of the Integral

Below, we assume that all functions are integrable whenever applicable.

- Integration is linear,

$$\int_a^b (f_1 + f_2)\, d\alpha = \int_a^b f_1\, d\alpha + \int_a^b f_2\, d\alpha \quad \int_a^b cf\, d\alpha = c\int_a^b f\, d\alpha.$$

  Proof: first, we prove that $f_1 + f_2$ is integrable. For any partition, we have

$$L(P, f_1, \alpha) + L(P, f_2, \alpha) \le L(P, f, \alpha) \le U(P, f, \alpha) \le U(P, f_1, \alpha) + U(P, f_2, \alpha)$$

  Pick partitions for f_1 and f_2 with error $\epsilon/2$. Then by the inequality above, their common refinement P has error at most ϵ for $f_1 + f_2$, as desired. Moreover, using the inequality again,

$$\int f\, d\alpha \le U(P, f, \alpha) < \int f_1\, d\alpha + \int f_2\, d\alpha + \epsilon.$$

  Repeating this argument with f_i replaced with $-f_i$ gives the desired result.

- If $f_1(x) \le f_2(x)$ on $[a, b]$, then
$$\int_a^b f_1 \, d\alpha \le \int_a^b f_2 \, d\alpha.$$
- Integration ranges add
$$\int_a^c f \, d\alpha + \int_c^b d\alpha = \int_a^b d\alpha.$$
- ML inequality: if $|f(x)| \le M$ on $[a, b]$, then
$$\left| \int_a^b f \, d\alpha \right| \le M(\alpha(b) - \alpha(a)).$$
- Integration is also linear in α,
$$\int_a^b f \, d(\alpha_1 + \alpha_2) = \int_a^b f \, d\alpha_1 + \int_a^b f \, d\alpha_2, \quad \int_a^b f \, d(c\alpha) = c \int_a^b f \, d\alpha.$$
As before, the integrals on the left exist if the ones on the right do.
- Products of integrable functions are integrable.
Proof: we use an algebraic trick. Let these functions be f and g. Since $\phi(t) = t^2$ is continuous, f^2 and g^2 are integrable, but then so is
$$4fg = (f + g)^2 - (f - g)^2$$
A similar trick works with maximum and minima, as $\max(f, g) = (f + g)/2 + |f - g|/2$.
- If f is integrable, then so is $|f|$, and
$$\left| \int_a^b f \, d\alpha \right| \le \int_a^b |f| \, d\alpha$$
Proof: for integrability, compose with $\phi(t) = |t|$. The inequality follows from $f \le |f|$.

The reason we used the Riemann-Stieltjes integral is because the choice of α gives us more flexibility. In particular, the Riemann-Stieltjes integral contains infinite series as a special case.

- Define the unit step function I as
$$I(x) = \begin{cases} 0 & \text{if } x \le 0 \\ 1 & \text{if } x > 0. \end{cases}$$
- If $a < s < b$, and f is bounded on $[a, b]$ and continuous at s, and $\alpha(x) = I(x - s)$, then
$$\int_a^b f \, d\alpha = f(s).$$

- If $c_n \geq 0$ and $\sum_n c_n$ converges, and (s_n) is a sequence of distinct points in (a, b), and f is continuous on $[a, b]$, then

$$\alpha(x) = \sum_n c_n I(x - s_n) \quad \to \quad \int_a^b f \, d\alpha = \sum_n c_n f(s_n).$$

  Proof: the series on the right-hand side converges by comparison to $\sum_n M c_n$ where $M = \sup |f(x)|$. We need to show that it converges to the desired integral; to do this, consider truncating the series after N terms so that the rest of the terms add up to ϵ, and let

$$\alpha_N(x) = \sum_{n=0}^{N} c_n I(x - s_n).$$

  Then $\int f \, d\alpha_N$ is at most $M\epsilon$ away from $\int f \, d\alpha$ by the ML inequality, while the truncated series is at most ϵ away from the full series. Taking ϵ to zero gives the result.

Note. These results show that equations from physics like

$$I = \int x^2 \, dm$$

make sense; with the Riemann-Stieltjes integral, this equation holds whether the masses are continuous or discrete, or both; the function $m(x)$ is the amount of mass to the left of x.

- Let α increase monotonically and let α be differentiable with $\alpha' \in \mathcal{R}$ on $[a, b]$. Let f be bounded on $[a, b]$. Then

$$\int_a^b f d\alpha = \int_a^b f(x)\alpha'(x) \, dx$$

  where one integral exists if and only if the other does.

  Proof: we relate the integrals using the MVT. For all partitions P, we can use the MVT to choose t_i in each interval so that

$$\Delta\alpha_i = \alpha'(t_i)\Delta x_i.$$

  Now consider taking s_i in each interval to yield the upper sum

$$U(P, f, \alpha) = \sum_i f(s_i)\Delta\alpha_i = \sum_i f(s_i)\alpha'(t_i)\Delta x_i.$$

  Now, since α' is integrable, we can choose P so that $U(P, \alpha') - L(P, \alpha') < \epsilon$. Then we have

$$\sum_i |\alpha'(s_i) - \alpha'(t_i)|\Delta x_i < \epsilon$$

  which implies that

$$|U(P, f, \alpha) - U(P, f\alpha')| \leq M\epsilon$$

  where $M = \sup |f(x)|$. Therefore the upper integrals (if they exist) must coincide; similarly the lower integrals must, giving the result.

- Let φ be a strictly increasing continuous function that maps $[A, B]$ onto $[a, b]$. Let α be monotonically increasing on $[a, b]$ and $f \in \mathcal{R}(\alpha)$ on $[a, b]$. Let

$$\beta(y) = \alpha(\varphi(y)), \quad g(y) = f(\varphi(y)).$$

  Then $g \in \mathcal{R}(\beta)$ on $[A, B]$ with

$$\int_A^B g\, d\beta = \int_a^b f\, d\alpha$$

  Proof: φ gives a one-to-one correspondence between partitions of $[a, b]$ and $[A, B]$. Corresponding partitions have identical upper and lower sums, so the integrals must be equal.

Note. The first proof above shows another common use of the MVT: pinning down specific points where an 'on average' statement is true. Having these points in hand makes the rest of the proof more straightforward.

Note. A set $A \subset \mathbb{R}$ has measure zero if, for every $\epsilon > 0$ there exists a countable collection of open intervals $\{(a_i, b_i)\}$ such that

$$A \subset \bigcup_i (a_i, b_i), \quad \sum_i (b_i - a_i) < \epsilon$$

That is, the "length" of the set is arbitrarily small. Lebesgue's theorem states that a bounded real function is Riemann integrable if and only if its set of discontinuities has measure zero.

Next, we relate integration and differentiation.

- Let $f \in \mathcal{R}$ on $[a, b]$. For $x \in [a, b]$, let

$$F(x) = \int_a^x f(t)\, dt.$$

  Then F is continuous on $[a, b]$, and if f is continuous at x_0, then F is differentiable at x_0 with $F'(x_0) = f(x_0)$.

  Proof: F is continuous by the ML inequality, and the fact that f is bounded. The second part also follows by the ML inequality: by continuity, we can bound $f(u) - f(x_0)$ when u is close to x_0. Then the quantity $F'(x_0) - f(x_0)$ can be bounded by the ML inequality to zero.

- FTC: if $f \in \mathcal{R}$ on $[a, b]$ and F is differentiable on $[a, b]$ with $F' = f$, then

$$\int_a^b f(x)\, dx = F(b) - F(a).$$

  Proof: choose a partition P so that $U(P, f) - L(P, f) < \epsilon$. By the MVT, we can choose points in each interval so that

$$F(x_i) - F(x_{i-1}) = f(t_i)\Delta x_i \quad \to \quad \sum_i f(t_i)\Delta x_i = F(b) - F(a)$$

  Then both the upper and lower integrals are within ϵ of $F(b) - F(a)$, and taking ϵ to zero gives the result.

- Vector ML inequality: for $\mathbf{f} : [a, b] \to \mathbb{R}^k$ and $\mathbf{f} \in \mathcal{R}(\alpha)$, then

$$\left| \int_a^b \mathbf{f}\, d\alpha \right| \leq \int_a^b |\mathbf{f}|\, d\alpha.$$

  Proof: first, we must show that $|\mathbf{f}|$ is integrable; this follows because it can be built from squaring, addition, square root, and norm, all of which are continuous. (The square root function is continuous because it is the inverse of the square on the compact interval $[0, M]$ for any M.) To show the bound, let $\mathbf{y} = \int \mathbf{f}\, d\alpha$. Then

$$|\mathbf{y}|^2 = \int \sum_i y_i f_i \, d\alpha \leq \int |\mathbf{y}||\mathbf{f}|\, d\alpha$$

  by Cauchy-Schwartz. Canceling $|\mathbf{y}|$ from both sides gives the result.

Note. The proofs above show some common techniques: using the ML inequality to bound an error to zero, and using the MVT to get concrete points to work with.

2.5 Uniform Convergence

Next, we establish some useful technical results using uniform convergence.

- A sequence of functions $f_n : X \to Y$ converges pointwise to $f : X \to Y$ if, for every $x \in X$,

$$\lim_{n\to\infty} f_n(x) = f(x).$$

  One must treat pointwise convergence with caution; the problems boil down to the fact that two limits may not commute. For instance, the pointwise limit of continuous functions may not be continuous.

- Integration and pointwise convergence don't commute. Let $f_n : [0, 1] \to \mathbb{R}$ where $f_n(x) = n^2$ on $(0, 1/n)$ and 0 otherwise. Then

$$\lim_{n\to\infty} \int_0^1 f_n(x)\, dx = \lim_{n\to\infty} n = \infty, \quad \int_0^1 \lim_{n\to\infty} f_n(x)\, dx = 0.$$

  An analogous statement holds for integration and series summation.

- Differentiation and pointwise convergence don't commute. Let $f_n(x) = \sin(n^2 x)/n$, so $f_n \to 0$ pointwise. But $f_n'(x) = -n\cos(n^2 x)$, so $f_n'(\pi/4) \to \infty$.

  An analogous statement holds for differentiation and series summation.

- A sequence of functions $f_n : X \to Y$ converges uniformly on X to $f : X \to Y$ if, for all $\epsilon > 0$, there exists an N so that for $n > N$, we have

$$d_Y(f_n(x), f(x)) < \epsilon$$

  for all $x \in X$. That is, just as in the definition of uniform continuity, we may use the same N for all points. For concreteness, we specialize to $X \subset \mathbb{R}$ and $Y = \mathbb{R}$ with the standard metric.

- An alternative way to write the criterion for uniform convergence is that

$$\alpha_n = \sup_{x \in X} |f_n(x) - f(x)| \to 0$$

as $n \to \infty$. It is clear that uniform convergence implies pointwise convergence.

We now establish properties of uniform convergence. All of our functions below map $X \to \mathbb{R}$.

- If (f_n) converges uniformly on X to f and the f_n are continuous, f is.

Proof: we will show f is continuous at $p \in X$. Fix $\epsilon > 0$. For x near p so that $|x - p| < \delta$,

$$|f(x) - f(p)| \leq |f(x) - f_N(x)| + |f_N(x) - f_N(p)| + |f_N(p) - f(p)|$$

and we are done if we can show the right-hand side is bounded by ϵ. We may first choose N so the first and third terms are bounded by $\epsilon/3$, by the definition of uniform continuity. Next, we choose δ so the second term is bounded by $\epsilon/3$, since f_N is continuous, giving the result.

- Uniform convergence also comes in a "Cauchy" variant: (f_n) converges uniformly on X if and only if, for all $\epsilon > 0$, there exists an N so that for $n, m > N$,

$$|f_n(x) - f_m(x)| < \epsilon$$

for all $x \in X$. This follows from the completeness of $\mathbb{R}$.

- If (f_n) converges uniformly to X on f and the f_n are integrable, f is.

Proof: for any $\epsilon > 0$, f is within ϵ of f_n for sufficiently large n. Then the upper and lower integrals of f are within $(b - a)\epsilon$ of each other, giving the result. In particular, the integral of f must be the limit of the sequence $(\int f_n)$.

- If (f_n) converges pointwise to $[a, b]$ on f, the f_n are differentiable on (a, b), and the f_n' are continuous and converge uniformly to a bounded function g on (a, b), then f is differentiable and $f' = g$.

Proof: the simplest proof uses integration. Taking the result

$$\int_a^x f_n'(t)\, dt = f_n(x) - f_n(a)$$

in the limit $n \to \infty$, and using the previous result, we have

$$\int_a^x g(t)\, dt = f(x) - f(a).$$

On the other hand, since g is continuous, the left-hand side is a differentiable function $F(x)$ with $F' = g$. Hence by differentiating both sides, $g = f'$ as desired.

We now apply these results to power series.

- Similarly, uniform convergence can be defined for series. For a set of real-valued functions u_k, the series $\sum u_k$ converges pointwise/uniformly on X if and only if (f_n) does, where

$$f_n = u_1 + \ldots + u_n.$$

By the above, if $\sum u_k$ converges uniformly and the u_k are continuous, then $\sum u_k$ is continuous. The same holds with "integrable" in place of "continuous", as well as "differentiable" if the u_k' are continuous and $\sum u_k'$ converges uniformly. In these cases, differentiation and integration can be performed term by term.

- Uniform convergence for series also comes in a Cauchy variant. The series $\sum u_k$ converges uniformly on X if and only if, for all $\epsilon > 0$, there exists an N so that for $n > m > N$,

$$|u_{m+1}(x) + \ldots + u_n(x)| < \epsilon$$

  for all $x \in X$.

- Weierstrass M-test: the series $\sum u_k$ converges uniformly on X if there exist real constants M_k so that for all k and $x \in X$,

$$|u_k(x)| \leq M_k, \quad \sum M_k \text{ converges.}$$

  This follows directly from the previous point, because $\sum M_k$ is a Cauchy series. This condition is stronger than necessary, because each M_k depends on the largest value $u_k(x)$ takes anywhere, but in practice is quite useful.

- As we saw earlier, a power series $\sum_k c_k x^k$ has a radius of convergence R so that it converges absolutely for $|x| < R$, and diverges for $|x| > R$.

- For any δ with $0 < \delta < R$, $\sum_k c_k x^k$ converges uniformly on $[-R+\delta, R-\delta]$. This simply follows from the Weierstrass M-test, using $M_k = |c_k(R-\delta)^k|$, where $\sum M_k$ converges by the root test.

  Note that the power series does *not* necessarily converge uniformly on $(-R, R)$. One simple example is $\sum x^k$, which has $R = 1$. However, the "up to δ" result here will be good enough because we can take δ arbitrarily small.

- As a result, the power series $\sum_k c_k x^k$ defines a continuous function f on $(-R, R)$. In particular, this establishes the continuity of various functions such as $\exp(x)$, $\sin(x)$, and $\cos(x)$. The reason that the "up to δ" issue above doesn't matter is that continuity is a local condition, which holds at individual points, while uniform convergence is global. Another way of saying this is that a function is continuous on an arbitrary union of domains where it is continuous, but this doesn't hold for uniform convergence.

- Similarly, the power series $\sum_k c_k x^k$ defines a differentiable function f on $(-R, R)$ which can be differentiated term by term. This takes some technical work, as we must show $\sum_k k c_k x^{k-1}$ converges uniformly. Repeating this argument, f is infinitely differentiable on $(-R, R)$.

- Weierstrass's polynomial approximation theorem states that for any continuous $f : [a, b] \to \mathbb{R}$, there exists a sequence (P_n) of real polynomials which converges uniformly to f.

3 Complex Analysis

3.1 Analytic Functions

Let $f(z) = u(x, y) + iv(x, y)$ where $z = x + iy$ and u and v are real.

- The derivative of a complex function $f(z)$ is defined by the usual limit definition. We say a complex function is analytic/holomorphic at a point z_0 if it is differentiable in a neighborhood of z_0.
- Approaching along the x and y-directions respectively, we have

  $$f'(z) = u_x + iv_x, \quad f'(z) = -iu_y + v_y.$$

  Thus, for the derivative to be defined we must have

  $$u_x = v_y, \quad v_x = -u_y$$

  which are the Cauchy-Riemann equations. These are also a sufficient condition, as any other directional derivative can be computed by a linear combination of these two.
- Assuming that f is twice differentiable, both u and v are harmonic functions, $u_{xx} + u_{yy} = v_{xx} + v_{yy} = 0$, by the equality of mixed partials.
- The level curves of u and v are orthogonal, because

  $$\nabla u \cdot \nabla v = u_x v_x + u_y v_y = -u_x u_y + u_x u_y = 0.$$

  In particular, this means that u solves Laplace's equation when conductor surfaces are given by level curves of v.
- Changing coordinates to polar gives an alternate form of the Cauchy-Riemann equations,

  $$u_r = \frac{1}{r} v_\theta, \quad v_r = -\frac{1}{r} u_\theta$$

  where the derivative is

  $$f'(z) = e^{-i\theta}(u_r + iv_r).$$

- Locally, a complex function differentiable at z_0 satisfies $\Delta f = f'(z_0)\Delta z$. Thus the function looks like a local 'scale and twist' of the complex plane, which provides some intuition. For example, $\overline{z}$ is not differentiable because it behaves like a 'flip' and twist.
- The mapping $z \mapsto f(z)$ takes harmonic functions to harmonic functions as long as f is differentiable with $f'(z) \neq 0$. This is because the harmonic property ('function value equal to average of neighbors') is invariant under rotation and scaling.
- Conformal transformations are maps of the plane which preserve angles; all holomorphic functions with nonzero derivative produce such a transformation.
- A domain is an open, simply connected region in the complex plane. We say a complex function is analytic in a region if it is analytic in a domain containing that region. If a function is analytic everywhere, it is called entire.

- Using the formal coordinate transformation from (x, y) to $(z, \overline{z})$ yields the Wirtinger derivatives,

$$\partial_z = \frac{1}{2}(\partial_x - i\partial_y), \quad \partial_{\overline{z}} = \frac{1}{2}(\partial_x + i\partial_y).$$

  The Cauchy-Riemann equations are equivalent to

$$\partial_{\overline{z}} f = 0.$$

  Similarly, we say f is antiholomorphic if $\partial_z f = 0$. The Wirtinger derivatives satisfy a number of intuitive properties, such as $\partial_z(zz^*) = z^*$.

As an example, we consider ideal fluid flow.

- The flow of a fluid is described by a velocity field $\mathbf{v} = (v_1, v_2)$. Ideal fluid flow is steady, nonviscous, incompressible, and irrotational. The latter two conditions translate to $\nabla \cdot v = \nabla \times v = 0$, which in terms of components are

$$\partial_x v_1 + \partial_y v_2 = \partial_x v_2 - \partial_y v_1 = 0.$$

  We are switching our derivative notation to avoid confusion with the subscripts.

- The zero curl condition can be satisfied automatically by using a velocity potential, $\mathbf{v} = \nabla\phi$. It is also useful to define a stream function ψ, so that

$$v_1 = \partial_x \phi = \partial_y \psi, \quad v_2 = \partial_y \phi = -\partial_x \psi$$

  in which case incompressibility is also automatic.

- Since ϕ and ψ satisfy the Cauchy-Riemann equations, they can be combined into an analytic complex velocity potential

$$\Omega(z) = \phi(x, y) + i\psi(x, y).$$

- Since the level sets of ψ are orthogonal to those of ϕ, level sets of the stream function ψ are streamlines. The derivative of Ω is the complex velocity,

$$\Omega'(z) = \partial_x \phi + i\partial_x \psi = \partial_x \phi - i\partial_y \phi = v_1 - iv_2.$$

  The boundary conditions are typically of the form 'constant velocity at infinity' (which requires ϕ to approach a linear function) and 'zero velocity normal to an obstacle' (which requires ψ to be constant on its surface).

Example. The uniform flow $\Omega(z) = v_0 e^{-i\theta_0} z$. The real part is

$$\phi(x, y) = v_0(\cos\theta_0 x + \sin\theta_0 y)$$

giving a velocity of $\mathbf{v} = v_0(\cos\theta_0, \sin\theta_0)$.

Example. Flow past a cylinder. Consider the velocity potential

$$\Omega(z) = v_0(z + a^2/z), \quad \phi = v_0(r + a^2/r)\cos\theta, \quad \psi = v_0(r - a^2/r)\sin\theta.$$

At infinity, the flow has uniform velocity v_0 to the right. Since $\psi = 0$ on $r = a$, this potential describes flow past a cylindrical obstacle. To get intuition for this result, note that ϕ also serves as an electric potential in the case of a cylindrical conductor at $r = a$, in a uniform background field. We know that the cylinder is polarized, producing a dipole moment, and corresponding dipole potential $\cos\theta/r^2 = x/r^3$. For the fluid flow there is one less power of r since the situation is two-dimensional.

Example. Using a conformal transformation. The complex potential $\Omega(z) = z^2$ has stream function $2xy$, and hence $xy = 0$ is a streamline; hence this potential describes flow at a rectangular corner. An alternate solution is to apply conformal transformation to the boundary condition. If we define $z = w^{1/2}$, with $z = x + iy$ and $w = u + iv$, then the boundary $x = 0$, $y = 0$ is mapped to $v = 0$. This problem is solved by the uniform flow $\Omega(w) = w$, and transforming back gives the result.

3.2 Multivalued Functions

Multivalued functions arise in complex analysis as the inverses of single-valued functions.

- Consider $w = z^{1/2}$, defined to be the 'inverse' of $z = w^2$. For every z, there are two values of w, which are opposites of each other. In polar coordinates,

  $$w = r^{1/2} e^{i\theta_p/2} e^{n\pi i}$$

  where θ_p is restricted to lie in $[0, 2\pi)$ and $n = 0, 1$ indexes the two possible values. The surprise is that if we go in a loop around the origin, we can move from $n = 0$ to $n = 1$, and vice versa!

- We say $z = 0$ is a branch point; a loop traversed around a branch point can change the value of a multivalued function. Similarly, the point $z = \infty$ is a branch point, as can be seen by taking $z = 1/t$ and going around the point $t = 0$.

- A multivalued function can be rendered single-valued and continuous in a subset of the plane by choosing a branch. Often this is done by removing a curve, called a 'branch cut', from the plane. In the case above, the branch cut is arbitrary, but must join the branch points $z = 0$ and $z = \infty$. This prevents curves from going around either of the branch points. (Generally, but not always, branch cuts connect pairs of branch points.)

- Using stereographic projection, the branch points for $w = z^{1/2}$ are the North and South poles, and the branch cut connects them.

- A second example is the logarithm function,

  $$\log z = \log |z| + i\theta_p + 2n\pi i$$

  where $n \in \mathbb{Z}$, and we take the logarithm of a real number to be single-valued. This function has infinitely many branches, with a branch point at $z = 0$. It also has a branch point at $z = \infty$, by considering $\log 1/z = -\log z$.

- For a rational power $z^{m/l}$ with m and l relatively prime, we have

  $$z^{m/l} = e^{(m/l)\log z} = \exp\left[\frac{m}{l}(\log r + i\theta_p)\right] \exp\left[2\pi i(mn/l)\right]$$

  so that there are l distinct branches. For an irrational power, there are infinitely many branches.

Example. An explicit branch of the logarithm. Defining

$$w = \log z, \quad z = x + iy, \quad w = u + iv$$

we have

$$e^{2u} = x^2 + y^2, \quad \tan v = \frac{y}{x}.$$

The first can be easily inverted to yield $u = \log(x^2 + y^2)/2$, which is single-valued because the real log is, but the second is more subtle. For the inverse tangent of a real number, we customarily take the branch so that the range is $(-\pi/2, \pi/2)$. Then to maintain continuity of v, we set

$$v = \tan^{-1}(y/x) + C_i, \quad C_1 = 0, \quad C_2 = C_3 = \pi, \quad C_4 = 2\pi$$

in the i^{th} quadrant. Then the branch cut is along the positive x axis. Finally, we differentiate, for

$$\frac{d}{dz}\log z = u_x + iv_x = \frac{x - iy}{x^2 + y^2} = \frac{1}{z}$$

as expected.

Example. A more complicated multivalued function. Let $w = \cos^{-1} z$. We have

$$\cos w = z = \frac{e^{iw} + e^{-iw}}{2}$$

and solving this as a quadratic in e^{iw} yields

$$e^{iw} = z + i(1 - z^2)^{1/2} \quad \rightarrow \quad w(z) = -i\log(z + i(1 - z^2)^{1/2}).$$

The function thus has two sources of multivaluedness. We have branch points at $z = \pm 1$ due to the square root. There are no branch points due to the logarithm at finite z, because its argument is never zero, but there is a branch point at infinity (as can be seen by substituting $t = 1/z$). Intuitively, these branch points come from the fact that the cosine of x is the same as the cosine of $2\pi - x$ (for the square root) and the cosine of $x + 2\pi$ (for the logarithm).

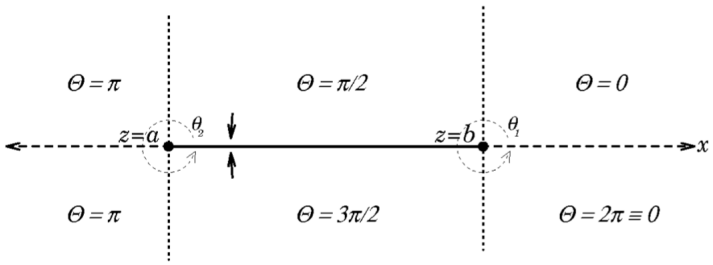
Example. Explicit construction of a more complicated branch. Consider

$$w = [(z - a)(z - b)]^{1/2}.$$

There are branch cuts at $z = a$ and $z = b$, though one can check by setting $t = 1/z$ that there is no branch cut at infinity. (Intuitively, going around the 'point at infinity' is the same as going around *both* finite branch points, each of which contribute a phase of π.) To explicitly set a branch, define

$$z - b = r_1 e^{i\theta_1}, \quad z - a = r_2 e^{i\theta_2}$$

so that $w \propto e^{i(\theta_1 + \theta_2)/2}$. A branch is thus specified by a choice of θ. For example, we may choose to restrict $0 \le \theta_i < 2\pi$, which gives a branch cut between a and b, as shown below.



An alternative choice can send the branch cut through the point at infinity, which is more easily visualized using stereographic projection. Similar reasoning can be used to handle any function made of products of $(z - x_i)^k$.

Note. Branches can be visualized geometrically as sheets of Riemann surfaces, which are generated by gluing copies of the complex plane together along branch cuts. The logarithm has an infinite 'spiral staircase' of such sheets, with each winding about the origin bringing us to the next.

Example. More flows. The potential $\Omega(z) = k \log(z)$ with k real describes a source or sink at the origin. Its derivative $1/z$ describes a dipole, i.e. a source and sink immediately adjacent.

By contrast, the potential $\Omega(z) = ik \log(z)$ describes circulation about the origin. Here, the multivaluedness of $\log(z)$ is crucial, because if the velocity potential were single-valued, then it would be impossible to have net circulation along any path; instead going around the origin takes us to another branch of the logarithm. (In multivariable calculus, we say that zero curl does not imply that a function is a gradient, if the domain is not simply connected. Here, we can retain the gradient function at the cost of making it multivalued.)

3.3 Contour Integration

Next, we turn to defining integration.

- A contour C in the complex plane can be parametrized as $z(t)$. We will choose to work with piecewise smooth contours, i.e. those where $z'(t)$ is piecewise continuous.
- For convenience, we may sometimes require that C be simple, i.e. that it does not intersect itself. This ensures that C winds about every point at most once.
- The contour integral of f along C is defined as

$$\int_C f(z)\,dz = \int_a^b f(z(t))z'(t)\,dt.$$

  All the usual properties of integration apply; in particular the result is independent of parametrization. In the piecewise smooth case, we simply define the integral by splitting C into smooth pieces.

- The ML inequality states that the magnitude any contour integral is bounded by the product of the supremum of $|f(z)|$ and the length of the contour.
- Cauchy's theorem: if f is analytic in a simply connected domain D, and f' is continuous, then along a simple closed contour C in D,

$$\oint_C f(z)\,dz = 0.$$

  Proof: in components, we have

$$\int f(z)\,dz = \int_C (u\,dx - v\,dy) + i(v\,dx + u\,dy).$$

  We can then apply Green's theorem to the real and imaginary parts. Applying the Cauchy-Riemann equations, the 'curl' is zero, giving the result. We need the simply connected hypothesis to ensure that C does not contain points outside of D.

- Goursat's theorem: Cauchy's theorem holds without the assumption that f' is continuous.

  Proof sketch: we break the integral down into the sum of integrals over contours with arbitrarily small size. By Taylor's theorem, the function can be expanded as the sum of a constant, linear, and sublinear term within each small contour. The integrals of the first two vanish, while the contributions of the sublinear terms go to zero in the limit of small contours.

- As a result, every analytic f in a simply connected domain has a primitive, i.e. a function F with $F' = f$, with
$$\int_C f(z)\,dz = F(b) - F(a).$$
We can construct the function F by simply choosing any contour connecting a to b.

Example. We integrate $f(z) = 1/z$ over an arbitrary closed contour which winds around the origin once. (Equivalently, any simple closed contour containing the origin.) Since f is analytic everywhere besides the origin, we may freely deform the contour so that it becomes a small circle of radius r about the origin. Then
$$\int_C \frac{dz}{z} = \int \frac{ire^{i\theta}}{re^{i\theta}}\,d\theta = 2\pi i.$$
This result can be thought of as due to having a multivalued primitive $F(z) = \log z$, or due to the hole at the origin. The analogous calculation for $1/z^n$ gives zero for $n \neq 1$, as there are single-valued primitives $1/z^{n-1}$.

Example. Complex fluid flow again. The circulation along a curve and flow out of a curve are
$$\Gamma = \int_C v_x\,dx + v_y\,dy, \quad \mathcal{F} = \int_C v_x\,dy - v_y\,dx.$$
Combining these, we find
$$\Gamma + i\mathcal{F} = \int_C \Omega'(z)\,dz$$
where Ω is the complex velocity potential. This also provides some general intuition: multiplying i makes the circulation and flux switch places.

Example. Let $P(z)$ be a polynomial with degree n and n simple roots, and let C be a simple closed contour. We wish to evaluate
$$I = \frac{1}{2\pi i} \oint_C \frac{P'(z)}{P(z)}\,dz.$$
First note that if $P(z) = A\prod_i (z - a_i)$, then
$$\frac{P'(z)}{P(z)} = \sum_i \frac{1}{z - a_i}.$$
Every root is thus a simple pole, so the integral is simply the number of roots in C. One way to think of this is that the integrand is really $d(\log P)$, and here $\log P$ has logarithmic branch points at every root, each of which gives a change of $2\pi i$.

Example. Consider the integral

$$I = \int_0^\infty e^{ix^2}\,dx.$$

We consider a contour integral of e^{iz^2} with a line from the origin to R, an arc to $Re^{i\pi/4}$, and a line back to the origin. The arc is exponentially suppressed and does not contribute in the limit $R \to \infty$, while the total integral is zero since the integrand is analytic. Thus

$$I = \int_0^\infty e^{i\pi/4} e^{-r^2}\,dr = e^{i\pi/4}\sqrt{\pi}/2.$$

More generally, this shows that the standard Gaussian integral formula holds for any complex σ^2 as long as the integral converges.

Next, we introduce some more theoretical results.

- Cauchy's integral formula states that if $f(z)$ is analytic in and on a simple closed contour C,

  $$f(z) = \frac{1}{2\pi i} \oint_C \frac{f(\xi)}{\xi - z}\,d\xi.$$

  Then the value of an analytic function is determined by the values of points around it. The proof is to deform the contour to a small circle about $\xi = z$, where the pole gives $f(z)$. The error term goes to zero by continuity and the ML inequality.

- As a corollary, if $f(z)$ is analytic in and on C, then all of its derivatives exist, with

  $$f^{(k)}(z) = \frac{k!}{2\pi i} \oint_C \frac{f(\xi)}{(\xi - z)^{k+1}}\,d\xi.$$

  Proof: we consider $k = 1$ first. The difference quotient is

  $$\frac{f(z+h) - f(z)}{h} = \frac{1}{2\pi i}\frac{1}{h} \oint_C f(\xi) \left(\frac{1}{\xi - (z+h)} - \frac{1}{\xi - z} \right) d\xi.$$

  This gives the desired result, plus an error term

  $$R = \frac{h}{2\pi i} \oint_C \frac{f(\xi)\,d\xi}{(\xi - z)^2(\xi - z - h)}.$$

  For $|\xi - z| > \delta$ and $|h| < \delta/2$, the integral is bounded by ML. Since h goes to zero, R goes to zero as well. This also serves as a proof that $f'(z)$ exists. The cases $k > 1$ are handed inductively by similar reasoning.

- Intuitively, if we represent a complex function as a Taylor series, the general formulas above simply pluck out individual terms of this series by shifting them over to $1/z$.

- Applying the ML inequality above yields the bound

  $$|f^{(n)}(z)| \le \frac{n!M}{R^n}$$

  where M is the maximum of $|f(z)|$ on C.

- Liouville's theorem: a bounded entire function must be constant.

  Proof: suppose f is bounded and apply the bound above for $n = 1$. Then $|f'(z)| \leq M/R$, and taking R to infinity shows that $f'(z) = 0$, so f is constant.

- Morera: if $f(z)$ is continuous in a domain D, and all contour integrals of f are zero, then $f(z)$ is analytic in D.

  Proof: we may construct a primitive $F(z)$ by integration, with $F'(z) = f(z)$. Since F is automatically twice-differentiable, f is analytic.

- Fundamental theorem of algebra: every nonconstant polynomial $P(z)$ has a root in $\mathbb{C}$.

  Proof: assume P has no roots. Since $|P(z)| \to \infty$ for $|z| \to \infty$, the function $1/P(z)$ is bounded and entire, and hence constant by Liouville's theorem. Then $P(z)$ is constant.

- Mean value property: if $f(z)$ is analytic on the set $|z - z_0| \leq r$, then

$$f(z_0) = \frac{1}{2\pi} \int_0^{2\pi} f(z_0 + re^{i\theta})\, d\theta.$$

  Intuitively, this is because the components of f are harmonic functions. It also follows directly from Cauchy's integral formula; the contour integral along the boundary is

$$f(z_0) = \frac{1}{2\pi i} \int_C \frac{f(z)}{z - z_0}\, dz = \frac{1}{2\pi i} \int_0^{2\pi} \frac{f(z_0 + re^{i\theta})}{re^{i\theta}}\, ire^{i\theta}\, d\theta = \frac{1}{2\pi} \int_0^{2\pi} f(z_0 + re^{i\theta})\, d\theta.$$

  As a corollary, if $|f|$ has a relative maximum at some point, then f must be constant in a neighborhood of that point.

- Maximum modulus: suppose $f(z)$ is analytic in a bounded connected region A. If f is continuous on A and its boundary, then either f is constant or the maximum of $|f|$ occurs only on the boundary of A.

  Proof: the assumptions ensure $|f|$ has an absolute maximum on A and its boundary by the extreme value theorem. If the maximum is in the interior of A, then f is constant by the mean value property.

Example. We evaluate the integral

$$\int_C \frac{dz}{z^2(1 - z)}$$

around a small counterclockwise circle centered at $z = 0$. Naively, one might think the answer is zero since the root at $z = 0$ is a double root, but $1/(1 - z)$ expands to $1 + z + \ldots$. Then the piece with a simple root is z/z^2, giving $2\pi i$. Another approach is to use Cauchy's integral formula with $f(z) = 1/(1 - z)$, which gives

$$\frac{1}{2\pi i} \int_C \frac{f(z)\, dz}{z^2}\, dz = f'(0) = 1$$

as expected.

3.4 Laurent Series

We begin by reviewing Taylor series. For simplicity, we center all series about $z = 0$.

- Previously, we have shown that a power series

$$f(z) = \sum_{n=0}^{\infty} a_n z^n, \quad \alpha = \limsup_{n\to\infty} \sqrt[n]{|a_n|}$$

converges for $|z| < R = 1/\alpha$ and diverges for $|z| > R$. It is uniformly convergent for $|z| < R$, so we may perform term-by-term integration and differentiation. For example, the power series

$$\sum_{n=1}^{\infty} n a_n z^{n-1}$$

converges to $f'(z)$, and also has radius of convergence R.

- We would like to show that a function's Taylor series converges to the function itself. For an infinitely-differentiable real function, Taylor's theorem states that the error of omitting the n^{th} and higher terms is bounded as

$$\text{error at } x \leq \max_{x'\in[0,x]} \frac{|f^{(n)}(x')|}{n!} x^n.$$

One can show this error goes to zero as n goes to infinity for common real functions, such as the exponential.

- For a complex differentiable function f, the Taylor series of f automatically converges to f within its radius of convergence. This is a consequence of Cauchy's integral formula.

To see this, let the Taylor series of f centered at zero have radius of convergence R. We consider a circular contour of radius $r_2 < R$ and let $|z| < r_1 < r_2$. Then

$$f(z) = \frac{1}{2\pi i} \oint \frac{f(\xi)}{\xi - z}\, d\xi = \frac{1}{2\pi i} \oint \sum_{n=0}^{\infty} \frac{f(\xi)}{\xi^{n+1}} z^n\, d\xi$$

where the geometric series is convergent since $r_1 < r_2$. In particular, it is uniformly convergent, so we can exchange the sum and the integral, giving

$$f(z) = \sum_{n=0}^{\infty} \frac{1}{2\pi i} \oint \frac{f(\xi)}{\xi^{n+1}} z^n\, d\xi = \sum_{n=0}^{\infty} \frac{f^{(n)}(0)}{n!} z^n.$$

Taking r_1 arbitrarily close to R gives the result.

- Therefore, we say a function is analytic at a point if it admits a Taylor series about that point with positive radius of convergence, and is equal to its Taylor series in a neighborhood of that point. We have shown that a complex differentiable function is automatically analytic and thus use the terms interchangeably.

- A function is singular if it is not analytic at a point.
  - The function $\log z$ has a singularity at $z = 0$ since it diverges there.

  - More subtly, e^{-1/z^2} has a singularity at $z = 0$ since it is not equal to its Taylor series in any neighborhood of $z = 0$.

- A singularity of a function is isolated if there is a neighborhood of that point, excluding the singular point itself, where the function is analytic.
  - The function $1/\sin(\pi/z)$ has singularities at $z = 0$ and $z = 1/n$ for integer n, and hence the singularity at $z = 0$ is not isolated.
  - As a real function, the singularity at $x = 0$ of $\log x$ is not isolated since $\log x$ is not defined for $x < 0$. As a single-valued complex function, the same holds because $\log z$ requires a branch cut starting at $z = 0$.

- More generally, suppose that $f(z)$ is complex differentiable in a region R, $z_0 \in R$, and the disk of radius r about z_0 is contained in R. Then f converges to its Taylor series about z_0 inside this disk. The proof of this statement is the same as above, just for general z_0.

- The zeroes of an analytic function, real or complex, are isolated. We simply expand in a Taylor series about the zero at $z = z_0$ and pull out factors of $z - z_0$ until the series is nonzero at z_0. The remaining series is nonzero in a neighborhood of z_0 by continuity.

Next, we turn to Laurent series.

- Suppose $f(z)$ is analytic on the annulus $A = \{r_1 < |z| < r_2\}$. Then we claim $f(z)$ may be written as a Laurent series
$$f(z) = \sum_{n=1}^{\infty} \frac{b_n}{z^n} + \sum_{n=0}^{\infty} a_n z^n$$
where the two parts are called the singular/principal and analytic/regular parts, respectively, and converge to analytic functions for $|z| < r_2$ and $|z| > r_1$, respectively.

- The proof is similar to our earlier proof for Taylor series. Let $z \in A$ and consider the contour consisting of a counterclockwise circle C_1 of radius greater than $|z|$ and a clockwise circle C_2 of radius less than $|z|$, both lying within the annulus. By Cauchy's integral formula,
$$f(z) = \frac{1}{2\pi i} \oint_{C_1 - C_2} \frac{f(\xi)}{\xi - z}\, d\xi = \frac{1}{2\pi i} \oint_{C_1} \sum_{n=0}^{\infty} \frac{f(\xi)}{\xi^{n+1}} z^n\, d\xi - \frac{1}{2\pi i} \oint_{C_2} \sum_{n=0}^{\infty} \frac{f(\xi)}{z^{n+1}} \xi^n\, d\xi$$
where both geometric series are convergent. These give the analytic and singular pieces of the Laurent series, respectively.

- From this proof we also read off integral expressions for the coefficients,
$$a_n = \frac{1}{2\pi i} \oint \frac{f(\xi)}{\xi^{n+1}}\, d\xi, \quad b_n = \frac{1}{2\pi i} \oint f(\xi) \xi^{n-1}\, d\xi.$$
Unlike for Taylor series, none of these coefficients can be expressed in terms of derivatives of f.

- In practice, we use series expansions and algebraic manipulations to determine Laurent series, though we must use series that converge in the desired annulus.

- Suppose $f(z)$ has an isolated singularity at z_0, so it has a Laurent series expansion about z_0.

- If all of the b_n are zero, then z_0 is a removable singularity. We may define $f(z_0) = a_0$ to make f analytic at z_0. Note that this is guaranteed if f is bounded by the ML inequality.
- If a finite number of the b_n are nonzero, we say z_0 is a finite pole of $f(z)$. If b_k is the highest nonzero coefficient, the pole has order k. A finite pole with order 1 is a simple pole, or simply a pole. The residue $\text{Res}(f, z_0)$ of a finite pole is b_1.
- Finite poles are nice, because functions with only finite poles can be made analytic by multiplying them with polynomials.
- If an infinite number of the b_n are nonzero, z_0 is an essential singularity. For example, $z = 0$ for e^{-1/z^2} is an essential singularity. Essential singularities behave very badly; Picard's theorem states that they take on all possible values infinitely many times, with at most one exception, for any neighborhood of z_0.
- A function that is analytic on some region with the exception of a set of poles of finite order is called meromorphic.
- Note that all of these definitions apply only to isolated poles. For example, the logarithm has a branch cut starting at $z = 0$, so the order of this singularity is not defined.

Example. The function $f(z) = 1/(z(z-1))$ has poles at $z = 0$ and $z = 1$, and hence has a Laurent series about $z = 0$ for $0 < |z| < 1$ and $1 < |z| < \infty$. In the first case, the result can be found by geometric series,

$$f(z) = -\frac{1}{z}\frac{1}{1-z} = -\frac{1}{z}(1 + z + z^2 + \ldots).$$

We see that the residue of the pole at $z = 0$ is -1. In the second case, this series expansion does not converge; we instead expand in $1/z$ for the completely different series

$$f(z) = \frac{1}{z}\frac{1}{z(1-1/z)} = \frac{1}{z^2}\left(1 + \frac{1}{z} + \frac{1}{z^2} + \ldots\right).$$

In particular, note that there is no $1/z$ term because the residues of the two (simple) poles cancel out, as can be seen by partial fractions; we cannot use this Laurent series to compute the residue of the $z = 0$ pole.

Example. Going to the complex plane gives insight into why some real Taylor series fail. First, consider $f(x) = 1/(1 + x^2)$ about $x = 0$. This Taylor series breaks down for $|x| \geq 1$ even though the function itself is not singular at all. This is explained by the poles at $z = \pm i$ in the complex plane, which set the radius of convergence.

As another example, e^{-1/x^2} does not appear to be pathological on the real line at first glance. One can see that it is not analytic because its high-order derivatives blow up, but an easier way is to note that when approached along the imaginary axis, the function becomes e^{1/x^2}, which diverges very severely at $x = 0$.

Next, we give some methods for computing residues, all proven with Laurent series.

- If f has a finite pole at z_0, then

$$\text{Res}(f, z_0) = \lim_{z \to z_0} (z - z_0) f(z) = \lim_{z \to z_0} \frac{1}{(n-1)!}\left(\frac{d}{dx}\right)^{n-1} (z - z_0)^n f(z).$$

- If f has a simple pole at z_0 and g is analytic at z_0, then
$$\mathrm{Res}(fg, z_0) = g(z_0)\mathrm{Res}(f, z_0).$$
- If $g(z)$ has a simple zero at z_0, then $1/g(z)$ has a simple pole at z_0 with residue $1/g'(z_0)$.
- In practice, we can find the residue of a function defined from functions with Laurent series expansions by taking the Laurent series of everything, expanding, and finding the $1/z$ term.
- Suppose that f is analytic in a region R except for a set of isolated singularities. Then if C is a closed curve in A that doesn't go through any of the singularities,
$$\oint_C f(z)\,dz = 2\pi i \sum \text{residues of } f \text{ in } C \text{ counted with multiplicity.}$$
This is the residue theorem, and it can be shown by deforming the contour to a set of small circles about each singularity, and expanding in Laurent series about each one and using the Cauchy integral formula.

Example. Find the residue at $z = 0$ of $f(z) = \sinh(z)e^z/z^5$. The answer is the z^4 term of the Laurent series of $\sinh(z)e^z$, and
$$\sinh(z)e^z = \left(z + \frac{z^3}{3!} + \dots\right)\left(1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \dots\right) = \dots + \left(\frac{1}{4!} + \frac{1}{3!}\right)z^4 + \dots$$
giving the residue $5/24$.

Example. The function $\cot(z) = \cos(z)/\sin(z)$ has a simple pole of residue 1 at $z = n\pi$ for all integers π. To see this, note that $\sin(z)$ has simple zeroes at $z = n\pi$ and its derivative is $\cos(z)$, so $1/\sin(z)$ has residues of $1/\cos(n\pi)$. Multiplying by $\cos(z)$, which is analytic everywhere, cancels these factors giving a residue of 1.

Example. Compute the contour integral along the unit circle of $z^2\sin(1/z)$. There is an essential singularity at $z = 0$, but this doesn't change the computation. The Laurent series for $\sin(1/z)$ is
$$\sin(1/z) = \frac{1}{z} - \frac{1}{3!}\frac{1}{z^3} + \dots$$
which gives a residue of $-1/6$.

Example. The residue at infinity. Suppose that f is analytic in $\mathbb{C}$ with a finite number of singularities, and a curve C encloses every singularity once. Then the contour integral along C is the sum of all the residues. On the other hand, we can formally think of the interior of the contour as the exterior; then we get the same result if we postulate a pole at infinity with residue
$$\mathrm{Res}(f, \infty) = -\frac{1}{2\pi i}\oint_C f(z)\,dz.$$
To compute this quantity, substitute $z = 1/w$ to find
$$\mathrm{Res}(f, \infty) = \frac{1}{2\pi i}\oint_C \frac{f(1/w)}{w^2}\,dw$$
where C is now negatively oriented. Now, $f(1/w)$ has no poles inside the curve C, so the only possible pole is at $w = 0$. Then
$$\mathrm{Res}(f, \infty) = -\mathrm{Res}(f(1/w)/w^2, 0)$$
which may be much easier to compute. Under this language, z has a simple pole at infinity, while e^z has an essential singularity at infinity.

3.5 Application to Real Integrals

In this section, we apply our theory to the evaluation of real integrals.

- In order to express real integrals over the real line in terms of contour integrals, we will have to close the contour. This is easy if the decay is faster than $1/|z|$ in either the upper or lower-half plane by the ML inequality.
- Another common situation is when the function is oscillatory, e.g. it is of the form $f(z)e^{ikz}$. If $|f(z)|$ does not decay faster than $1/|z|$, the ML inequality does not suffice. However, since the oscillation on the real axis translates to decay in the imaginary direction, if we use a square contour bounded by $z = \pm L$, the vertical sides are founded by $|f(L)|/k$ and the top side is exponentially small, so the contributions vanish as $L \to \infty$ as desired.

Example. We compute

$$I = \int_{-\infty}^{\infty} \frac{dx}{x^4 + 1}.$$

We close the contour in the upper-half plane; the decay is $O(1/|z|^4)$ so the semicircle does not contribute. The two poles are at $z_1 = e^{i\pi/4}$ and $z_2 = e^{3i\pi/4}$. An easy way to compute the residues is with L'Hopital's rule,

$$\operatorname{Res}(f, z_1) = \lim_{z \to z_1} \frac{z - z_1}{1 + z^4} = \lim_{z \to z_1} \frac{1}{4z^3} = \frac{e^{-3i\pi/4}}{4}, \quad \operatorname{Res}(f, z_2) = \frac{e^{-i\pi/4}}{4}, \quad I = \frac{\pi}{\sqrt{2}}.$$

Example. For $b > 0$, we compute

$$I = \int_{-\infty}^{\infty} \frac{\cos(x)}{x^2 + b^2}\, dx.$$

For convenience we replace $\cos(x)$ with e^{ix} and take the real part at the end. Now, the function decays faster than $1/|z|$ in the upper-half plane, so we close the contour there. The contour contains the pole at $z = ib$ which has residue $e^{-b}/2ib$, giving $I = \pi e^{-b}/b$.

Example. Integrals over angles can be replaced with contour integrals over the unit circle. We let

$$z = e^{i\theta}, \quad dz = iz d\theta, \quad \cos\theta = \frac{z + 1/z}{2}, \quad \sin\theta = \frac{z - 1/z}{2i}.$$

For example, we can compute

$$I = \int_0^{2\pi} \frac{d\theta}{1 + a^2 - 2a\cos\theta}, \quad |a| \neq 1.$$

Making the above substitutions and some simplifications, we have

$$I = \int_C \frac{dz}{-ia(z - a)(z - 1/a)} = \begin{cases} 2\pi/(a^2 - 1) & |a| > 1, \\ -2\pi/(a^2 - 1) & |a| < 1. \end{cases}$$

It is clear this method works for any trigonometric integral over $[0, 2\pi)$.

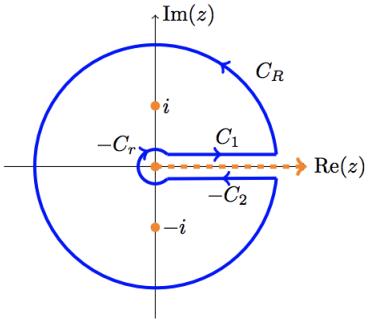
Example. An integral with a branch cut. Consider

$$I = \int_0^\infty \frac{x^{1/3}}{1+x^2}\,dx.$$

We will place the branch cut on the positive real axis, so that for $z = re^{i\theta}$, we have

$$z^{1/3} = r^{1/3}e^{i\theta/3}, \quad 0 \le \theta < 2\pi.$$

We choose a keyhole contour that avoids the branch cut.



The desired integral I is the integral over C_1, while the integrals over C_r and C_R go to zero. The integral over C_2 is on the other end of the branch cut, and hence is $-e^{2\pi i/3}I$. Finally, including the contributions of the two poles gives $I = \pi/\sqrt{3}$.

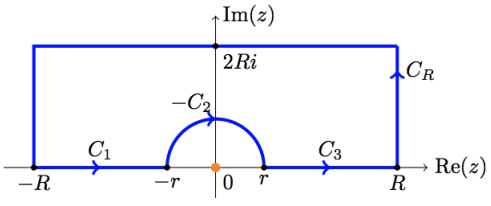
Example. The Cauchy principal value. We compute

$$I = \int_{-\infty}^\infty \frac{\sin(x)}{x}\,dx.$$

This is the imaginary part of the contour integral

$$I' = \int_C \frac{e^{iz}}{z}\,dz$$

where the contour along the real line is closed by a semicircle. The integrand blows up along the contour, since it goes through a pole; to fix this, we define the principal value of the integral I' to be the limit $\lim_{r\to 0} I'(r)$ where a circle of radius r about the origin is deleted from the contour. This is equal to I because the integrand $\sin(x)/x$ is not singular at the origin; in more general cases where the original integrand is singular, the value of the integral is defined as the principal value.



Now consider the contour above. In the limit $r \to 0$, we have $I' = \pi i$ because it picks up "half of the pole", giving $I = \pi$. More generally, if the naive contour "slices" through a pole, the principal value picks up i times the residue times the angle subtended.

Note. The idea of a principal value works for both real and complex integrals. In the case of a real integral, we delete a small segment centered about the divergence. The principal value also exists for integrals with bounds at $\pm\infty$, by setting the bounds to $-R$ and R and taking $R \to \infty$.

3.6 Conformal Transformations

In this section, we apply conformal transformations.

- A conformal map on the complex plane $f(z)$ is a map so that the tangent vectors at any point z_0 are mapped to the tangent vectors at $f(z_0)$ by a nonzero scaling and proper rotation. Informally, this means that conformal maps preserve angles.
- As we've seen, $f(z)$ is automatically conformal if it is holomorphic with nonzero derivative; the scaling and rotation factor is $f'(z_0)$.
- The Riemann mapping theorem states that if a region A is simply connected, and not the entire complex plane, then there exists a bijective conformal map between A and the unit disk; we say the regions are conformally equivalent.
- The proof is rather technical, but is useful to note a few specific features.
  - We cannot take $A = \mathbb{C}$, by Liouville's theorem.
  - There are three real degrees of freedom in the map, which corresponds to the fact that there is a three-parameter family of maps from the unit disk to itself.
  - If A is bounded by a simple closed curve, which may pass through the point at infinity, we may use these degrees of freedom to specify the images of three boundary points.
  - Alternatively, we may specify the image of one interior point of A, and the image of a direction at that point.
  - In practice, we could use "canonical domains" other than the unit disc; one common one is the upper half-plane, in which case we usually fix points to map to 0, 1, and ∞.
  - The theorem guarantees the mapping is conformal in the interior of A, but not necessarily its boundary, where singularities are needed to smooth out corners and cusps.
  - Since conformal maps preserve angles, including their orientation, a curve traversing ∂A with the interior of A to its right maps to a curve traversing the image of ∂A satisfying the same property.
- A useful set of conformal transformations are the fraction linear transformations, or Mobius transformations
$$T(z) = \frac{az+b}{cz+d}, \quad ad - bc \neq 0.$$
Note that when $ad - bc = 0$, then $T(z)$ is constant. Mobius transformations can also be taken to act on the extended complex plane, with
$$T(\infty) = \frac{a}{c}, \quad T(-d/c) = \infty.$$
They are bijective on the extended complex plane, and conformal everywhere except $z = -d/c$.
- When $c = 0$, we get scalings and rotations. The map $T(z) = 1/z$ flips circles inside and outside of the unit circle. As another example,
$$T(z) = \frac{z-i}{z+i}$$
maps the real axis to the unit circle, and hence maps the upper half-plane to the unit disk.

- In general, a Mobius transformation maps generalized circles to generalized circles, where generalized circles include straight lines. To show this, note that it is true for scaling and rotation, so we only need to prove it for inversions, which can be done by components. For example, inversion maps a circle passing through the origin to a linear that doesn't.

- A very useful fact is that Mobius transformations can be identified with matrices,

$$T(z) = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

  so that composition of Mobius transformations is matrix multiplication. Since we can always scale $ad - bc$ to one, and then further multiply all the coefficients by -1, the set of Mobius transformations is $PSL_2(\mathbb{C}) = SL_2(\mathbb{C})/\{\pm I\}$.

- The subset of Mobius transformations that map the upper half-plane to itself turn out to be the ones where a, b, c, and d are all real, and $ad - bc = 1$. Then the group of conformal automorphisms of the upper half-plane contains $PSL_2(\mathbb{R})$.

- In fact, these are all of the conformal automorphisms of the upper half-plane. To prove this, one typically shows using the Schwartz lemma that the conformal automorphisms of the disk take the form

$$T(z) = \lambda \frac{z - a}{\overline{a}z - 1}, \quad |\lambda| = 1, \quad |a| < 1$$

  and then notes that the upper half-plane is conformally equivalent to the disk.

- Given any three distinct points (z_1, z_2, z_3), there exists a Mobius transformation that maps them to (w_1, w_2, w_3). To see this, note that we can map (z_1, z_2, z_3) to $(0, 1, \infty)$ by

$$T(z) = \frac{(z - z_1)(z_2 - z_3)}{(z - z_3)(z_2 - z_1)}$$

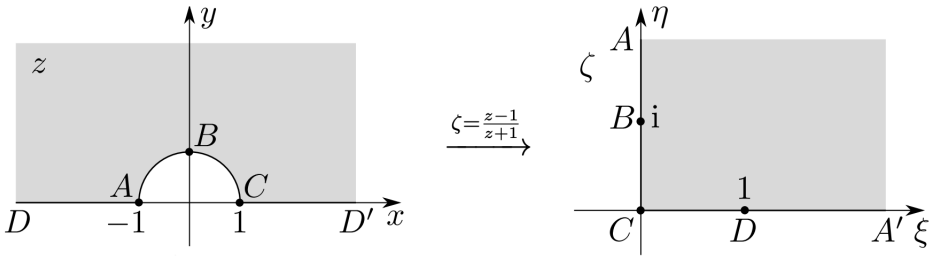
  and this map is invertible, giving the result.

Note. A little geometry. The reflection of a point in a line is the unique point so that any generalized circle that goes through both points intersects the line perpendicularly. We define the reflection of a point in a generalized circle in the same way. To prove that this reflection is unique, note that since Mobius transformations preserve generalized circles and angles, they preserve the reflection property; however we can use a Mobius transformation to map a given circle to a line, then use uniqueness of reflection in a line.

Reflection in a circle is called inversion in the context of Euclidean geometry. Our "inversion" map $z \mapsto 1/z$ is close, but it actually corresponds to an inversion about the unit circle followed by a reflection about the real axis. The inversion along would be $z \mapsto 1/\overline{z}$.

Example. Suppose two circles C_1 and C_2 do not intersect; we would like to construct a conformal mapping that makes them concentric. To do this, let z_1 and z_2 be reflections of each other in both circles – it is easier to see such points exist by mapping C_1 to a line and then mapping back. Now, by a conformal transformation we can map z_1 to zero and z_2 to infinity, which means both centers must end up centered at zero.

Example. Find a map from the upper half-plane with a semicircle removed to a quarter-plane.



We will use a Mobius transformation. The trick is to look at how the boundary must be mapped. We have right angles at A and C, but only one right angle in the image; we can achieve this by mapping A to infinity and C to zero, so

$$z \mapsto \zeta = \frac{z-1}{z+1}.$$

To verify the boundary is correct, we note that ABC and CDA are still generalized circles after the mapping, and verify that B and D are mapped into the imaginary and real axes, respectively. More generally, if we need to change the angle at the origin by a factor α, we can compose by a monomial $z \mapsto z^\alpha$.

Example. Map the upper half plane to itself, permuting the points $(0, 1, \infty)$. We must use Mobius maps with real coefficients. Since orientation is preserved, we can only perform even permutations. The answers are

$$\zeta = \frac{1}{1-z}, \quad (0, 1, \infty) \mapsto (1, \infty, 0)$$

and

$$\zeta = \frac{z-1}{z}, \quad (0, 1, \infty) \mapsto (\infty, 0, 1).$$

Example. The Dirichlet problem is to find a harmonic function on a region A given specified values on the boundary of A. For example, let A be the unit disk with boundary condition

$$u(e^{i\theta}) = \begin{cases} 1 & 0 < \theta < \pi, \\ 0 & \pi < \theta < 2\pi. \end{cases}$$

The problem can be solved by conformal mapping. We apply $T(z) = (z-i)/(z+i)$, which maps the real axis to the unit circle. Then A maps to the upper half-plane with boundary condition $u(x) = \theta(-x)$, and an explicit solution is $u(x, y) = \theta/\pi = \mathrm{Im}(\log(z))/\pi$.

More generally, consider a piecewise constant boundary condition $u(e^{i\theta})$. Then the conformally transformed solution is a sum of pieces of the form $\log(z - x_0)$. An arbitrary boundary condition translates to a weighted integral of $\log(z - x)$ over real x.

Example. The general case of flow around a circle. Suppose $f(z)$ is a complex velocity potential. Singularities of the potential correspond to sources or vertices. If there are no singularities for $|z| < R$, then the Milne-Thomson circle theorem states that

$$\Phi(z) = f(z) + \overline{f(R^2/\overline{z})}$$

is a potential for a flow with a streamline on $|z| = R$ and the same singularities; it is what the potential would be if we introduced a circular obstacle but kept everything else the same. We've already seen the specific example of uniform flow around a circle, where $f(z) = z$.

To see this, note that $f(z)$ may be expanded in a Taylor series

$$f(z) = a_0 + a_1 z + a_2 z^2 + \dots$$

which converges for $|z| \le R$. Then $\overline{f(R^2/\overline{z})}$ has a Laurent series

$$\overline{f(R^2/\overline{z})} = \overline{a_0} + \overline{a_1}\frac{R^2}{z} + \overline{a_2}\left(\frac{R^2}{z}\right)^2 + \dots$$

which converges for $|z| \ge R$, so no new physical singularities are introduced by adding it. To see that $|z| = R$ is a streamline, note that

$$\Phi(Re^{i\theta}) = f(Re^{i\theta}) + \overline{f(Re^{i\theta})} \in \mathbb{R}.$$

Then the stream function $\operatorname{Im}\Phi$ has a level set on $|z| = R$, namely zero.

Example. The map $f(z) = e^{iz}$ takes the half-strip $\operatorname{Im}(z) > 0$, $\operatorname{Re}(z) \in (-\pi/2, \pi/2)$ to the right half-disc. In general, since the complex exponential is periodic in 2π, it is useful for mapping from strips. The logarithm $f(z) = \log z$ maps to strips. For example, it takes the upper half-plane to the strip $\operatorname{Im}(z) \in (0, \pi)$. It also maps the upper half-disc to the left half of this strip.

Example. The Joukowski map is

$$f(z) = \frac{1}{2}\left(z + \frac{1}{z}\right).$$

This map takes the unit disc to the entire complex plane; to see this, we simply note that the unit circle is mapped to the slit $x \in (-1, 1)$. This does not contradict the Riemann mapping theorem, because $f(z)$ is singular at $z = 0$. We create corners at $z = \pm 1$, which is acceptable because f' vanishes at these points. Since the Joukowski map obeys $f(z) = f(1/z)$, the region outside the unit disc is also mapped to the complex plane. The Joukowski transform is useful in aerodynamics, because it maps off-center circles to shapes that look like airfoils. The flow past these airfoils can be solved by applying the inverse transform, since the flow around a sphere is known analytically.

3.7 Additional Topics

Next, we introduce the argument principle, which is useful for counting poles and zeroes.

- Previously, we have restricted to simple closed curves, as these wind about any point at most once. However, we may now define the winding number or index

  $$\operatorname{Ind}(\gamma, z_0) = \frac{1}{2\pi i}\int_\gamma \frac{dz}{z - z_0}$$

  for any closed curve γ that does not contain z_0. This follows from Cauchy's integral theorem; intuitively, the integrand is $d(\log(z - z_0))$ and hence counts the number of windings by the net phase change.

- For any integer power $f(z) = z^n$, we have

  $$\int_\gamma \frac{f'(z)}{f(z)}\,dz = 2\pi n, \quad \operatorname{Ind}(\gamma, 0) = 1.$$

  This is because the integrand is df/f, so it counts the winding number of f about the origin along the curve. Moreover, $(fg)'/(fg) = f'/f + g'/g$, so other zeroes or poles contribute additively.

- Formalizing this result, for a meromorphic function f and a simple closed curve γ not going through any of the poles, we have the argument principle

$$\int_\gamma \frac{f'(z)}{f(z)}\,dz = 2\pi(\text{zeroes minus poles}) = 2\pi i \operatorname{Ind}(f \circ \gamma, 0)$$

  where the zeroes and poles are weighted by their order.

- Rouche's theorem states that for meromorphic functions f and h and a simple closed curve γ not going through any of their poles, if $|h| < |f|$ everywhere on γ, then

$$\operatorname{Ind}(f \circ \gamma, 0) = \operatorname{Ind}((f+h) \circ \gamma, 0).$$

  Intuitively, this follows from the picture of a 'dog on a short leash' held by a person walking around a tree. It can be shown using the argument principle; interpolating between f and $f+h$, the integral varies continuously, so the index must stay the same.

- A useful corollary of Rouche's theorem is the case of holomorphic f and h, which gives

$$\text{zeroes of } f \text{ in } \gamma = \text{zeroes of } f+h \text{ in } \gamma.$$

  For example, suppose we wish to show that $z^5 + 3z + 1$ has all five of its zeroes within $|z| = 2$.

- This same reasoning provides a different proof of the fundamental theorem of algebra. We let $f(z) \propto z^n$ be the higher-order term in the polynomial and let h be the rest. Then within a sufficiently large circle, $f + h$ must contain n zeroes.

Next, we discuss analytic continuation.

- Suppose that f is holomorphic in a connected region R and vanishes in a sequence of distinct points $\{w_i\}$ with a limit point in R. Then f is zero.

  To see this, suppose that f is nonzero. Then it has a Taylor series expansion about the limit point, but we've shown that zeroes of functions with Taylor series are isolated by continuity.

- As a corollary, if f and g are holomorphic on a connected region R and agree on a set of points with a limit point in R, then they are equal. An analytic continuation of a real function is a holomorphic function that agrees with it on the real axis; this result ensures that analytic continuation is unique, at least locally.

- One must be more careful globally. For example, consider the two branches of the logarithm with a branch cut along the positive and negative real axis. The two functions agree in the first quadrant, but we cannot conclude they agree in the fourth quadrant, because the region where they are both defined is the complex plane minus the real axis, which is not connected.

- These global issues are addressed by the monodromy theorem, which states that analytic continuation is unique (i.e. independent of path) if the domain we use is simply connected. This does not hold for the logarithm, because it is nonanalytic at the origin.

- As another example, the factorial function doesn't have a unique analytic continuation, because the set of positive integers doesn't have a limit point. But the gamma function, defined as an integral expression for positive real arguments, does have a unique analytic continuation. (This statement is sometimes mangled to "the gamma function is the unique analytic continuation of the factorial function", which is incorrect.)

- Consider a Taylor series with radius of convergence R. This defines a holomorphic function within a disk of radius R and hence can be analytically continued, e.g. by taking the Taylor series about a different point in the disk.
- As an example where this fails, consider $f(z) = z+z^2+z^4+\ldots$, which has radius of convergence 1. The function satisfies the recurrence relation $f(z) = z + f(z^2)$, which implies that $f(1)$ is divergent. By repeatedly applying this relation, we see that $f(z)$ is divergent if $z^{2^n} = 1$, so the divergences are dense on the boundary of the unit disk. These divergences form a 'natural boundary' beyond which analytic continuation is not possible.

4 Linear Algebra

4.1 Exact Sequences

In this section, we rewrite basic linear algebra results using exact sequences. For simplicity, we only work with finite-dimensional vector spaces.

- Consider vector spaces V_i and maps $\varphi_i : V_i \to V_{i+1}$, which define a sequence

$$\dots \to V_{i-1} \xrightarrow{\varphi_{i-1}} V_i \xrightarrow{\varphi_i} V_{i+1} \to \dots.$$

  We say the map is exact at V_i if $\operatorname{im} \varphi_{i-1} = \ker \varphi_i$. The general intuition is that this means V_i is 'made up' of its neighbors V_{i-1} and V_{i+1}.

- We write 0 for the zero-dimensional vector space. For any other vector space V, there is only one possible linear map from V to 0, or from 0 to V.

- A short exact sequence is an exact sequence of the form

$$0 \to V_1 \xrightarrow{\varphi_1} V_2 \xrightarrow{\varphi_2} V_3 \to 0.$$

  The sequence is exact at V_1 iff φ_1 is injective and exact at V_2 iff φ_2 is surjective.

- As an example, the exact sequence

$$0 \to V_1 \xrightarrow{\varphi} V_2 \to 0$$

  requires φ to be an isomorphism.

- If $T : V \to W$ is surjective, then we have the exact sequence

$$0 \to \ker T \xrightarrow{i} V \xrightarrow{T} W \to 0$$

  where i is the inclusion map.

- Given this short exact sequence, there exists a linear map $S : W \to V$ so that $T \circ S = 1$. We say the exact sequence splits, and that S is a section of T. To see why S exists, take any basis $\{f_i\}$ of W. Then there exist e_i so that $T(e_i) = f_i$, and we simply define $S(f_i) = e_i$.

- Using the splitting, we have the identity

$$V = \ker T \oplus S(W)$$

  which is a refinement of the rank-nullity theorem; this makes it clear exactly how V is determined by its neighbors in the short exact sequence. Note that we always have $\dim V_i = \dim V_{i-1} + \dim V_{i+1}$, but by using the splitting, we get a direct decomposition of V itself.

- It is tempting to write $V = \ker T \oplus W$, but this is technically incorrect because W is not a subspace of V. We will often ignore this distinction below.

Example. Quotient spaces. Given a subspace $W \subset V$ we define the equivalence relation $v \sim w$ if $v - w \in W$. The set of equivalence classes $[v]$ is called V/W and we define the projection map $\pi : V \to V/U$ by $\pi(v) = [v]$. Then we have an exact sequence

$$0 \to U \xrightarrow{i} V \xrightarrow{\pi} V/U \to 0$$

which implies $\dim(V/U) = \dim V - \dim U$.

Example. The kernel of $T : V \to W$ measures the failure of T to be injective; the cokernel $\operatorname{coker} T = W/\operatorname{im} T$ measures the failure of T to be surjective. Then we have the exact sequence

$$0 \to \ker T \xrightarrow{i} V \xrightarrow{T} W \xrightarrow{\pi} \operatorname{coker} T \to 0$$

where π projects out $\operatorname{im} T$.

Example. Exact sequences and chain complexes. Consider the chain complex with boundary operator ∂. The condition $\operatorname{im} \varphi_{i-1} \subset \ker \varphi_i$ states that the composition $\varphi_i \circ \varphi_{i-1}$ takes everything to zero, so $\partial^2 = 0$. The condition $\ker \varphi_i \subset \operatorname{im} \varphi_{i-1}$ implies that the homology is trivial. Thus, homology measures the failure of the chain complex to be exact.

Next, we prove a familiar theorem using the language of exact sequences.

Example. We claim every space with a symmetric nondegenerate bilinear form g has an orthonormal basis, i.e. a set $\{v_i\}$ where $g(v_i, v_j) = \pm\delta_{ij}$. We prove only the real case for simplicity. Let $\dim V = k$ and suppose we have an orthonormal set of $k-1$ vectors e_i. Defining the projection map

$$\pi(v) = \sum_{i=1}^{k-1} g(e_i, e_i) g(e_i, v) e_i$$

we have the exact sequence

$$0 \to W^\perp \xrightarrow{i} V \xrightarrow{\pi} W \to 0$$

where $W^\perp = \ker \pi$ is the orthogonal complement of W. Now, we claim that g is nondegenerate when restricted to $W^\perp$. To see this, note that if $g(w_1, w_2) = 0$ for all $w_2 \in W$, then $g(w_1, v) = 0$ for all vectors $v \in V$, so w_1 must be zero by nondegeneracy. The result follows by induction.

We can also give a more direct proof. Given a set of vectors $\{v_i\}$, define the Gram matrix G to have components

$$g_{ij} = g(e_i, e_j).$$

In the context of physics, this is simply the metric in matrix form. Then the form is nondegenerate if and only if G has trivial nullspace, as

$$G\mathbf{v} = 0 \leftrightarrow g(v_i e_i, e_j) = 0.$$

By the spectral theorem, we can choose a basis so that G is diagonal; by the result above, its diagonal entries are nonzero, so we can scale them to be ± 1. This yields the desired basis. Sylvester's theorem states that the total number of 1's and -1's in the final form of G is unique. We say g is an inner product if it is positive definite, i.e. there are no -1's.

The determinate of the gram matrix, called the Grammian, is a useful concept. For example, for any collection of vectors $\{v_i\}$, the vectors are independent if and only if the Grammian is nonzero.

4.2 The Dual Space

Next, we consider dual spaces and dual maps.

- Let the dual space V^* be the set of linear functionals f on V. For finite-dimensional V, V and V^* are isomorphic but there is no natural map between them.
- For infinite-dimensional V, V and V^* are generally not isomorphic. One important exception is when V is a Hilbert space, which is crucial in quantum mechanics.
- We always have $V^{**} = V$, with the natural isomorphism $v \mapsto (f \mapsto f(v))$.
- When an inner product is given, we can define an isomorphism ψ between V and V^* by

  $$v \mapsto f_v, \quad f_v(\cdot) = g(v, \cdot).$$

  By nondegeneracy, ψ is injective; since V and V^* have the same dimension, this implies it is surjective as well.
- In the context of a complex vector space, there are some extra subtleties: the form can only be linear in one argument, say the second, and is antilinear in the other. Then the map ψ indeed maps vectors to linear functionals, but it does so at the cost of being antilinear itself.
- The result above also holds for (infinite-dimensional) Hilbert spaces, where it is called the Riesz lemma; it is useful in quantum mechanics.
- Given a linear map $A : V \to W$, there is a dual map $A^* : W^* \to V^*$ defined by

  $$(A^* f)(v) = f(Av).$$

  The dual map is often called the transpose map. To see why, pick arbitrary bases of V and W with the corresponding dual bases of V^* and W^*. Then in components,

  $$f_i A_{ij} v_j = (A^* f)_j v_j = (A^*_{ji} f_i) v_j$$

  which implies that $A_{ij} = A^*_{ji}$. That is, expressed in terms of matrices in the appropriate bases, they are transposes of each other.
- Given an inner product g on V and a linear map $A : V \to V$, there is another linear map $A^\dagger : V \to V$ called the adjoint of A, defined by

  $$g(A^\dagger w, v) = g(w, Av).$$

  By working in an orthonormal basis and expanding in components, the matrix elements satisfy

  $$A^\dagger_{ij} = A^*_{ji}$$

  so that the matrices are conjugate transposes.
- In the case where $V = W$ and V is a real vector space, the matrix representations of the dual and adjoint coincide, but they are very different objects. In quantum mechanics, we switch between a map and its dual constantly, but taking the adjoint has a nontrivial effect.

4.3 Determinants

We now review some facts about determinants.

- Defining the ij minor of a matrix A_{ij} to be $\widetilde{A}_{ij} = \det A(i|j)$ where $A(i|j)$ is A with its i^{th} row and j^{th} column removed. Define the adjugate matrix $\operatorname{adj} A$ to have elements
$$(\operatorname{adj} A)_{ij} = \widetilde{A}_{ji}.$$
- By induction, we can show that the determinate satisfies the Laplace expansion formula
$$\det A = \sum_{j=1}^{n} A_{ij}\widetilde{A}_{ij}.$$
More generally, we have
$$\sum_{j=1}^{n} A_{ij}\widetilde{A}_{kj} = \delta_{jk} \det A$$
where we get a zero result when $j \neq k$ because we are effectively taking the determinant of a matrix with identical rows.
- Therefore, removing the components, we have
$$A(\operatorname{adj} A) = (\operatorname{adj} A)A = (\det A)I$$
so that the adjugate gives a formula for the inverse, when it exists! When it doesn't exist, $\det A = 0$, so both sides are simply zero.
- Applying this formula to $A\mathbf{x} = \mathbf{b}$, we have $\mathbf{x} = (\operatorname{adj} A)\mathbf{b}/\det A$. Taking components gives Cramer's rule
$$x_i = \det A^{(i)}/\det A$$
where $A^{(i)}$ is A with the i^{th} column replaced with $\mathbf{b}$.
- The Laplace expansion formula gives us a formula for the derivative of the determinant,
$$\frac{\partial}{\partial A_{ij}}(\det A) = \widetilde{A}_{ij}.$$
In the case $\det A \neq 0$, this gives the useful result
$$\frac{\partial}{\partial A_{ij}}(\det A) = (\det A)(A^{-1})_{ji}.$$

Note. The final result above can also be derived by the identity
$$\log\det A = \operatorname{tr}\log A.$$
Taking the variation of both sides,
$$\delta(\log\det A) = \operatorname{tr}\log(A + \delta A) = \operatorname{tr} A^{-1}\delta A$$
which implies
$$\frac{\partial}{\partial A}(\log\det A) = (A^{-1})^T$$
in agreement with our result. The crucial step is the simplification of the log, which is not valid in general, but works because of the cyclic property of the trace. More precisely, if we expand the logarithm order by order (keeping only terms up to first-order in δA), the cyclic property always allows us to bring the factor of δA to the back, so A and δA effectively commute.

4.4 Endomorphisms

An endomorphism is a linear map from a vector space V to itself. The set of such endomorphisms is called $\mathrm{End}(V)$ in math, and the set of linear operators on V in physics. We write abstract endomorphisms with Greek letters; for example, the identity map ι has matrix representation I.

- Two matrix representations of an endomorphism differ by conjugation by a change of basis matrix, and any two matrices related this way are called similar.
- We define the trace and determinant of an endomorphism by the trace and determinant of any matrix representation; this does not depend on the basis chosen.
- We define the λ-eigenspace of α as $E(\lambda) = \ker(\alpha - \lambda\iota)$.
- We define the characteristic polynomial of α by

  $$\chi_\alpha(t) = \det(t\iota - \alpha).$$

  This is a monic polynomial with degree $\dim V$, and its roots correspond to eigenvalues. Similarly, we can define the characteristic polynomial of a matrix as $\chi_A(t) = \det(tI - A)$, and it is invariant under basis change.
- The eigenspaces $E(\lambda_i)$ are independent. To prove this, suppose that $\sum_i x_i = 0$ where $x_i \in E(\lambda_i)$. Then we may project out all but one component,

  $$\sum_i x_i = \prod_{k\neq j}(\alpha - \lambda_k\iota)(x_i) = \prod_{k\neq j}(\lambda_j - \lambda_k)x_j \propto x_j.$$

  For the left-hand side to be zero, x_j must be zero for all j, giving the result.
- We say α is diagonalizable when its eigenspaces span all of V, i.e. $V = \oplus_i E_i$. Equivalently, α has a diagonal matrix representation, produced by choosing a basis of eigenvectors.

Diagonalizability is an important property. To approach it, we introduce the minimal polynomial.

- Polynomial division: for any polynomials f and g, we may write $f(t) = q(t)g(t) + r(t)$ where $\deg r < \deg g$.
- As a corollary, whenever f has a root λ, we can extract a linear factor $f(t) = (t - \lambda)g(t)$. The fundamental theorem of algebra tells us that f will always have at least one root; repeating this shows that all polynomials split into linear factors in $\mathbb{C}$.
- The endomorphism α is diagonalizable if and only if there is a nonzero polynomial $p(t)$ with distinct linear factors such that $p(\alpha) = 0$. Intuitively, each such linear factor $(x - \lambda_i)$ projects away the eigenspace E_i, and since $p(\alpha) = 0$, the E_i must span all of V.

  Proof: The backward direction is simple. To prove the forward direction, we define projection operators. Let the roots be λ_i and let

  $$q_j(t) = \prod_{i\neq j}\frac{t - \lambda_i}{\lambda_j - \lambda_i} \quad \to \quad q_j(\lambda_i) = \delta_{ij}.$$

  Then $q(t) = \sum_j q_j(t) = 1$. Now define the operators $\pi_j = q_j(\alpha)$. Since $(\alpha - \lambda_j\iota)\pi_j \propto p(\alpha) = 0$, π_j projects onto the λ_j eigenspace. Since the projectors sum to $\pi_j(v) = q(\alpha) = \iota$, the eigenspaces span V.

- Define the minimal polynomial of α to be the non-zero monic polynomial $m_\alpha(t)$ of least degree so that $m_\alpha(\alpha) = 0$. Such polynomials exist with degree bounded by n^2, since $\mathrm{End}(V)$ has dimension n^2.
- For any polynomial p, $p(\alpha) = 0$ if and only if m_α divides p.

  Proof: using division, we have $p(t) = q(t)m_\alpha(t) + r(t)$. Plugging in α, we have $r(\alpha) = 0$, but r has smaller degree than m_α, so it must be zero, giving the result.
- As a direct corollary, the endomorphism α is diagonalizable if and only if $m_\alpha(t)$ is a product of distinct linear factors.
- Every eigenvalue is a root of the minimal polynomial, and vice versa.

Example. Intuition for the above results. Consider the matrices

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

Then A satisfies $t - 1$, but B does not; instead its minimal polynomial is $(t-1)^2$. To understand this, note that

$$C = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$$

has minimal polynomial t^2, and its action consists of taking the basis vectors $\hat{e}_2 \to \hat{e}_1 \to 0$, which is why it requires two powers of t to vanish. This matrix is not diagonalizable because the only possible eigenvalue is zero, but only $\hat{e}_1$ is an eigenvector; $\hat{e}_2$ is a 'generalized eigenvector' that instead eventually maps to zero. As we'll see below, such generalized eigenvectors are the only obstacle to diagonalizability.

Prop (Schur)**.** Let V be a finite-dimensional complex vector space and let $\alpha \in \mathrm{End}(V)$. Then there is a basis where α is upper triangular.

Proof. By the FTA, the characteristic polynomial has a root, and hence there is an eigenvector. By taking this as our first basis element, all entries in the first column are zero except for the first. Quotienting out the eigenspace gives the result by induction.

Theorem (Cayley-Hamilton)**.** Let V be a finite-dimensional vector space over $\mathbb{F}$ and let $\alpha \in \mathrm{End}(V)$. Then $\chi_\alpha(\alpha) = 0$, so m_α divides χ_α.

Proof. [$\mathbb{F} = \mathbb{C}$] We use Schur's theorem, and let α be represented with A, which has diagonal elements λ_i. Then $\chi_\alpha(t) = \prod_i (t - \lambda_i)$. Applying the factor $(\alpha - \lambda_n)$ sets the basis vector $\hat{e}_n$ to zero. Subsequently applying the factor $(\alpha - \lambda_{n-1})$ sets the basis vector $\hat{e}_{n-1}$ to zero, and does not map anything to $\hat{e}_n$ since A is upper triangular. Repeating this logic, $\chi_\alpha(\alpha)$ sets every basis vector to zero, giving the result. This also proves the Cayley-Hamilton theorem for $\mathbb{F} = \mathbb{R}$, because every real polynomial can be regarded as a complex one.

Proof. [General $\mathbb{F}$] A tempting false proof of the Cayley-Hamilton theorem is to simply directly substitute $t = A$ in $\det(tI - A)$. This doesn't make sense, but we can make it make sense by explicitly expanding the characteristic polynomial. Let $B = tI - A$. Then

$$\mathrm{adj}\, B = B_{n-1}t^{n-1} + \ldots + B_1 t + B_0.$$

Using $B(\operatorname{adj} B) = (\det B)I - \chi_A(t)I$, we have

$$(tI - A)(B_{n-1}t^{n-1} + \ldots + B_0) = (t^n + a_{n-1}t^{n-1} + \ldots + a_0)I_n$$

where the a_i are the coefficients of the characteristic polynomial. Expanding term by term,

$$A^n B_{n-1} = A^n, \quad A^{n-1}B_{n-2} - A^n B_{n-1} = a_{n-1}A^{n-1}, \quad \ldots \quad , -AB_0 = a_0 I_n.$$

Adding these equations together, the left-hand sides telescope, giving the result.

Proof. [Continuity] Use the fact that Cayley-Hamilton is obvious for diagonalizable matrices, continuity of χ_α, and the fact that diagonalizable matrices are dense in the space of matrices. This is the shortest proof, but has the disadvantage of requiring much more setup.

Example. The minimal polynomial of

$$A = \begin{pmatrix} 1 & 0 & -2 \\ 0 & 1 & 1 \\ 0 & 0 & 2 \end{pmatrix}.$$

We know the characteristic polynomial is $(t-1)^2(t-2)$, and that both 1 and 2 are eigenvalues. Thus by Cayley-Hamilton the minimal polynomial is $(t-1)^a(t-2)$ where a is 1 or 2. A direct calculation shows that $a = 1$ works; hence A is diagonalizable.

Next, we move to Jordan normal form.

- Let λ be an eigenvalue of α. Its algebraic multiplicity a_λ is its multiplicity as a root of $\chi_\alpha(t)$. Its geometric multiplicity is $g_\lambda = \dim E_\alpha(\lambda)$. We also define c_λ as its multiplicity as a root of $m_\alpha(t)$.
- If $a_\lambda = g_\lambda$ for all λ, then α is diagonalizable. As shown earlier, this is equivalent to $c_\lambda = 1$ for all eigenvalues λ.
- As we'll see, the source of nondiagonalizability is Jordan blocks, i.e. matrices of the form

  $$J_n(\lambda) = \lambda I_n + K_n$$

  where K_n has ones directly above the main diagonal. These blocks have $g_\lambda = 1$ but $a_\lambda = c_\lambda = n$. A matrix is in Jordan normal form if it is block diagonal with Jordan blocks.
- It can be shown that every matrix is similar to one in Jordan normal form. A sketch of the proof is to split the vector space into 'generalized eigenspaces' (the nullspaces of $(A - \lambda I)^k$ for sufficiently high k), so that we can focus on a single eigenvalue, which can be shifted to zero without loss of generality.

Example. All possible Jordan normal forms of 3×3 matrices. We have the diagonalizable examples,

$$\operatorname{diag}(\lambda_1, \lambda_2, \lambda_3), \quad \operatorname{diag}(\lambda_1, \lambda_2, \lambda_2), \quad \operatorname{diag}(\lambda_1, \lambda_1, \lambda_1),$$

as well as

$$\begin{pmatrix} \lambda_1 & & \\ & \lambda_2 & 1 \\ & & \lambda_2 \end{pmatrix}, \quad \begin{pmatrix} \lambda_1 & & \\ & \lambda_1 & 1 \\ & & \lambda_1 \end{pmatrix}, \quad \begin{pmatrix} \lambda_1 & 1 & \\ & \lambda_1 & 1 \\ & & \lambda_1 \end{pmatrix}.$$

The minimal polynomials are $(t-\lambda_1)(t-\lambda_2)^2$, $(t-\lambda_1)^2$, and $(t-\lambda_1)^3$, while the characteristic polynomials can be read off the main diagonal. In general, a_λ is the total dimension of all Jordan blocks with eigenvalue λ, c_λ is the dimension of the largest Jordan block, and g_λ is the number of Jordan blocks. The dimension of the λ eigenspace is g_λ, while the dimension of the λ generalized eigenspace is a_λ.

Example. The prototype for a Jordan block is a nilpotent endomorphism that takes

$$\hat{e}_1 \mapsto \hat{e}_2 \mapsto \hat{e}_3 \mapsto 0$$

for basis vectors $\hat{e}_i$. Now consider an endomorphism that takes

$$\hat{e}_1, \hat{e}_2 \mapsto \hat{e}_3 \to 0.$$

At first glance it seems this can't be put in Jordan form, but it can because it takes $\hat{e}_1 - \hat{e}_2 \to 0$. Thus there are actually two Jordan blocks!

Example. Solving the differential equation $\dot{x} = Ax$ for a general matrix A. The method of normal modes is to diagonalize A, from which we can read off the solution $x(t) = e^{At}x(0)$. More generally, the best we can do is Jordan normal form, and the exponential of a Jordan block contains powers of t, so generally the amplitude will grow polynomially. Note that this doesn't happen for mass-spring systems, because there the equivalent of A must be antisymmetric by Newton's third law, so it is diagonalizable.

5 Groups

5.1 Fundamentals

We begin with the basic definitions.

- A group G is a set with an associative binary operation, so that there is an identity e which satisfies $ea = ae = a$ for all $a \in G$, and for every element a there is an inverse a^{-1} so that $aa^{-1} = a^{-1}a = e$. A group is abelian if the operation is commutative.
- There are many important basic examples of groups.
  - Any field $\mathbb{F}$ is a abelian group under addition, while $\mathbb{F}^*$, which omits the zero element, is a abelian group under multiplication.
  - The set of $n \times n$ invertible real matrices forms the group $GL(n, \mathbb{R})$ under matrix multiplication, and it is not abelian.
  - A group is cyclic if all elements are powers g^k of a fixed group element g. The n^{th} cyclic group C_n is the cyclic group with n elements.
  - The dihedral group D_{2n} is the set of symmetries of a regular n-gon. It is generated by rotations r by $2\pi/n$ and a reflection s and hence has $2n$ elements, of the form r^k or sr^k. We may show this using the relations $r^n = s^2 = 1$ and $srs = r^{-1}$.
- We can construct new groups from old.
  - The direct product group $G \times H$ has the operation
    $$(g_1, h_1)(g_2, h_2) = (g_1 g_2, h_1 h_2).$$
    For example, there are two groups of order 4, which are C_4 and the Klein four group $C_2 \times C_2$.
  - A subgroup $H \subseteq G$ is a subset of G closed under the group operations. For example, $C_n \subseteq D_{2n}$ and $C_2 \subseteq D_{2n}$.
  - Note that intersections of subgroups are subgroups. The subgroup generated by a subset S of G, called $\langle S \rangle$ is the smallest subgroup of G that contains S. One may also consider the subgroup generated by a group element, $\langle g \rangle$.
- A group isomorphism $\phi : G \to H$ is a bijection so that $\phi(g_1 g_2) = \phi(g_1)\phi(g_2)$.
- The order of a group $|G|$ is the number of elements it contains, while the order of a group element g is the smallest integer k so that $g^k = e$.
- An equivalence relation $\sim$ on a set S is a binary relation that is reflexive, symmetric, and transitive. The set is thus partitioned into equivalence classes; the equivalence class of $a \in S$ is written as $\overline{a}$ or $[a]$.
- Two elements in a group g_1 and g_2 are conjugate if there is a group element h so that $g_1 = hg_2h^{-1}$. Conjugacy is an equivalence relation and hence splits the group into conjugacy classes.

One of the most important examples is the permutation group.

- The symmetric group S_n is the set of bijections $S \to S$ of a set S with n elements, conventionally written as $S = \{1, 2, \ldots, n\}$, where the group operation is composition.

- An element σ of S_n can be written in the notation

$$\begin{pmatrix} 1 & 2 & \dots & n \\ \sigma(1) & \sigma(2) & \dots & \sigma(n) \end{pmatrix}.$$

  There is an ambiguity of notation, because for $\sigma, \tau \in S_n$ the product $\sigma\tau$ can refer to doing the permutation σ first, as one would expect naively, or to doing τ first, because one would write $\sigma(\tau(i))$ for the image of element i. We choose the former option.

- It is easier to write permutations using cycle notation. For example, a 3-cycle (123) denotes a permutation that maps $1 \to 2 \to 3 \to 1$ and fixes everything else. All group elements are generated by 2-cycles, also called transpositions.

- Any permutation can be written as a product of disjoint cycles. The cycle type is the set of lengths of these cycles, and conjugacy classes in S_n are specified by cycle type, because conjugation merely 'relabels the numbers'.

- Specifically, suppose there are k_i cycles of length ℓ_i. Then the number of permutations with this cycle type is

$$\frac{n!}{\prod_i \ell_i^{k_i} k_i!}$$

  where the first term in the denominator accounts for shuffling within a cycle (since (123) is equivalent to (231)) and the second accounts for exchanging cycles of the same length (since (12)(34) is equivalent to (34)(12)).

- Every permutation can be represented by a permutation matrix. A permutation matrix is even if its permutation matrix has determinant $+1$. Hence by properties of determinants, even and odd permutations are products of an even or odd number of transpositions.

- The subgroup of even permutations is the alternating group $A_n \subseteq S_n$. Note that every even permutation is paired with an odd one, by multiplying by an arbitrary transposition, so $|A_n| = n!/2$. For $n \geq 4$, A_n is not abelian since (123) and (124) don't commute.

- Finally, some conjugacy classes break in half when passing from S_n to A_n. For example, (123) and (132) are not conjugate in A_4, because if $\sigma^{-1}(123)\sigma = (132)$, then $(1\sigma\, 2\sigma\, 3\sigma) = (132)$, which means σ is odd.

Next, we turn to the group theory of the integers $\mathbb{Z}$.

- The integers are the cyclic group of infinite order. To make this very explicit, we may define an isomorphism $\phi(g^k) = k$ for generator g.

- Any subgroup of a cyclic group is cyclic. Let $G = \langle g \rangle$ and $H \subseteq G$. Then if n is the minimum natural number so that $g^n \in H$, we claim $H = \langle g^n \rangle$. For an arbitrary element $g^a \in H$, we may use the division algorithm to write $a = qn + r$, and hence $g^r \in H$. Then we have a contradiction unless $r = 0$.

- In particular, this means the subgroups of $\mathbb{Z}$ are $n\mathbb{Z}$. We define

$$\langle m, n \rangle = \langle \gcf(m, n) \rangle, \quad \langle m \rangle \cap \langle n \rangle = \langle \operatorname{lcm}(m, n) \rangle.$$

We then immediately have Bezout's lemma, i.e. there exist integers u and v so that

$$um + vn = \gcf(m, n).$$

We can then establish the usual properties, e.g. if $x|m$ and $x|n$ then $x|\gcf(m, n)$.

- The Chinese remainder theorem states that if $\gcf(m, n) = 1$, then

  $$C_{mn} \cong C_m \times C_n.$$

  Specifically, if g and h generate C_m and C_n, we claim (g, h) generates $C_m \times C_n$. It suffices to show (g, h) has order mn. Clearly its order divides mn. Now suppose that $(g^k, h^k) = e$. Then $m|k$ and $n|k$, and by Bezout's lemma $um + vn = 1$. But then we have

  $$mn|umk + vnk = k$$

  so mn divides the order, and hence they are equal.

- We write $\mathbb{Z}_n$ for the set of equivalence classes where $a \sim b$ if $n|(a - b)$. Both addition and multiplication are well defined on these classes. Under addition, $\mathbb{Z}_n$ is simply a cyclic group C_n.

- Multiplication is more complicated. By Bezout's lemma, $m \in \mathbb{Z}_n$ has a multiplicative inverse if and only if $\gcf(m, n) = 1$, and we call m a unit. Hence if $\mathbb{Z}_n$ is prime, then it is a field. In general the set of units forms a group $\mathbb{Z}_n^*$ under multiplication.

Next, we consider Lagrange's theorem.

- Let H be a subgroup of G. We define the left and right cosets

  $$gH = \{gh : h \in H\}, \quad Hg = \{hg : h \in H\}$$

  and write G/H to denote the set of (left) cosets. In general, $gH \neq Hg$.

- We see gH and kH are the same coset if $k^{-1}g \in H$. This is an equivalence relation, so the cosets partition the group. Moreover, all cosets have the same size because the map $h \mapsto gh$ is a bijection between H and gH. Thus we have

  $$|G| = |G/H| \cdot |H|.$$

  In particular, we have Lagrange's theorem, $|H|$ divides $|G|$.

- By considering the cyclic group generated by any group element, the order of any group element divides $|G|$. In particular, all groups with prime order are cyclic.

- Fermat's little theorem states that for a prime p where p does not divide a,

  $$a^{p-1} \equiv 1 \bmod p.$$

  This is simply because the order of a in $\mathbb{Z}_p^*$ divides $p - 1$.

- In general, $|\mathbb{Z}_n^*| = \phi(n)$ where ϕ is the totient function, which satisfies

$$\phi(p) = p - 1, \quad \phi(p^k) = p^{k-1}(p-1), \quad \phi(mn) = \phi(m)\phi(n) \text{ if } \gcf(m,n) = 1.$$

  Then Euler's theorem generalizes Fermat's little theorem to

$$a^{\phi(n)} \equiv 1 \bmod n$$

  where $\gcf(a, n) = 1$.

- Wilson's theorem states that for a prime p,

$$(p-1)! \equiv -1 \bmod p.$$

  To see this, note that the only elements that are their own inverses are ± 1. All other elements pair off with their inverses and contribute 1 to the product.

- If G has even order, then it has an element of order 2, by the same reasoning as before: some element must be its own inverse by parity.

- This result allows us to classify groups of order $2p$ for prime $p \geq 3$. There must be an element x of order 2. Furthermore, not all elements can have order 2, or else the group would be $(\mathbb{Z}_2)^n$, so there is an element y of order p. Since p is odd, $x \notin \langle y \rangle$, so the group is $G = \langle y \rangle \cup x\langle y \rangle$.

  The product yx must be one of these elements, and it can't be a power of y, so $yx = xy^j$. Then odd powers of yx all carry a power of x, so yx must have even order. If it has order $2p$, then $G \cong C_{2p}$. Otherwise, it has order 2, so $(yx)^2 = y^{j+1} = 1$, implying $j = p - 1$, so $G \cong D_{2p}$.

- The group D_{2n} can be presented in terms of generators and relations,

$$D_{2n} = \langle r, s : r^n = s^2 = e, sr = r^{-1}s \rangle.$$

  In general, when one is given a group in this form, one simply uses the relations to reduce strings of the generators, called words, as far as possible. The remaining set that cannot be reduced form the group elements.

Example. So far we've classified all groups up to order 7, where order 6 follows from the work above. The groups of order 8 are

$$C_8, \quad C_2 \times C_4, \quad C_2 \times C_2 \times C_2, \quad D_8, \quad Q_8$$

where Q_8 is the quaternion group. The quaternions are numbers of the form

$$q = a + b\mathbf{i} + c\mathbf{j} + d\mathbf{k}, \quad a, b, c, d \in \mathbb{R}$$

obeying the rules

$$\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = \mathbf{ijk} = -1.$$

The group Q_8 is identified with the subset $\{\pm 1, \pm\mathbf{i}, \pm\mathbf{j}, \pm\mathbf{k}\}$.

5.2 Group Homomorphisms

Next, we consider maps between groups.

- A group homomorphism $\phi : G \to H$ is a map so that

$$\phi(g_1 g_2) = \phi(g_1)\phi(g_2)$$

  and an isomorphism is simply a bijective homomorphism. An automorphism of G is an isomorphism from G to G, and form a group $\mathrm{Aut}(G)$ under composition. An endomorphism of G is a homomorphism from G to G. We say a monomorphism is an injective homomorphism and an epimorphism is a surjective homomorphism.

- There are many basic examples of homomorphisms.
  - If $H \subseteq G$, we have inclusion $\iota : H \to G$ with $\iota(h) = h$.
  - The trivial map $\phi(g) = e$.
  - The projections $\pi_1 : G_1 \times G_2 \to G_1$, $(g_1, g_2) \mapsto g_1$, and $\pi_2 : G_1 \times G_2 \to G_2$, $(g_1, g_2) \mapsto g_2$.
  - The sign map $\mathrm{sgn} : S_n \to \{\pm 1\}$ which gives the sign of a permutation.
  - The determinant $\det : GL(n, \mathbb{R}) \to \mathbb{R}^*$, and the trace $\mathrm{tr} : M_n(\mathbb{R}) \to \mathbb{R}$ where the operation on $M_n(\mathbb{R})$ is addition.
  - The map $\log : (0, \infty) \to \mathbb{R}$, which is moreover an isomorphism.
  - The map $\phi : G \to G$ given by $\phi(g) = g^2$, if and only if G is abelian.
  - Conjugation is an automorphism, $\phi_h(g) = hgh^{-1}$.
  - All homomorphisms of $\phi : \mathbb{Z} \to \mathbb{Z}$ are of the form $\phi(x) = nx$, because homomorphisms are completely determined by how they map the generators.

- We say H is a normal subgroup of G, and write $H \trianglelefteq G$ if

$$gH = Hg \text{ for all } g \in G$$

  or equivalently if $g^{-1}hg \in H$ for all $g \in G$, $h \in H$. Since conjugation is akin to a "basis change", a normal subgroup "looks the same from all directions". Normality depends on how H is embedded in G, not just on H itself. A group is simple if it has no proper normal subgroups. In an abelian group, all subgroups are normal.

- For a group homomorphism $\phi : G \to H$, define the kernel and image by

$$\ker \phi = \{g \in G : \phi(g) = e\} \trianglelefteq G, \quad \mathrm{im}\, \phi = \{\phi(g) : g \in G\} \subseteq H.$$

  Note that ϕ is constant on cosets of $\ker \phi$.

- Normal subgroups are unions of conjugacy classes. This can place strong constraints on normal subgroups by counting arguments.

- If $|G/H| = 2$ then $H \trianglelefteq G$. This is because the left and right cosets eH and He must coincide, and hence the other left and right coset also coincide. For example, $A_n \trianglelefteq S_n$ and $SO(n) \trianglelefteq O(n)$.

- We define the center of G as

$$Z(G) = \{g \in G : gh = hg \text{ for all } h \in G\}.$$

  Then $Z(G) \trianglelefteq G$.

Next, we construct quotient groups.

- For $H \trianglelefteq G$, we may define a group operation on G/H by

$$(g_1H)(g_2H) = (g_1g_2)H$$

  and hence make G/H into a quotient group. This rule is consistent because

$$(g_1H)(g_2H) = g_1HHg_2 = g_1Hg_2 = g_1g_2H.$$

  Conversely, the consistency of this rule implies $H \trianglelefteq G$, because

$$(g^{-1}hg)H = (g^{-1}H)(hH)(gH) = (g^{-1}H)(eH)(gH) = (g^{-1}g)H = H$$

  which implies that $g^{-1}hg \in H$.

- The idea of a quotient construction is to 'mod out' by H, leaving a simpler structure, or equivalently identify elements of G by an equivalence relation. In terms of sets, there are no restrictions, but we need $H \trianglelefteq G$ to preserve group structure.

- If $H \trianglelefteq G$, it is the kernel of a homomorphism from G, namely

$$\pi : G \to G/H, \quad \pi(g) = gH.$$

- We give a few examples of quotient groups below.
  - We have $\mathbb{Z}/n\mathbb{Z} \cong \mathbb{Z}_n$ almost by definition.
  - We have $S_n/A_n \cong C_2$.
  - For the rotation generator r of D_{2n}, $D_{2n}/\langle r \rangle \cong C_2$.
  - We have $\mathbb{C}^*/S^1 \cong (0, \infty)$ because we remove the complex phase.
  - Let $AGL(n, \mathbb{R})$ denote the group of affine maps $f(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$ where $A \in GL(n, \mathbb{R})$. If T is the subgroup of translations, $G/T \cong GL(n, \mathbb{R})$.

- The first isomorphism theorem states that for a group homomorphism $\phi : G \to H$,

$$G/\ker\phi \cong \operatorname{im}\phi$$

  via the isomorphism

$$g(\ker\phi) \mapsto \phi(g).$$

  It is straightforward to verify this is indeed an isomorphism. As a corollary,

$$|G| = |\ker\phi| \cdot |\operatorname{im}\phi|.$$

- We give a few examples of this theorem below.

  - For $\det : GL(n,\mathbb{R}) \to \mathbb{R}^*$ we have $GL(n,\mathbb{R})/SL(n,\mathbb{R}) \cong \mathbb{R}^*$.
  - For $\phi : \mathbb{Z} \to \mathbb{Z}$ with $\phi(x) = nx$), we have $\mathbb{Z} \cong n\mathbb{Z}$.
  - For $\phi : \mathbb{Z} \to \mathbb{Z}_n$ given by $\phi(x) = \overline{x}$, we have $\mathbb{Z}/n\mathbb{Z} \cong \mathbb{Z}_n$.

- The first isomorphism theorem can also be used to classify all homomorphisms $\phi : G \to H$. We first determine the normal subgroups of G, as these are the potential kernels. For each normal subgroup N, we count the number $n(N)$ of subgroups in H isomorphic to G/N. Finally, we determine $\mathrm{Aut}(G/N)$. Then the number of homomorphisms is

$$\sum_N n(N) \cdot |\mathrm{Aut}(G/N)|.$$

  This is because all such homomorphisms have the form

$$G \xrightarrow{\pi} G/N \xrightarrow{\iota} I$$

  where π maps $g \mapsto gN$ and ι is an isomorphism from G/N to $I \subseteq H \cong G/N$, or which there are $\mathrm{Aut}(G/N)$ possibilities.

There are also additional isomorphism theorems.

- For a group G, if $H \subseteq G$ and $N \trianglelefteq G$, then $HN = \{hn | h \in H, n \in N\}$ is a subgroup of G. This is because $NH = HN$, and $HNHN = HNNH = HNH = HHN = HN$.

- The second isomorphism theorem states that for $H \subseteq G$ and $N \trianglelefteq G$, then $H \cap N \trianglelefteq H$ and

$$\frac{HN}{N} \cong \frac{H}{H \cap N}.$$

  The first statement follows because both N and H are closed under conjugation by elements of H. As for the second, we consider

$$H \xrightarrow{i} HN \to HN/N$$

  where i is the inclusion map and the second map is a quotient. The composition is surjective with kernel $H \cap N$, so the result follows from the first isomorphism theorem.

- Let $N \trianglelefteq G$ and $K \trianglelefteq G$ with $K \subseteq N$. Then $N/K \trianglelefteq G/K$ and

$$(G/K)/(N/K) \cong G/N.$$

  The first statement follows because

$$(gK)^{-1}(nK)(gK) = g^{-1}KnKgK = g^{-1}ngK \in N/K$$

  since K is normal in G. Now consider the composition of quotient maps

$$G \to G/K \to (G/K)/(N/K).$$

  The composition is surjective with kernel N, giving the result.

- Conversely, let $K \trianglelefteq G$ and let $\overline{G} = G/K$ with $\overline{H} \trianglelefteq \overline{G}$. Then there exists $H \subseteq G$ with $\overline{H} = H/K$, defined by
$$H = \{h \in G | hK \in \overline{H}\}.$$
Note that in this definition, $\overline{H}$ is comprised of cosets. However, if $\overline{H} \trianglelefteq \overline{G}$ then $H \trianglelefteq G$.
- As a corollary, given $K \trianglelefteq G$ there is a one-to-one correspondence $H \mapsto \overline{H} = H/K$ between subgroups of G containing K, and subgroups of G/K, which preserves normality. This is a sense in which structure is preserved upon quotienting.

Example. We will use the running example of $G = S_4$. Let $H = S_3 \subseteq S_4$ by acting on the first three elements only, and let $N = V_4 \trianglelefteq S_4$. Then $HN = S_4$ and $H \cap N = \{e\}$, so the second isomorphism theorem states
$$S_4/V_4 \cong S_3.$$
Next, let $N = A_4 \trianglelefteq S_4$ and $K = V_4 \trianglelefteq S_4$. We may compute $G/K \cong S_3$ and $N/K \cong A_3$, so the third isomorphism theorem states
$$S_3/A_3 \cong C_2.$$

Example. The symmetric groups S_n are not simple, because $A_n \trianglelefteq S_n$. However, A_n is simple for $n \geq 5$. For example, for A_5 the conjugacy classes have sizes
$$60 = 1 + 20 + 15 + 12 + 12$$
where the factors of 12 come from splitting the 24 5-cycles. There is no way to pick a subset of these numbers to sum to 30. In fact, A_5 is the smallest non-abelian simple group.

Note. As we'll see below, the simple groups are the "atoms" of group theory. The finite simple groups have been classified; the only possibilities are:

- A cyclic group of prime order C_p.
- An alternating group A_n for $n \geq 5$.
- A finite group of Lie type such as $\mathrm{PSL}(n, q)$ for $n > 2$ or $q > 3$.
- One of 26 sporadic groups, including the Monster and Baby Monster groups.

5.3 Group Actions

Next, we consider group actions.

- A left action of a group G on a set S is a map
$$\rho : G \times S \to S, \quad g \cdot s \equiv \rho(g, s)$$
obeying the axioms
$$e \cdot s = s, \quad g \cdot (h \cdot s) = (gh) \cdot s$$
for all $s \in S$ and $g, h \in G$. A right action would have the order in the second axiom reversed.
- All groups have a left action on themselves by $g \cdot h = gh$ and by conjugation, $g \cdot h = ghg^{-1}$. As we've already seen, there is a left action of G on the left cosets G/H by $g_1 \cdot (g_2 H) = (g_1 g_2)H$, though this only descends to a left action of G/H on itself when $H \trianglelefteq G$.

- The orbit and stabilizer of $s \in S$ are defined as

$$\mathrm{Orb}(s) = \{g \cdot s : g \in G\} \subset S, \quad \mathrm{Stab}(s) = \{g \in G : g \cdot s = s\} \subseteq G.$$

  In particular, $\mathrm{Stab}(s)$ is a subgroup of G, and the orbits partition S. If there is only one orbit, we say the action is transitive. Also, if two elements lie in the same orbit, their stabilizers are conjugate.

- For example, $GL(n, \mathbb{R})$ acts on matrices and column vectors $\mathbb{R}^n$ by matrix multiplication, and on matrices by conjugation; in the latter case the orbits correspond to Jordan normal forms. Also note that $GL(n, \mathbb{R})$ has a left action on column vectors but a right action on row vectors.

- The symmetry group D_{2n} acts on the vertices of a regular n-gon. Affine transformations of the plane act on shapes in the plane, and the orbits are congruence classes. Geometric group actions such as these were the original motivation for group theory.

- The orbit-stabilizer theorem states that

$$|G| = |\mathrm{Stab}(s)| \cdot |\mathrm{Orb}(s)|.$$

  This is because there is an isomorphism between the cosets of $\mathrm{Stab}(s)$ and the elements of $\mathrm{Orb}(s)$, explicitly by $g\,\mathrm{Stab}(s) \mapsto g \cdot s$, which implies $|G|/|\mathrm{Stab}(s)| = |\mathrm{Orb}(s)|$. That is, a transitive group action corresponds to a group action on the set of cosets of the stabilizer.

- This is a generalization of Lagrange's theorem, because in the case $H \subseteq G$, the action of G on G/H by $g \cdot (kH) = (gk)H$ has $\mathrm{Stab}(H) = H$ and $\mathrm{Orb}(H) = G/H$, so $|G| = |G/H| \cdot |H|$. What we've additionally learned is that in the general case, $|\mathrm{Orb}(s)|$ divides $|G|$.

- Define the centralizer of $g \in G$ by

$$C_G(g) = \{h \in G : gh = hg\}.$$

  Also let $C(g)$ be the conjugacy class of g. Applying the orbit-stabilizer theorem to the group action of conjugation,

$$|G| = |C_G(g)| \cdot |C(g)|.$$

  This gives an alternate method for finding $|C(g)|$, or for finding $|G|$.

Example. Let G_T be the tetrahedral group, the set of rotational symmetries of the four vertices of a tetrahedron. The stabilizer of a particular vertex v consists of the identity and two rotations, and the action is transitive, so

$$|G_T| = 3 \cdot 4 = 12.$$

Similarly, for the cube, the stabilizer of a vertex consists of the identity and the 120° and 240° rotations about a space diagonal through the vertex, so

$$|G_C| = 3 \cdot 8 = 24.$$

We could also have done the calculation looking at the orbit and stabilizer of edges or faces.

Example. If $|G| = p^r$, then G has a nontrivial center. The conjugacy class sizes are powers of p, and the class of the identity has size 1, so there must be more classes of size 1, yielding a nontrivial center. In the case $|G| = p^2$, let x be a nontrivial element in the center. If the order of x is p^2, then $G \cong C_{p^2}$. If not, it has order p. Consider another element y with order p, not generated by x. Then the p^2 group elements $x^i y^j$ form the whole group, so $G \cong C_p \times C_p$.

Example. Cauchy's theorem states that for any finite group G and prime p dividing $|G|$, G has an element of order p. To see this, consider the set

$$S = \{(g_1, g_2, \ldots, g_p) \in G^p | g_1 g_2 \ldots g_p = e\}.$$

Then $|S| = |G|^{p-1}$, because the first $p-1$ elements can be chosen freely, while the last element is determined by the others. The group C_p with generator σ acts on S by

$$\sigma \cdot (g_1, g_2, \ldots, g_p) = (g_2, \ldots, g_p, g_1).$$

By the Orbit-Stabilizer theorem, the orbits have size 1 or p, and the orbits partition the set. Since $(e, \ldots, e)$ is an orbit of size 1, there must be other orbits of size 1, corresponding to an element g with $g^p = e$.

Orbits can also be used in counting problems.

- Let G act on S and let N be the number of orbits O_i. Then

  $$N = \frac{1}{|G|} \sum_{g \in G} |\mathrm{fix}(g)|, \quad \mathrm{fix}(g) = \{s \in S : g \cdot s = s\}.$$

  To see this, note that we can count the pairs (g, s) so that $g \cdot s = s$ by summing over group elements or set elements, giving

  $$\sum_{g \in G} |\mathrm{fix}(g)| = \sum_{s \in S} |\operatorname{Stab}(s)|.$$

  Next, applying the Orbit-Stabilizer theorem,

  $$\sum_{s \in S} |\operatorname{Stab}(s)| = \sum_{i=1}^{N} \sum_{s \in O_i} |\operatorname{Stab}(s)| = \sum_{i=1}^{N} \sum_{s \in O_i} \frac{|G|}{|O_i|} = N|G|$$

  as desired. This result is called Burnside's lemma.

- Note that if g and h are conjugate, then $|\mathrm{fix}(g)| = |\mathrm{fix}(h)|$, so the right-hand side can also be evaluated by summing over conjugacy classes.

- Note that every action of G on a set S is associated with a homomorphism

  $$\rho : G \to \operatorname{Sym}(S)$$

  which is called a representation of G. For example, when S is a vector space and G acts by linear transformations, then ρ is a representation as used in physics.

- The representation is faithful if G is isomorphic to $\operatorname{im} \rho$. Equivalently, it is faithful if only the identity element acts trivially.

- A group's action on itself by left multiplication is faithful, so every finite group G is isomorphic to a subgroup of $S_{|G|}$. This is called Cayley's theorem.

Example. Find the number of ways to color a triangle's edges with n colors, up to rotation and reflection. We consider rotations D_6 acting on the triangle, and want to find the number of orbits. Burnside's lemma gives

$$N = \frac{1}{3}\left(n^3 + 3n^2 + 2n\right)$$

where we summed over the trivial conjugacy class, the conjugacy class of the rotation, and the conjugacy class of the reflection. This is indeed the correct answer, with no casework required.

Example. Find the number of ways to paint the faces of a rectangular box black or white, where the three side lengths are distinct. The rotational symmetries are $C_2 \times C_2$, corresponding to the identity and 180° rotations about the x, y, and z axes. Then

$$N = \frac{1}{4}(2^6 + 2^4) = 28.$$

Example. Find the number of ways to make a bracelet with 3 red beads, 2 blue beads, and 2 white beads. Here the symmetry group is D_{14}, imagining the beads as occupying the vertices of a regular heptagon, and there are $7!/3!2!2! = 210$ bracelets without accounting for the symmetry. Then

$$N = \frac{1}{14}(210 + 6(0) + 7(3!)) = 18.$$

Example. Find the number of ways to color the faces of a cube with n colors. The relevant symmetry group is G_C. Note that we have a homomorphism $\rho : G_C \to S^4$ by considering how G_C acts on the four space diagonals of the cube. In fact, it is straightforward to check that ρ is an isomorphism, so $G_C \cong S^4$. This makes it easy to count the conjugacy classes. We have

$$24 = 1 + 3 + 6 + 6 + 8$$

where the 3 corresponds to double transpositions or rotations of π about opposing faces' midpoints, the first 6 corresponds to 4-cycles or rotations of $\pi/2$ about opposing faces' midpoints, the second 6 corresponds to transpositions or rotations of π about opposing edges' midpoints, and the 8 corresponds to 3-cycles or rotations of $\pi/3$ about space diagonals. By Burnside's lemma,

$$N = \frac{1}{24}(n^6 + 3n^4 + 6n^3 + 6n^3 + 8n^2).$$

By similar reasoning, we have a homomorphism $\rho : G_T \to S_4$ by considering how G_T acts on the four vertices of the tetrahedron, and $|G_T| = 12$, so $G_T \cong A_4$.

5.4 Composition Series

First, we look more carefully at generators and relations.

- For a group G and a subset S of G, we defined the subgroup $\langle S \rangle \subseteq G$ to be the smallest subgroup of G containing S. However, it is not clear how this definition works for infinite groups, nor immediately clear why it is unique. A better definition is to let $\langle S \rangle$ be the intersection of all subgroups of G that contain S.

- We say a group G is finitely generated if there exists a finite subset S of G so that $\langle S \rangle = G$. All groups of uncountable order are not finitely generated. Also, $\mathbb{Q}$ under multiplication is countable but not finitely generated because there are infinitely many primes.
- Suppose we have a set S called an alphabet, and define a corresponding set S^{-1}, so the element $x \in S$ corresponds to $x^{-1} \in S^{-1}$. A word w is a finite sequence $w = x_1 \dots x_n$ where each $x_i \in S \cup S^{-1}$. The empty sequence is denoted by $\varnothing$.
- We may contract words by canceling adjacent pairs of the form xx^{-1} for $x \in S \cup x^{-1}$. It is somewhat fiddly to prove, but intuitively clear, that every word w can be uniquely transformed into a reduced word $[w]$ which does not admit any such contractions.
- The set of reduced words is a group under concatenation, called the free group $F(S)$ generated by S. Here $F(S)$ is indeed a group because

  $$[[ww']w''] = [w[w'w'']]$$

  by the uniqueness of reduced words; both are equal to $[ww'w'']$.

Free groups are useful because we can use them to formalize group presentations.

- Given any set S, group G, and mapping $f : S \to G$, there is a unique homomorphism $\phi : F(S) \to G$ so that the diagram

  $$\begin{array}{ccc} S & \xrightarrow{f} & G \\ \downarrow{\scriptstyle i} & \nearrow{\scriptstyle \phi} & \\ F(S) & & \end{array}$$

  commutes, where $i : S \to F(S)$ is the canonical inclusion which takes $x \in S$ to the corresponding generator of $F(S)$.
- To see this, we define

  $$\phi(x_1^{\epsilon_1} \dots x_n^{\epsilon_n}) = f(x_1)^{\epsilon_1} \dots f(x_2)^{\epsilon_2}$$

  where $\epsilon_i = \pm 1$. It is clear this is a homomorphism, and it is unique because $\phi(x) = f(x)$ for every $x \in S$, and a homomorphism is determined by its action on the generators.
- Taking S to be a generating set for G, and f to be inclusion, this implies every group is a quotient of a free group.
- Let B be a subset of a group G. The normal subgroup generated by B is the intersection of all normal subgroups of G that contain B, and is denoted by $\langle\langle B \rangle\rangle$.
- More precisely, we have

  $$\langle\langle B \rangle\rangle = \langle gbg^{-1} : g \in G, b \in B \rangle$$

  which explicitly means that $\langle\langle B \rangle\rangle$ consists of elements of the form

  $$\prod_{i=1}^{n} g_i b_i^{\epsilon_i} g_i^{-1}.$$

To prove this, denote this set as N. It is clear that $N \subseteq \langle\langle B \rangle\rangle$, so it suffices to show that $N \trianglelefteq G$. The only nontrivial check is closure under conjugation, which works because

$$g\left(\prod_{i=1}^{n} g_i b_i^{\epsilon_i} g_i^{-1}\right) g^{-1} = \prod_{i=1}^{n} (gg_i) b_i^{\epsilon_i} (gg_i)^{-1}$$

which lies in N.

- Let X be a set and let R be a subset of $F(X)$. We define the group with presentation $\langle X|R\rangle$ to be $F(X)/\langle\langle R \rangle\rangle$. We need to use $\langle\langle R \rangle\rangle$ because the relation $w = e$ implies $gwg^{-1} = e$.
- For any group G, there is a canonical homomorphism $F(G) \to G$ by sending every generator of $F(G)$ to the corresponding group element. Letting $R(G)$ be the kernel, we have $G \cong F(G)/R(G)$, and hence we define the canonical presentation for G to be

  $$\langle G|R(G)\rangle.$$

  This is a very inefficient presentation, which we mention because it uses no arbitrary choices.
- Free groups also characterize homomorphisms. Let $\langle X|R\rangle$ and H be groups. A map $f : X \to R$ induces a homomorphism $\phi : F(X) \to H$. This descends to a homomorphism $\langle X|R\rangle \to H$ if and only if $R \subset \ker \phi$.

Next, we turn to composition series.

- A composition series for a group G is a sequence of subgroups

  $$\{e\} \trianglelefteq G_1 \trianglelefteq \ldots \trianglelefteq G_{n-1} \trianglelefteq G_n = G$$

  so that each composition factor G_{i+1}/G_i is simple, or equivalently each G_i is a maximal proper normal subgroup of G_{i+1}. By induction, every finite group has a composition series.
- Composition series are not unique. For example, we have

  $$\{e\} \trianglelefteq C_2 \trianglelefteq C_4 \trianglelefteq C_{12}, \quad \{e\} \trianglelefteq C_3 \trianglelefteq C_6 \trianglelefteq C_{12}, \quad \{e\} \trianglelefteq C_2 \trianglelefteq C_6 \trianglelefteq C_{12}.$$

  The composition factors are C_2, C_2, and C_3 in each case, but in a different order.
- Composition series do not determine the group. For example, A_4 has composition series

  $$\{e\} \trianglelefteq C_2 \trianglelefteq V_4 \trianglelefteq A_4$$

  with composition factors C_2, C_2, and C_3. There are actually three distinct composition series here, since V_4 has three C_2 subgroups. The composition factors don't say how they fit together.
- The group $\mathbb{Z}$, which is infinite, does not have a composition series.
- The Jordan-Holder theorem states that all composition series for a finite group G have the same length, with the same composition factors. Consider the two composition series

  $$\{e\} \trianglelefteq G_1 \trianglelefteq \ldots \trianglelefteq G_{r-1} \trianglelefteq G_r = G, \quad \{e\} \trianglelefteq H_1 \trianglelefteq \ldots \trianglelefteq H_{s-1} \trianglelefteq H_s = G.$$

  We prove the theorem by induction on r. If $G_{r-1} = H_{s-1}$, then we are done. Otherwise, note that $G_{r-1}H_{s-1} \trianglelefteq G$. Now, by the definition of a composition series G_{r-1} cannot contain H_{s-1}, so $G_{r-1}H_{s-1}$ must be strictly larger than G_{r-1}. But by the definition of a composition series again, that means we must have $G_{r-1}H_{s-1} = G$. Let $K = G_{r-1} \cap H_{s-1} \trianglelefteq G$.

- The next step in the proof is to 'quotient out' by K. By the second isomorphism theorem,

$$G/G_{r-1} \cong H_{s-1}/K, \quad G/H_{s-1} \cong G_{r-1}/K$$

so G_{r-1}/K and H_{s-1}/K are simple. Since K has a composition series, we have composition series

$$\{e\} \trianglelefteq K_1 \trianglelefteq \ldots \trianglelefteq K_{t-1} \trianglelefteq K \trianglelefteq G_{r-1}, \quad \{e\} \trianglelefteq K_1 \trianglelefteq \ldots \trianglelefteq K_{t-1} \trianglelefteq K \trianglelefteq H_{s-1}.$$

By induction, the former series is equivalent to

$$\{e\} \trianglelefteq G_1 \trianglelefteq \ldots \trianglelefteq G_{r-1}$$

which means that $t = r - 2$. By induction again, the latter series is equivalent to

$$\{e\} \trianglelefteq H_1 \trianglelefteq \ldots \trianglelefteq H_{s-1}$$

which proves that $r = s$.

- Next, we append the factor G to the end of these series. By the second isomorphism theorem, the composition series

$$\{e\} \trianglelefteq K_1 \trianglelefteq \ldots \trianglelefteq K_{t-1} \trianglelefteq K \trianglelefteq G_{r-1} \trianglelefteq G, \quad \{e\} \trianglelefteq K_1 \trianglelefteq \ldots \trianglelefteq K_{t-1} \trianglelefteq K \trianglelefteq H_{s-1} \trianglelefteq G$$

are equivalent. Then our original two composition series are equivalent, completing the proof.

- Note that if G is finite and abelian, its composition factors are also, and hence must be cyclic of prime order. In particular, for $G = C_n$ this proves the fundamental theorem of arithmetic.

- If $H \trianglelefteq G$ with G finite, then the composition factors of G are the union of those of H and G/H. We showed this as a corollary when discussing the isomorphism theorems. In particular, if X and Y are simple, the only two composition series of $X \times Y$ are

$$\{e\} \trianglelefteq X \trianglelefteq X \times Y, \quad \{e\} \trianglelefteq Y \trianglelefteq X \times Y.$$

- A finite group G is solvable if every composition factor is a cyclic group of prime order, or equivalently, abelian. Burnside's theorem states that all groups of order $p^n q^m$ for primes p and q are solvable, while the Feit-Thompson theorem states that all groups of odd order are solvable.

5.5 Semidirect Products

Finally, as a kind of converse, we see how groups can be built up by combining groups.

- We already know how to combine groups using the direct product, but this is uninteresting. Suppose a group were of the form $G = G_1 G_2$ for two disjoint subgroups G_1 and G_2. Then every group element can be written in the form $g_1 g_2$, but it is unclear how we would write the product of two elements $(g_1 g_2)(g_1' g_2')$ in this form. The problem is resolved if one of the G_i is normal in G, motivating the following definition.

- Let G be a group with $H \subseteq G$ and $N \trianglelefteq G$. We say G is an internal semi-direct product of H and N and write
$$G = N \rtimes H$$
if $G = NH$ and $H \cap N = \{e\}$.

- The semidirect product generalizes the direct product. If we also have $H \trianglelefteq G$, then $G \cong N \times H$. To see this, note that every group element can be written uniquely in the form nh. Letting $nh = (n_1h_1)(n_2h_2)$, we have
$$nh = (n_1h_1n_2h_1^{-1})(h_1h_2) = (n_1n_2)(n_2^{-1}h_1n_2h_2).$$
By normality of N and H, both these expressions are already in the form nh. Then we have $n = n_1h_1n_2h_1^{-1} = n_1n_2$, which implies $h_1n_2 = n_2h_1$, giving the result.

- We've already seen several examples of the semidirect product.
  - We have $D_{2n} = \langle\sigma\rangle \rtimes \langle\tau\rangle$ where σ generates rotations and τ is a reflection. Note that a nonabelian group arises from the semidirect product of abelian groups.
  - We have $S_n = A_n \rtimes \langle\sigma\rangle$ for any transposition σ.
  - We have $S_4 = V_4 \rtimes S_3$, which we found earlier.

- To understand the multiplication rule in a semidirect product, letting $nh = (n_1h_1)(n_2h_2)$ again,
$$nh = n_1h_1n_2h_2 = (n_1h_1n_2h_1^{-1})h_1h_2$$
which implies that
$$(n_1, h_1) \circ (n_2, h_2) = (n_1\phi_{h_1}(n_2), h_1h_2), \quad \phi_h(g) = hgh^{-1}.$$
That is, the multiplication law is like that of a direct product, but the multiplication in N is "twisted" by conjugation by H. The map $h \mapsto \phi_h$ gives a group homomorphism $H \to \mathrm{Aut}(N)$.

- This allows us to define the semidirect product of two groups without referring to a larger group, i.e. an external semidirect product. Specifically, for two groups H and N and a homomorphism
$$\phi : H \to \mathrm{Aut}(N)$$
we may define $(N \rtimes H, \circ)$ to consist of the set of pairs (n, h) with group operation
$$(n_1, h_1) \circ (n_2, h_2) = (n_1\phi(h_1)(n_2), h_1h_2).$$
Then it is straightforward to check that $N \trianglelefteq H$ is an internal semi-direct product of the subgroups $\tilde{H} = \{(e, h)\}$ and $\tilde{N} = \{(n, e)\}$. The direct product is just the case of trivial ϕ.

Example. Let $C_n = \langle a\rangle$ and $C_2 = \langle b\rangle$. Let $\phi : C_2 \to \mathrm{Aut}(C_n)$ satisfy $\phi(b)(a) = a^{-1}$. Then $C_n \rtimes_\phi C_2 \cong D_{2n}$. To see this, note that $a^n = b^2 = e$ and
$$ba = (e, b) \circ (a, e) = (\phi(b)(a), b) = a^{-1}b$$
which is the other relation of D_{2n}.

Example. An automorphism of $\mathbb{Z}_n$ must map 1 to another generator, so

$$\mathrm{Aut}(\mathbb{Z}_n) \cong U(\mathbb{Z}_n)$$

where $U(\mathbb{Z}_n)$ is the group of units of the ring $\mathbb{Z}_n$, i.e. the numbers k with $\gcf(k, n) = 1$. For example, suppose we classify semidirect products $\mathbb{Z}_3 \rtimes \mathbb{Z}_3$. Then

$$\mathrm{Aut}(\mathbb{Z}_3) \cong \{1, 2\} \cong \mathbb{Z}_2$$

since the automorphism that maps $1 \mapsto 2$ is negation. However, since the only homomorphism $H : \mathbb{Z}_3 \to \mathbb{Z}_2$ is the trivial map, the only possible semidirect product is $\mathbb{Z}_3 \times \mathbb{Z}_3$.

Next consider $\mathbb{Z}_3 \rtimes \mathbb{Z}_4$. There is one nontrivial homomorphism $H : \mathbb{Z}_4 \to \mathbb{Z}_2$, which maps $1 \bmod 4$ to negation. Hence

$$(n_1 \bmod 3, h_1 \bmod 4) \circ (n_2 \bmod 3, h_2 \bmod 4) = (n_1 + (-1)^{h_1} n_2 \bmod 3, h_1 + h_2 \bmod 4).$$

This is easier to understand in terms of generators. Defining

$$x = (1 \bmod 3, 0 \bmod 4), \quad y = (0 \bmod 3, 1 \bmod 4)$$

we have relations $x^3 = y^4 = e$ and $yx = x^{-1}y$. This is a group of order 12 we haven't seen before.

Example. We know that $S_4 = V_4 \rtimes S_3$. To see this as an external direct product, note that

$$\mathrm{Aut}(V_4) \cong S_3 = \mathrm{Sym}(\{1, 2, 3\})$$

since the three non-identity elements a, b, and c can be permuted. Writing the other factor of S_3 as $\mathrm{Sym}(\{a, b, c\})$, the required homomorphism is the one induced by mapping $a \leftrightarrow 1$, $b \leftrightarrow 2$, $c \leftrightarrow 3$.

We now discuss the group extension problem.

- Let A, B, and G be groups. Then

  $$\{e\} \to A \xrightarrow{i} G \xrightarrow{\pi} B \to \{e\}$$

  is a short exact sequence if i is injective, π is surjective, and $\mathrm{im}\, i = \ker \pi$. Note that $i(A) = \ker \pi \trianglelefteq G$ and by the first isomorphism theorem, $B \cong G/A$.

- In general, we say that an extension of A by B is a group G with a normal subgroup $K \cong A$, with $G/K \cong B$. This is equivalent to the exactness of the above sequence. Hence the classification of extensions of A by B is equivalent to classifying groups G where we know $G/A \cong B$.

- The short exact sequence shown above splits if there is a group homomorphism $j : B \to G$ so that $\pi \circ j = \mathrm{id}_B$, and this occurs if and only if $G \cong A \rtimes B$. For the forward direction, note that if the sequence splits, then j is injective and $\mathrm{im}\, J \cong B$. Since $\mathrm{im}\, i \cap \mathrm{im}\, j = \{e\}$, $G \cong A \rtimes B$. To show explicitly that G is an external semidirect product, we use

  $$\phi : B \to \mathrm{Aut}(A), \quad \phi(b)(a) = i^{-1}(j(b)i(a)j(b^{-1})).$$

Example. The extensions of $C_2 = \langle a \rangle$ by $C_2 = \langle b \rangle$ are

$$\{e\} \to C_2 \to C_2 \times C_2 \times C_2 \to \{e\}$$

along with the nontrivial extension

$$\{e\} \to C_2 \xrightarrow{i} C_4 = \langle c \rangle \xrightarrow{\pi} C_2 \to \{e\}$$

where $i(a) = c^2$ and $\pi(c) = b$. The short exact sequence does not split. Hence even very simple extensions can fail to be semidirect products.

6 Rings

6.1 Fundamentals

We begin with the basic definitions.

- A ring R is a set with two binary operations $+$ and $\times$, so that R is an abelian group under the operation $+$ with identity element $0 \in R$, and $\times$ is associative and distributes over $+$,

$$(a+b)c = ac + bc, \quad a(b+c) = ab + ac$$

  for all $a, b, c \in \mathbb{R}$. If multiplication is commutative, we say the ring is commutative. Most intuitive rules of arithmetic hold, with the notable exception that multiplication is not invertible.

- A ring R has an identity if there is an element $1 \in R$ where $a1 = 1a = a$, and $1 \neq 0$. If the latter were not true, then everything would collapse down to the zero element. Most rings we study will be commutative rings with an identity (CRIs).

- Here we give some fundamental examples of rings.
  - Any field $\mathbb{F}$ is a CRI. The polynomials $\mathbb{F}[x]$ also form a CRI. More generally given any ring R, the polynomials $R[x]$ also form a ring. We may also define polynomial rings with several variables, $R[x_1, \ldots, x_n]$.
  - The integers $\mathbb{Z}$, the Gaussian integers $\mathbb{Z}[i]$, and $\mathbb{Z}_n$ are CRIs. The quaternions $\mathbb{H}$ form a noncommutative ring.
  - The set $M_n(\mathbb{F})$ of $n \times n$ matrices over $\mathbb{F}$ is a ring, which implies $\mathrm{End}(V) = \mathrm{Hom}(V, V)$ is a ring for a vector space V.
  - For an $n \times n$ matrix A, the set of polynomials evaluated on A, denoted $\mathbb{F}[A]$, is a commutative subring of $M_n(\mathbb{F})$. Note that the matrix A may satisfy nontrivial relations; for instance if $A^2 = -I$, then $\mathbb{R}[A] \cong \mathbb{C}$.
  - The space of bounded real sequences ℓ^∞ is a CRI under componentwise addition and multiplication, as does the set of continuous functions $C(\mathbb{R})$. In general for a set S and ring R we may form a ring R^S out of functions $f : S \to R$.
  - The power set $\mathcal{P}(X)$ of a set X is a CRI where the multiplication operation is intersection, and the addition operation is XOR, written as $A \Delta B = (A \setminus B) \cup (B \setminus A)$. Then the additive inverse of each subset is itself. For a finite set, $\mathcal{P}(X) \cong (\mathbb{Z}_2)^{|X|}$.

- Polynomial rings over fields are familiar. However, we will be interested in polynomial rings over rings, which are more subtle. For example, in $\mathbb{Z}_8[x]$ we have

$$(2x)(4x) = 8x^2 = 0$$

  so multiplication is not invertible. Moreover the quadratic $x^2 - 1$ has four roots $1, 3, 5, 7$, and hence can be factored in two ways,

$$x^2 - 1 = (x-1)(x+1) = (x-3)(x-5).$$

  Much of our effort will be directed at finding when properties of $\mathbb{C}[x]$ carry over to general polynomial rings.

- A subring $S \subseteq R$ is a subset of a ring R that is closed under $+$ and $\times$. This implies $0 \in S$. For example, as in group theory, we always have the trivial subrings $\{0\}$ and R. Given any subset $X \subset R$, the subring generated by X is the smaller subring containing it.
- In a ring R, we say a nonzero element $a \in R$ is a zero divisor if there exist nonzero $b, c \in R$ so that $ab = ca = 0$. An integral domain R is a CRI with no zero divisors.
- If R is an integral domain, then cancellation works: if $a \neq 0$ and $ab = ac$, then $b = c$. This is because $0 = ab - ac = a(b - c)$, which implies $b - c = 0$.
- In a ring R with identity, an element $a \in R$ is a unit if there exists a $b \in R$ so that $ab = ba = 1$. If such a b exists, we write it as a^{-1}. The set of units R^* forms a group under multiplication.
- We now give a few examples of these definitions.
  - All fields are integral domains where every element is a unit.
  - The integers $\mathbb{Z}$ form an integral domain with units ± 1. The Gaussian integers $\mathbb{Z}[i]$ form an integral domain with units $\pm 1, \pm i$.
  - In $\mathbb{H}$ there no zero divisors but it is not an integral domain, because it is not commutative.
  - In $M_n(\mathbb{R})$, the nonzero singular matrices are zero divisors, and the invertible matrices are the units.
  - In $\mathcal{P}(X)$, every nonempty proper set is a zero divisor and the only unit is X.

6.2 Quotient Rings and Field Extensions

6.3 Factorization

6.4 Modules

6.5 The Structure Theorem

7 Point-Set Topology

7.1 Definitions

We begin with the fundamentals, skipping content covered when we considered metric spaces.

Definition. A topological space is a set X and a topology $\mathcal{T}$ of subsets of X, whose elements are called the open sets of X. The topology must include $\varnothing$ and X and be closed under finite intersections and arbitrary unions.

Example. The topology containing all subsets of X is called the discrete topology, and the one containing only X and $\varnothing$ is called the indiscrete/trivial topology.

Example. The finite complement topology $\mathcal{T}_f$ is the set of subsets U of X such that $X - U$ is either finite or all of X. The set of finite subsets U of X (plus X itself) fails to be a topology, since it's instead closed under arbitrary intersections and finite unions; taking the complement flips this.

Definition. Let $\mathcal{T}$ and $\mathcal{T}'$ be two topologies on X. If $\mathcal{T}' \supset \mathcal{T}$, then $\mathcal{T}'$ is finer than $\mathcal{T}$. If the reverse is true, we say $\mathcal{T}'$ is coarser than $\mathcal{T}$. If either is true, we say $\mathcal{T}$ and $\mathcal{T}'$ are comparable.

Definition. A basis $\mathcal{B}$ for a topology on X is a set of subsets of X, called basis elements, such that

- For every $x \in X$, there is at least one basis element B containing x.
- If x belongs to the intersection of two basis elements B_1 and B_2, then there is a basis element B_3 containing x such that $B_3 \subset B_1 \cap B_2$.

The topology $\mathcal{T}$ generated by $\mathcal{B}$ is the set of unions of elements of $\mathcal{B}$. Conversely, $\mathcal{B}$ is a basis for $\mathcal{T}$ if every element of $\mathcal{T}$ can be written as a union of elements of $\mathcal{B}$.

Prop. The set of subsets generated by a basis $\mathcal{B}$ is a topology.

Proof. Most properties hold automatically, except for closure under finite intersections. It suffices to consider the intersection of two sets, $U_1, U_2 \in \mathcal{T}$. Let $x \in U_1 \cap U_2$. We know there is a basis element $B_1 \subset U_1$ that contains x, and a basis element $B_2 \subset U_2$ that contains x. Then there is a B_3 containing x contained in $B_1 \cap B_2$, which is in $U_1 \cap U_2$. Then $U_1 \cap U_2 \in \mathcal{T}$, as desired.

Describing a topological space by a basis fits better with our intuitions. For example, the topology generated by $\mathcal{B}'$ is finer than the topology generated by $\mathcal{B}$ is every element of $\mathcal{B}$ can be written as the union of elements of $\mathcal{B}'$. Intuitively, we "smash rocks (basis elements) into pebbles".

Example. The collection of one-point subsets is a basis for the discrete topology. The collection of (open) circles is a basis for the "usual" topology of $\mathbb{R}^2$, as is the collection of open rectangles. We'll formally show this later.

Example. Topologies on $\mathbb{R}$. The standard topology on $\mathbb{R}$ has basis (a, b) for all real $a < b$, and we'll implicitly mean this topology whenever we write $\mathbb{R}$. The lower limit topology on $\mathbb{R}$, written $\mathbb{R}_l$, is generated by basis $[a, b)$. The K-topology on $\mathbb{R}$, written $\mathbb{R}_K$, is generated by open intervals (a, b) and sets $(a, b) - K$, where $K = \{1/n \mid n \in \mathbb{Z}_+\}$.

Both of these topologies are strictly finer than $\mathbb{R}$. For $x \in (a, b)$, we have $x \in [x, b) \subset (a, b)$, so $\mathbb{R}_l$ is finer; since there is no open interval containing x in $[x, b)$, it is strictly finer. Similarly, there is no open interval containing 0 in $(-1, 1) - K$, so $\mathbb{R}_K$ is strictly finer.

Definition. A subbasis $\mathcal{S}$ for a topology on X is a set of subsets of X whose union is $\mathcal{S}$. The topology it generates is the set of unions and finite intersections of elements of $\mathcal{S}$.

Definition. Let X be an ordered set with more than one element. The order topology on X is generated by a basis $\mathcal{B}$ containing all open intervals (a, b), and the intervals $[a_0, b)$ and $(a, b_0]$ where a_0 and b_0 are the smallest and largest elements of X, if they exist.

It's easy to check $\mathcal{B}$ is a basis, as the intersection of two intervals is either empty or another interval.

Prop. The order topology on X contains the open rays

$$(a, +\infty) = \{x \mid x > a\}, \quad (-\infty, a) = \{x \mid x < a\}.$$

Proof. Consider $(a, +\infty)$. If X has a largest element, we're done. Otherwise, it is the union of all basis elements of the form (a, x) for $x > a$.

Example. The order topology on $\mathbb{R}$ is just the usual topology. The order topology on $\mathbb{R}^2$ in the dictionary order contains all open intervals of the form $(a \times b, c \times d)$ where $a < c$ or $a = c$ and $b < d$. It's sufficient to take the intervals of the second type as a basis, since we can recover intervals of the first type by taking unions of rays.

Example. The set $X = \{1, 2\} \times \mathbb{Z}_+$ in the dictionary order looks like $a_1, a_2 \ldots; b_1, b_2, \ldots$. However, the order topology on X is not the discrete topology, because it doesn't contain $\{b_1\}$! All open sets containing b_1 must contain some a_i.

Definition. If X and Y are topological spaces, the product topology on $X \times Y$ is generated by the basis $\mathcal{B}$ containing all sets of the form $U \times V$, where U and V are open in X and Y.

We can't use $\mathcal{B}$ itself as the topology, since the union of product sets is generally not a product set.

Prop. If $\mathcal{B}$ and $\mathcal{C}$ are bases for X and Y, the set of products $\mathcal{D} = \{B \times C \mid B \in \mathcal{B}, C \in \mathcal{C}\}$ is a basis for the product topology on $X \times Y$.

Proof. We must show that any $U \times V$ can be written as the union of members of $\mathcal{D}$. For any $x \times y \in U \times V$, we have basis elements $B \subset U$ containing x and $C \subset V$ containing y. Then $B \times C \subset U \times V$ and contains x, as desired.

Example. The standard topology on $\mathbb{R}^2$ is the product topology on $\mathbb{R} \times \mathbb{R}$.

We can also find a subbasis for the product topology. Let $\pi_1 : X \times Y \to X$ denote projection onto the first factor and let $\pi_2 : X \times Y \to Y$ be projection onto the second factor. If U is open in X, then $\pi_1^{-1}(U) = U \times Y$ is open in $X \times Y$.

Prop. The collection

$$\mathcal{S} = \{\pi_1^{-1}(U) \mid U \text{ open in } X\} \cup \{\pi_2^{-1}(V) \mid V \text{ open in } Y\}$$

is a subbasis for the product topology on $X \times Y$. Intuitively, the basis contains rectangles, and the subbasis contains strips.

Proof. Since every element of $\mathcal{S}$ is open in the product topology, we don't get any extra open sets. We know we get every open set because intersecting two strips gives a rectangle, so we can get every basis element.

Definition. Let X be a topological space with topology $\mathcal{T}$ and let $Y \subset X$. Then

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

is the subspace topology on Y. Under this topology, Y is called a subspace of X.

We show $\mathcal{T}_Y$ is a topology using the distributive properties of $\cap$ and $\cup$. We have to be careful about phrasing; if $U \subset Y$, we say U is open relative to Y if $U \in \mathcal{T}_Y$ and U is open relative to X if $U \in \mathcal{T}$.

Lemma. If $Y \subset X$ and $\mathcal{B}$ is a (sub)basis for $\mathcal{T}$ on X, $\mathcal{B}_Y = \{B \cap Y \,|\, B \in \mathcal{B}\}$ is a (sub)basis for $\mathcal{T}_Y$.

Lemma. Let Y be a subspace of X. If U is open in Y and Y is open in X, then U is open in X.

Prop. If A is a subspace of X and B is a subspace of Y, then the product topology on $A \times B$ is the same as the topology $A \times B$ inherits as a subspace of $X \times Y$. (Product and subspace commute.)

Proof. We show their bases are equal. Every basis element of the topology $X \times Y$ is of the form $U \times V$ for U open in X and V open in Y. Then the basis elements for the subspace topology $A \times B$ of the form

$$(U \times V) \cap (A \times B) = (U \cap A) \times (V \cap B).$$

But the basis elements of X are of the form $U \cap A$ by our lemma, so this is just the basis for the product topology $A \times B$. Thus the topologies are the same.

The same result doesn't hold for the order topology. If X has the order topology and Y is a subset of X, the subspace topology on Y is not the same as the order topology it inherits from X.

Example. Let Y be the subset $[0, 1]$ of $\mathbb{R}$ in the subspace topology. Then the basis has elements of the form (a, b) for $a, b \in Y$, but also elements of the form $[0, b)$ and $(a, 1]$, which are not open in $\mathbb{R}$. This illustrates our above lemma. However, the order topology on Y does coincide with its subspace topology.

Now let Y be the subset $[0, 1) \cup \{2\}$ of $\mathbb{R}$. Then $\{2\}$ is an open set in the subspace topology, but it isn't open in the order topology. (But it would be if Y were the subset $[0, 1] \cup \{2\}$.)

Example. Let $I = [0, 1]$. The set $I \times I$ in the dictionary order topology is called the ordered square, denoted I_o^2. However, it is not the same as the subspace topology on $I \times I$ (as a subspace of the dictionary order topology on $\mathbb{R} \times \mathbb{R}$), since in the latter, $\{1/2\} \times (1/2, 1]$ is open.

In both examples above, the subspace topology looks strange because the intersection operation chops up open sets into closed ones. We will show that if this never happens, the topologies coincide.

Prop. Let a subset Y of X be convex in X if, for every pair of points $a < b$ in Y, all points in the interval (x, y) of X are in Y. If Y is convex in an ordered set X, the order topology and subspace topology on Y coincide.

Proof. We will show they contain each others' subbases. We know Y_{ord} has a subbasis of rays in Y, and Y_{sub} has a subbasis consisting of the intersection of Y with rays in X.

Consider the intersection of ray $(a, +\infty)$ in X with Y. If $a \in Y$, we get a ray in Y. If $a \notin Y$, then by convexity, a is either a lower or upper bound on Y, in which case we get all of Y or nothing. Thus Y_{ord} contains Y_{sub}.

Now consider a ray in Y, $(a, +\infty)$. This is just the intersection of Y with the ray $(a, +\infty)$ in X, so Y_{sub} contains Y_{ord}, giving the result.

In the future, we'll assume that a subset Y of X is given the subspace topology, regardless of the topology on X.

7.2 Closed Sets and Limit Points

Prop. Let Y be a subspace of X. If A is closed in Y and Y is closed in X, then A is closed in X.

Prop. Let Y be a subspace of X and let $A \subset Y$. Then the closure of A in Y is $\overline{A} \cap Y$.

Proof. Let B denote the closure of A in Y. Since B is closed in Y, $B = Y \cap U$ where U is closed in X and contains A. Then $\overline{A} \subset U$, so $\overline{A} \cap Y \subset B$. Next, since $\overline{A}$ is closed in X, $\overline{A} \cap Y$ is closed in Y and contains A, so $B \subset \overline{A} \cap Y$. These two inclusions show the result.

Now we give a convenient way to find the closure of a set. Say that a set A intersects a set B if $A \cap B$ is not empty, and say U is a neighborhood of a point x if U is an open point containing x.

Theorem. Let $A \subset X$. Then $x \in \overline{A}$ iff every neighborhood of x intersects A. If X has a basis, the theorem is also true if we only use basis elements as neighborhoods.

Proof. Consider the contrapositive. Suppose x has a neighborhood U that doesn't intersect A. Then $X - U$ is closed, so $\overline{A} \subset X - U$, so $x \notin A$. Conversely, if $x \notin A$, then $X - \overline{A}$ is a neighborhood of x that doesn't intersect A.

Restricting to basis elements works because if U is a neighborhood of x, then by definition, there is a basis element $B \subset U$ that contains x.

Definition. If $A \subset X$, we say $x \in X$ is a limit point of A if it belongs to the closure of $A - \{x\}$.

Equivalently, every neighborhood of x intersects an element of A, besides itself; intuitively, there are points of A "arbitrarily close" to x.

Theorem. Let $A \subset X$ and let A' be the set of limit points of A. Then $\overline{A} = A \cup A'$.

Proof. The limit point criterion is stricter than the closure criterion above, so $A' \subset \overline{A}$, giving $A \cup A' \subset \overline{A}$. To show the reverse, let $x \in \overline{A}$. If $x \in A$, we're done; otherwise, every neighborhood of x intersects an element of A that isn't itself, so $x \in A'$. Then $\overline{A} \subset A \cup A'$.

Corollary. A subset of a topological space is closed iff it contains all its limit points.

Example. If $A \subset \mathbb{R}$ is the interval $(0, 1]$, then $\overline{A} = [0, 1]$, but the closure of A in the subspace $Y = (0, 2)$ is $(0, 1]$. We can also show that $\overline{\mathbb{Q}} = \mathbb{R}$, and $\overline{\mathbb{Z}_+} = \mathbb{Z}_+$. Note that $\mathbb{Z}_+$ has no limit points.

In a general topological space, intuitive statements about closed sets that hold in $\mathbb{R}$ may not hold. For example, let $X = \{a, b\}$ and $\mathcal{T} = \{\{\}, \{a\}, \{a, b\}\}$. Then the one-point set $\{a\}$ isn't closed, since it has b as a limit point!

Similarly, statements about convergence fail. Given a sequence of points $x_i \in X$, we say the sequence converges to $x \in X$ if, for every neighborhood U of x, there is a positive integer N so that $x_n \in U$ for all $n \geq N$. Then the one-point sequence $a, a, \ldots$ converges to both a and b!

The problem is that the points a and b are "too close together", so close that we can't topologically tell them apart. We add a new, mild axiom to prevent this from happening.

Definition. A topological space X is Hausdorff if, for every two distinct points $x_1, x_2 \in X$, there exist disjoint neighborhoods of x_1 and x_2. Then the points are "housed off" from each other.

Prop. Every finite point set in a Hausdorff space is closed.

Proof. It suffices to show this for a one-point set, $\{x_0\}$. If $x \neq x_0$, then x has a neighborhood that doesn't contain x_0. Then it's not in the closure of $\{x_0\}$, by definition.

This condition, called the T_1 axiom, is even weaker than the Hausdorff axiom.

Prop. Let X satisfy the T_1 axiom and let $A \subset X$. Then x is a limit point of A iff every neighborhood of x contains infinitely many points of A.

Proof. Suppose the neighborhood U of x contains finitely many points of $A - \{x\}$, and call this finite set A'. Since A' is closed, $U \cap (X - A')$ is a neighborhood of x disjoint from $X - \{x\}$, so x is not a limit point of A.

If every neighborhood U of x contains infinitely many points of A, then every such neighborhood contains at least one point of $A - \{x\}$, so x is a limit point of A.

Prop. If X is a Hausdorff space, sequences in X have unique limits.

Proof. Let $x_n \to x$ and $y \neq x$. Then x and y have disjoint neighborhoods U and V. Since all but finitely many x_n are in U, the same cannot be true of V, so x_n does not converge to y.

Prop. Every order topology is Hausdorff, and the Hausdorff property is preserved by products and subspaces.

7.3 Continuous Functions

Example. Let $f : \mathbb{R} \to \mathbb{R}$ be continuous. Then given $x_0 \in \mathbb{R}$ and $\epsilon > 0$, $f^{-1}((f(x_0) - \epsilon, f(x_0) + \epsilon))$ is open in $\mathbb{R}$. Since this set contains x_0, it must contain a basis element (a, b) about x_0, so it contains $(x_0 - \delta, x_0 + \delta)$ for some δ. Thus, if f is continuous, $|x - x_0| < \delta$ implies $|f(x) - f(x_0)| < \epsilon$, the standard continuity criterion. The two are equivalent.

Example. Let $f : \mathbb{R} \to \mathbb{R}_l$ be the identity function $f(x) = x$. Then f is not continuous, because the inverse image of the open set $[a, b)$ of $\mathbb{R}_0$ is not open in $\mathbb{R}$.

Definition. Let $f : X \to Y$ be injective and continuous and let $Z = f(X)$, so the restriction $f' : X \to Z$ is bijective. If f' is a homeomorphism, we say f is a topological imbedding of X in Y.

Example. The topological spaces $(-1, 1)$ and $\mathbb{R}$ are isomorphic. Define $F : (-1, 1) \to \mathbb{R}$ and its inverse G as

$$F(x) = \frac{x}{1 - x^2}, \quad G(y) = \frac{2y}{1 + (1 + 4y^2)^{1/2}}.$$

Because F is order-preserving and bijective, it corresponds basis elements of $(-1, 1)$ and $\mathbb{R}$, so it is a homeomorphism. Alternatively, we can show F and G are continuous using facts from calculus.

Example. Define $f : [0, 1) \to S^1$ by $f(t) = (\cos 2\pi t, \sin 2\pi t)$. Then f is bijective and continuous. However, f^{-1} is not, since f sends the open set $[0, 1/4)$ to a non-open set. This makes sense, since our two sets are topologically distinct.

As in real analysis, we now give rules for constructing continuous functions.

Prop. Let X and Y be topological spaces.

- The constant function is continuous.

- Compositions of continuous functions are continuous.
- Let A be a subspace of X. The inclusion function $j : A \to X$ is continuous, and the restriction of a continuous $f : X \to Y$ to A, $f|_A : A \to Y$, is continuous.
- (Range) Let $f : X \to Y$ be continuous. If Z is a subspace of Y containing $f(X)$, the function $g : X \to Z$ obtained by restricting the range of f is continuous. If Z is a space having Y as a subspace, the function $h : X \to Z$ obtained by expanding the range of f is also continuous.
- (Local criterion) The map $f : X \to Y$ is continuous if X can be written as the union of open sets U_α so that $f|_{U_\alpha}$ is continuous for each α.
- (Pasting) Let $X = A \cup B$ where A and B are closed in X. If $f : A \to Y$ and $g : B \to Y$ are continuous and agree on $A \cap B$, then they combine to yield a continuous function $h : X \to Y$.

Proof. Most of these properties are straightforward, so we only prove the last one. Let C be a closed subset of Y. Then $h^{-1}(C) = f^{-1}(C) \cup g^{-1}(C)$. These sets are closed in A and B respectively, and hence closed in X. Then $h^{-1}(C)$ is closed in X.

Example. The pasting lemma also works if A and B are both open, since the local criterion applies. However, it can fail if only A is closed and only B is open. Consider the real line and let $A = (-\infty, 0)$ and let $B = [0, \infty)$, with $f(x) = x - 2$ and $g(x) = x + 2$. These functions are continuous on A and B respectively, but pasting them yields a function discontinuous at $x = 0$.

Prop. Write $f : A \to X \times Y$ as $f(a) = (f_1(a), f_2(a))$. Then f is continuous iff the coordinate functions f_1 and f_2 are. This is another manifestation of the universal property of the product.

Proof. If f is continuous, the composition $f_i = \pi_i \circ f$ is continuous. Conversely, let f_1 and f_2 are continuous. We will show the inverse image of basis elements is open. By set theory, $f^{-1}(U \times V) = f_1^{-1}(U) \cap f_2^{-1}(V)$, which is open since it's the intersection of two open sets.

This theorem is useful in vector calculus; for example, a vector field is continuous iff its components are.

7.4 The Product Topology

We now generalize the product topology to arbitrary Cartesian products.

Definition. Given an index set J and a set X, a J-tuple of elements of X is a function $\mathbf{x} : J \to X$. We also write $\mathbf{x}$ as $(x_\alpha)_{\alpha \in J}$. Denote the set of such J-tuples as X^J.

Definition. Given an indexed family of sets $\{A_\alpha\}_{\alpha \in J}$, let $X = \bigcup_{\alpha \in J} A_\alpha$ and define their Cartesian product $\prod_{\alpha \in J} A_\alpha$ as the subset of X^J where $x_\alpha \in A_\alpha$ for each $\alpha \in J$.

Definition. Let $\{X_\alpha\}_{\alpha \in J}$ be an indexed family of topological spaces, and let U_α denote an arbitrary open set in X_α.

- The box topology on $\prod X_\alpha$ has basis elements of the form $\prod U_\alpha$.
- The product topology on $\prod X_\alpha$ has subbasis elements of the form $\pi_\alpha^{-1}(U_\alpha)$, for arbitrary α.

We've already seen that in the finite case, these two definitions are equivalent. However, they differ in the infinite case, because subbasis elements only generate open sets under *finite* intersections. Then the basis elements of the product topology are of the form $\prod U_\alpha$, where $U_\alpha = X_\alpha$ for all but finitely many values of α. We prefer the product topology, for the following reason.

Prop. Write $f : A \to \prod X_\alpha$ as $f(a) = (f_\alpha(a))_{\alpha \in J}$. If $\prod X_\alpha$ has the product topology, then f is continuous iff the coordinate functions f_α are.

Proof. If f is continuous, the composition $f_i = \pi_i \circ f$ is continuous. Conversely, let the f_α be continuous. We will show the inverse image of subbasis elements is open. The inverse image of $\pi_\beta^{-1}(U_\beta)$ is $f_\beta^{-1}(U_\beta)$, which is open in A by the continuity of f_β.

Example. The above proposition doesn't hold for the box topology. Consider $\mathbb{R}^\omega$ and let $f(t) = (t, t, \ldots)$. Then each coordinate function is continuous, but the inverse image of the basis element

$$B = (-1, 1) \times \left(-\frac{1}{2}, \frac{1}{2}\right) \times \left(-\frac{1}{3}, \frac{1}{3}\right) \times \cdots$$

is not open, because it contains the point zero, but no basis element $(-\delta, \delta)$ about the point zero. This is inherently because open sets are not closed under infinite intersections.

In the future, whenever we consider $\prod X_\alpha$, we will implicitly give it the product topology. The box topology will sometimes be used to construct counterexamples.

Prop. The following results hold for $\prod X_\alpha$ in either the box or product topologies.

- If A_α is a subspace of X_α, then $\prod A_\alpha$ is a subspace of $\prod X_\alpha$ if both are given the box or product topologies.
- If each X_α is Hausdorff, so is $\prod X_\alpha$.
- Let $A_\alpha \subset X_\alpha$. Then
$$\prod \overline{A_\alpha} = \overline{\prod A_\alpha}.$$
- Let X_α have basis $\mathcal{B}_\alpha$. Then $\prod B_\alpha$ where $B_\alpha \in \mathcal{B}_\alpha$ is a basis for the box topology. The same collection of sets, where $B_\alpha = X_\alpha$ for all but a finite number of α, is a basis for the product topology. Thus the box topology is finer than the product topology.

7.5 The Metric Topology

Definition. If X is a metric space with metric d, the collection of all ϵ-balls

$$B_d(x, \epsilon) = \{y \mid d(x, y) < \epsilon\}$$

is a basis for a topology on X, called the metric topology induced by d. We say a topological space is metrizable if it can be induced by a metric on the underlying set, and call a metrizable space together with its metric a metric space.

Metric spaces correspond nicely with our intuitions from analysis. For example, using a basis above, a set U is open if, for every $y \in U$, U contains an ϵ-ball centered at y. Different choices of metric may yield the same topology; properties dependent on such a choice are not topological properties.

Example. The metric $d(x, y) = 1$ (for $x \neq y$) generates the discrete topology.

Example. The metric $d(x, y) = |x - y|$ on $\mathbb{R}$ generates the standard topology on $\mathbb{R}$, because its basis elements $(x - \epsilon, x + \epsilon)$ are the same as those of the order topology, (a, b).

Example. Boundedness is not a topological property. Let X be a metric space with metric d. A subset A of X is bounded if the set of distances $d(a_1, a_2)$ with $a_1, a_2 \in A$ has an upper bound. If A is bounded, its diameter is

$$\operatorname{diam} A = \sup_{a_1, a_2 \in A} d(a_1, a_2).$$

The standard bounded metric on X is defined by

$$\overline{d}(x, y) = \min(d(x, y), 1).$$

Then every set is bounded if we use the metric $\overline{d}$, but d and $\overline{d}$ generate the same topology! Proof: we may use the set of ϵ-balls with $\epsilon < 1$ as a basis for the metric topology. These sets are identical for d and $\overline{d}$.

We now show that $\mathbb{R}^n$ is metrizable.

Definition. Given $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$, we define the Euclidean metric d_2 as

$$d_2(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2, \quad \|x\|_2 = \sqrt{x_1^2 + \ldots + x_n^2}.$$

We may also define other metric with a general exponent; in particular,

$$d_\infty(\mathbf{x}, \mathbf{y}) = \max\{|x_1 - y_1|, \ldots, |x_n - y_n|\}.$$

8 Algebraic Topology

9 Methods for ODEs

9.1 Differential Equations

> *I say, gentleman, hadn't we better kick over the whole show and scatter rationalism to the winds, simply to send these logarithms to the devil, and to enable us to live once more at our own sweet foolish will!*
>
> – Dostoevsky, Notes from Underground (1864)

In this section, we will focus on techniques for solving linear ordinary differential equations (ODEs).

- Our problems will be of the form

$$Ly(x) = f(x), \quad L = P_n\partial^n + \ldots + P_0, \quad a \leq x \leq b$$

  where L is a linear differential operator and f is the forcing function.

- There are several ways we can specify a solution. When the independent variable x represents time, we often use initial conditions, specifying y and its derivatives at $x = a$. When x represents space, we often use boundary conditions, which constrain y and its derivatives at $x = a$ or $x = b$.

- We will consider only linear boundary conditions, i.e. those of the form

$$\sum_n a_n y^{(n)}(x_0) = \gamma, \quad x_0 \in \{a, b\}.$$

  The boundary condition is homogeneous if γ is zero. Boundary value problems are more subtle than initial value problems, because a given set of boundary conditions may admit no solutions or infinitely many. As such, we will completely ignore the boundary conditions for now.

- By the linearity of L, the general solution consists of a solution to the equation plus any solution to the homogeneous equation, which has $f = 0$. The solutions to the homogeneous equation form an n-dimensional vector space. For simplicity we will focus on the case $n = 2$ below.

- The simplest way to check if a set of solutions to the homogeneous equation is linearly dependent is to evaluate the Wronskian. For $n = 2$ it is

$$W(y_1, y_2) = \det \begin{pmatrix} y_1 & y_2 \\ y_1' & y_2' \end{pmatrix} = y_1 y_2' - y_2 y_1'$$

  and the generalization to arbitrary n is straightforward. If the solutions are linearly dependent, then the Wronskian vanishes.

- The converse to the above statement is a bit subtle. It is clearly true if the P_i are all constants. However, if $P_2(x') = 0$ for some x', then y'' is not determined at that point; hence two solutions may be dependent for $x < x'$ but become independent for $x > x'$. If $P_2(x)$ never vanishes, the converse is indeed true.

- For constant coefficients, the homogeneous solutions may be found by guessing exponentials. In the case where $P_n \propto x^n$, all terms have the same power, so we may guess a power x^m.

- Another useful trick is reduction of order. Suppose one solution $y_1(x)$ is known. We guess a solution of the form
$$y(x) = v(x)y_1(x).$$
Plugging this in, all terms proportional to v cancel because y_1 satisfies the ODE, giving
$$P_2(2v'y_1' + v''y_1) + P_1v'y_1 = 0$$
which is a first-order ODE in v'.

Next, we introduce variation of parameters to solve the inhomogeneous equation.

- Given homogeneous solutions $y_1(x)$ and $y_2(x)$, we guess an inhomogeneous solution
$$y(x) = c_1(x)y_1(x) + c_2(x)y_2(x).$$
We impose the condition $c_1'y_1 + c_2'y_2 = 0$, so we have
$$y' = c_1y_1' + c_2y_2', \quad y'' = c_1y_1'' + c_2y_2'' + c_1'y_1' + c_2'y_2'$$
and the condition ensures that no second derivatives of the c_i appear.

- Plugging this into the ODE we find
$$Ly = P_2(c_1'y_1' + c_2'y_2') = f$$
where many terms drop out since y_1 and y_2 are homogeneous solutions.

- We are left with a system of two first-order ODEs for the c_i, which are solvable. By solving the system, we find
$$c_1' = -\frac{fy_2}{P_2W}, \quad c_2' = \frac{fy_1}{P_2W}$$
where W is again the Wronskian. Then the general solution is
$$y(x) = -y_1(x)\int^x \frac{f(t)y_2(t)}{P_2(t)W(t)}\,dt + y_2(x)\int^x \frac{f(t)y_1(t)}{P_2(t)W(t)}\,dt.$$
As before, there are issue if $P_2(t)$ ever vanishes, so we assume it doesn't. The constants of integration from the unspecified lower bounds allow the addition of an arbitrary homogeneous solution.

- So far we haven't accounted for boundary conditions. Consider the simple case $y(a) = y(b) = 0$. We choose homogeneous solutions obeying
$$y_1(a) = y_2(b) = 0.$$
Then the boundary conditions require
$$c_2(a) = c_1(b) = 0$$
which fixes the unique solution
$$y(x) = y_1(x)\int_x^b \frac{f(t)y_2(t)}{P_2(t)W(t)}\,dt + y_2(x)\int_a^x \frac{f(t)y_1(t)}{P_2(t)W(t)}\,dt.$$
We can also write this in terms of a Green's function $g(x,t)$,
$$y(x) = \int_a^b g(x,t)f(t)\,dt, \quad g(x,t) = \frac{1}{P_2(t)W(t)}\begin{cases} y_1(t)y_2(x) & t \le x \\ y_2(t)y_1(x) & x \le t \end{cases}.$$
Similar methods work for any homogeneous boundary conditions.

9.2 Eigenfunction Methods

We begin by reviewing Fourier series.

- Fourier series are defined for functions $f : S^1 \to \mathbb{C}$, parametrized by $\theta \in [-\pi, \pi)$. We define the Fourier coefficients
$$\hat{f}_n = \frac{1}{2\pi}(e^{in\theta}, f) \equiv \frac{1}{2\pi}\int_0^{2\pi} e^{-in\theta} f(\theta)\, d\theta.$$
We then claim that
$$f(\theta) = \sum_{n\in\mathbb{Z}} \hat{f}_n e^{in\theta}.$$
Before continuing, we investigate whether this sum converges to f, if it converges at all.

- One can show that the Fourier series converges to f for continuous functions with bounded continuous derivatives. Fejer's theorem states that one can always recover f from the $\hat{f}_n$ as long as f is continuous except at finitely many points, though it makes no statement about the convergence of the Fourier series. One can also show that the Fourier series converges to f as long as $\sum_n |\hat{f}_n|$ converges.

- The Fourier coefficients for the sawtooth function $f(\theta) = \theta$ are
$$\hat{f}_n = \begin{cases} 0 & n = 0, \\ (-1)^{n+1}/in & \text{otherwise.} \end{cases}$$
At the discontinuity, the Fourier series converges to the average of $f(\pi^+)$ and $f(\pi^-)$. This always happens: to show that, simply add the sawtooth to any function with a discontinuity to remove it, then apply linearity.

- Integration makes Fourier series 'nicer' by dividing $\hat{f}_n$ by in, while differentiation does the opposite. In particular, a discontinuity appears as $1/n$ decay of the Fourier coefficients (as shown for the sawtooth), so a discontinuity of $f^{(k)}$ appears as $1/n^{k+1}$ decay. For a smooth function, the Fourier coefficients fall faster than any power.

- Right next to a discontinuity, the truncated Fourier series displays an overshoot by about 18%, called the Gibbs-Wilbraham phenomenon. The width of the overshoot region goes to zero as more terms are added, but the maximum extent of the overshoot remains the same; this shows that the Fourier series converges pointwise rather than uniformly. (The phenomenon can be shown explicitly for the square wave; this extends to all other discontinuities by linearity.)

- Computing the norm-squared of f in position space and Fourier space gives Parseval's identity,
$$\int_{-\pi}^{\pi} |f(\theta)|^2\, d\theta = 2\pi \sum_{k\in\mathbb{Z}} |\hat{f}_k|^2.$$
This is simply the fact that the map $f(x) \to \hat{f}_n$ is unitary.

- Parseval's theorem also gives error bounds: the mean-squared error from cutting off a Fourier series is proportional to the length of the remaining Fourier coefficients. In particular, the best possible approximation of a function f (in terms of mean-squared error) using only a subset of the Fourier coefficients is obtained by simply truncating the Fourier series.

Fourier series are simply changes of basis in function space, and linear differential operators are linear operators in function space.

- We are interested in solving the eigenfunction problem

  $$Ly_i(x) = \lambda_i y_i(x)$$

  along with homogeneous boundary conditions. Generically, there will be infinitely many eigenfunctions, allowing us to construct a solution to the inhomogeneous problem by linearity.

- We define the inner product on the function space as

  $$(u, v) = \int_a^b u(x)v(x)\,dx.$$

  Note there is no conjugation because we only work with real functions.

- We wish to define the adjoint L^* of a linear operator L by

  $$(Ly, w) = (y, L^*w).$$

  We could then get an explicit expression for L^* using integration by parts. However, generally we end up with boundary terms, which don't have the correct form.

- Suppose that we have certain homogeneous boundary conditions on y. Demanding that the boundary terms vanish will induce homogeneous boundary conditions on w. If $L = L^*$ and the boundary conditions stay the same, the problem is self-adjoint. If only $L = L^*$, then we call L self-adjoint, or Hermitian.

Example. We take $L = \partial^2$ with $y(a) = 0$, $y'(b) - 3y(b) = 0$. Then we have

$$\int_a^b wy''\,dx = (wy' - w'y)\Big|_a^b + \int_a^b yw''\,dx.$$

Hence we have $L^* = \partial^2$, and the induced boundary conditions are

$$w'(b) - 3w(b) = 0, \quad w(a) = 0.$$

Hence the problem is self-adjoint.

Now we focus on the eigenfunctions.

- Eigenfunctions of the adjoint problem have the same eigenvalues as the original problem. That is, if $Ly = \lambda y$, there is a w so that $L^*w = \lambda w$. This is intuitive thinking of L^* as the transpose of L, though we can't formally prove it.

- Eigenfunctions with different eigenvalues are orthogonal. Specifically, let

  $$Ly_j = \lambda_j y_j, \quad Ly_k = \lambda_k y_k$$

  where the latter yields $L^*w_k = \lambda_k w_k$. Then if $\lambda_j \neq \lambda_k$, then $\langle y_j, w_k\rangle = 0$. This follows from the same proof as for matrices.

- To solve a general inhomogeneous boundary value problem, we solve the eigenvalue problem (subject to homogeneous boundary conditions) as well as the adjoint eigenvalue problem, to obtain (λ_j, y_j, w_j). To obtain a solution for $Ly = f(x)$ we assume

$$y = \sum_i c_i y_i(x).$$

  We then solve for the coefficients by projection,

$$\langle f, w_k \rangle = \langle Ly, w_k \rangle = \langle y, \lambda_k w_k \rangle = \lambda_k c_k \langle y_k, w_k \rangle$$

  from which we may find c_k.

- Finally, consider the case of inhomogeneous boundary conditions. Such a problem can always be split into an inhomogeneous problem with homogeneous boundary conditions, and a homogeneous problem with inhomogeneous boundary conditions. Since solving homogeneous problems tends to be easier, this case isn't much harder.

Example. Consider the inhomogeneous problem

$$y'' = f(x), \quad 0 \le x \le 1, \quad y(0) = \alpha, \quad y(1) = \beta.$$

Performing the decomposition described above, the homogeneous boundary conditions are simply $y(0) = y(1) = 0$, so the eigenfunctions are

$$y_k(x) = \sin(k\pi x), \quad \lambda_k = -k^2\pi^2, \quad k = 1, 2, \ldots.$$

The problem is self-adjoint, so $y_k = w_k$ and we have

$$c_k = \frac{\langle f, w_k \rangle}{\lambda_k \langle y_k, w_k \rangle} = -\frac{2\int_0^1 f(x)\sin(k\pi x)\,dx}{k^2\pi^2}.$$

To handle the inhomogeneous boundary conditions, we simply add on $(\beta - \alpha)x + \alpha$.

- For most applications, we're interested in second-order linear differential operators,

$$\mathcal{L} = P(x)\frac{d^2}{dx^2} + R(x)\frac{d}{dx} - Q(x), \quad \mathcal{L}y = 0.$$

- We may simplify $\mathcal{L}$ using the method of integrating factors,

$$\frac{1}{P(x)}\mathcal{L} = \frac{d^2}{dx^2} + \frac{R(x)}{P(x)}\frac{d}{dx} - \frac{Q(x)}{P(x)} = e^{-\int^x R(t)/P(t)\,dt}\frac{d}{dx}\left(e^{\int^x R(t)/P(t)\,dt}\frac{d}{dx}\right) - \frac{Q(x)}{P(x)}.$$

  Assuming $P(x) \neq 0$, the equation $\mathcal{L}y = 0$ is equivalent to $(1/P(x))\mathcal{L}y = 0$. Hence any $\mathcal{L}$ can be taken to have the form

$$\mathcal{L} = \frac{d}{dx}\left(p(x)\frac{d}{dx}\right) - q(x)$$

  without loss of generality. Operators in this form are called Sturm-Liouville operators.

- Sturm-Liouville operators are self-adjoint under the inner product

$$(f, g) = \int_a^b f(x)^* g(x)\, dx$$

provided that the functions on which they act obey appropriate boundary conditions. To see this, apply integration by parts for

$$(\mathcal{L}f, g) - (f, \mathcal{L}g) = \left[p(x) \left(\frac{df^*}{dx} g - f^* \frac{dg}{dx} \right) \right]_a^b.$$

- There are several possible boundary conditions that ensure the boundary term vanishes. For example, we can demand

$$f(a)/f'(a) = c_a, \quad f(b)/f'(b) = c_b$$

for constants c_a and c_b, for all functions f. Alternatively, we can demand periodicity,

$$f(a) = f(b), \quad f'(a) = f'(b).$$

Another possibility is that $p(a) = p(b) = 0$, in which case the term automatically vanishes. Naturally, we always assume the functions are smooth.

Next, we consider the eigenfunctions of the Sturm-Liouville operators.

- A function $y(x)$ is an eigenfunction of $\mathcal{L}$ with eigenvalue λ and weight function $w(x)$ if

$$\mathcal{L}y(x) = \lambda w(x) y(x).$$

The weight function must be real, nonnegative, and have finitely many zeroes on the domain $[a, b]$. It isn't necessary, as we can remove it by redefining y and $\mathcal{L}$, but it will be convenient.

- We define the inner product with weight w to be

$$(f, g)_w = \int_a^b f^*(x) g(x) w(x)\, dx$$

so that $(f, g)_w = (f, wg) = (wf, g)$. The conditions on the weight function are chosen so that the inner product remains nondegenerate, i.e. $(f, f)_w = 0$ implies $f = 0$. We take the weight function to be fixed for each problem.

- By the usual proof, if $\mathcal{L}$ is self-adjoint, then the eigenvalues λ are real. Moreover, since everything is real except for the functions themselves, f^* is an eigenfunction if f is. Thus we can always switch basis to $\operatorname{Re} f$ and $\operatorname{Im} f$, so the eigenfunctions can be chosen real.

- Moreover, eigenfunctions with different eigenvalues are orthogonal, as

$$\lambda_i (f_j, f_i)_w = (f_j, \mathcal{L}f_i) = (\mathcal{L}f_j, f_i) = \lambda_j (f_j, f_i)_w.$$

Thus we can construct an orthonormal set $\{Y_n(x)\}$ from eigenfunctions $y_n(x)$ by setting $Y_n = y_n / \sqrt{(y_n, y_n)}$.

- One can show that the eigenvalues form a countably infinite sequence $\{\lambda_n\}$ with $|\lambda_n| \to \infty$ as $n \to \infty$, and that the eigenfunctions $Y_n(x)$ form a complete set for functions satisfying the given boundary conditions. Thus we may always expand such a function f as

$$f(x) = \sum_{n=1}^{\infty} f_n Y_n(x), \quad f_n = (Y_n, f)_w = \int_a^b Y_n^*(x) f(x) w(x)\, dx.$$

  From now on we ignore convergence issues for infinite sums.

- Parseval's identity carries over, as

$$(f, f)_w = \sum_{n=1}^{\infty} |f_n|^2.$$

Example. We choose periodic boundary conditions on $[-L, L]$ with $\mathcal{L} = d^2/dx^2$ and $w(x) = 1$. Solving the eigenfunction equation

$$y''(x) = \lambda y(x)$$

gives solutions

$$y_n(x) = \exp(in\pi x/L), \quad \lambda_n = -\left(\frac{n\pi}{L}\right)^2, \quad n \in \mathbb{Z}.$$

Thus we've recovered the Fourier series.

Example. Consider the differential equation

$$\frac{1}{2}H'' - xH' = -\lambda H, \quad x \in \mathbb{R}$$

subject to the condition that $H(x)$ grows sufficiently slowly at infinity, to ensure inner products exist. Using the method of integrating factors, we rewrite the equation in Sturm-Liouville form,

$$\frac{d}{dx}\left(e^{-x^2}\frac{dH}{dx}\right) = -2\lambda e^{-x^2} H(x).$$

This is now an eigenfunction equation with weight function $w(x) = e^{-x^2}$. Thus weight functions naturally arise when converting general second-order linear differential operators to Sturm-Liouville form. The solutions are the Hermite polynomials,

$$H_n(x) = (-1)^n e^{x^2} \frac{d^n}{dx^n} e^{-x^2}$$

and they are orthogonal with respect to the weight function $w(x)$.

Example. Consider the inhomogeneous equation

$$\mathcal{L}\phi(x) = w(x)F(x)$$

where $F(x)$ is a forcing term. Expanding in the eigenfunctions yields the particular solution

$$\phi_p(x) = \sum_{n=1}^{\infty} \frac{(Y_n, F)_w}{\lambda_n} Y_n(x).$$

Alternatively, expanding this as an integral and defining $f(x) = w(x)F(x)$, we have

$$\phi_p(x) = \int_a^b G(x, \xi) f(\xi)\, d\xi, \quad G(x, \xi) = \sum_{n=1}^{\infty} \frac{Y_n(x) Y_n^*(\xi)}{\lambda_n}.$$

The function G is called a Green's function, and it provides a formal inverse to $\mathcal{L}$. It gives the response at x to forcing at ξ.

9.3 Distributions

We now take a detour by defining distributions, as the Dirac delta 'function' will be needed later.

- Given a domain Ω, we choose a class of test functions $D(\Omega)$. The test functions are required to be infinitely smooth and have compact support; one example is

$$\psi(x) = \begin{cases} e^{-1/(1-x^2)} & |x| < 1, \\ 0 & \text{otherwise.} \end{cases}$$

  A distribution T is a linear map $T : D(\Omega) \to \mathbb{R}$ given by $T : \phi \mapsto T[\phi]$. The set of distributions is written as $D'(\Omega)$, the dual space of $D(\Omega)$. It is a vector space under the usual operations.

- We can define the product of a distribution and a test function by

$$(\psi T)[\phi] = T[\psi\phi].$$

  However, there is no way to multiply distributions together.

- The simplest type of distribution is an integrable function $f : \Omega \to \mathbb{R}$, where we define the action by the usual inner product of functions,

$$f[\phi] = (f, \phi) = \int_\Omega f(x)\phi(x)\, dV.$$

  However, the most important example is the Dirac delta 'function',

$$\delta[\phi] = \phi(0)$$

  which cannot be thought of this way. Though we often write the Dirac δ-function under integrals, we always implicitly think of it as a functional of test functions.

- The Dirac δ-function can also be defined as the limit of a sequence of distributions, e.g.

$$G_n(x) = ne^{-n^2x^2}/\sqrt{\pi}.$$

  In terms of functions, the limit $\lim_{n\to\infty} G_n(x)$ does not exist. But if we view the functions as distributions, we have $\lim_{n\to\infty}(G_n, \phi) = \phi(0)$ for each ϕ, giving a limiting distribution, the Dirac delta.

- Next, we can define the derivative of a distribution by integration by parts,

$$T'[\phi] = -T[\phi'].$$

  This trick means that distributions are infinitely differentiable, despite being incredibly badly behaved! For example, $\delta'[\phi] = -\phi'(0)$. As another example, the step function $\Theta(x)$ is not differentiable as a function, but as a distribution,

$$\Theta'[\phi] = -\Theta[\phi'] = \phi(0) - \phi(\infty) = \phi(0)$$

  which gives $\Theta' = \delta$.

- The Dirac δ-function obeys
$$\delta(f(x)) = \sum_i \frac{\delta(x - x_i)}{|f'(x_i)|}$$
where the x_i are the roots of f. This can be shown nonrigorously by treating the delta function as an ordinary function and using integration rules; it can also be proven entirely within distribution theory.

- The Fourier series of the Dirac δ-function on $[-L, L]$ is
$$\delta(x) = \frac{1}{2L} \sum_{n \in \mathbb{Z}} e^{in\pi x/L}.$$
Again, the right-hand side must be thought of as a limit of a series of distributions. When integrated against a test function $\phi(x)$, it extracts the sum of the Fourier coefficients $\hat{\phi}_n$, which yields $\phi(0)$.

- Similarly, we can expand the Dirac δ-function in any basis of orthonormal functions,
$$\delta(x - \xi) = \sum_n c_n Y_n(x), \quad c_n = \int_a^b Y_n^*(x)\delta(x - \xi)w(x)\,dx = Y_n^*(\xi)w(\xi).$$
This gives the expansion
$$\delta(x - \xi) = w(\xi) \sum_n Y_n^*(\xi)Y_n(x) = w(x) \sum_n Y_n^*(\xi)Y_n(x)$$
where we can replace $w(\xi)$ with $w(x)$ since $\delta(x-\xi)$ is zero for all $x \neq \xi$. To check this expression, note that if $g(x) = \sum_m d_m Y_m(x)$, then
$$\int_a^b g^*(x)\delta(x - \xi) = \sum_{m,n} Y_n^*(\xi)d_m^* \int_a^b w(x)Y_m^*(x)Y_n(x)\,dx = \sum_m d_m^* Y_m^*(\xi) = g^*(\xi).$$
We will apply the eigenfunction expansion of the Dirac δ-function to Green's functions below.

Note. Principal value integrals. Suppose we wanted to view the function $1/x$ as a distribution. This isn't possible directly because of the divergence at $x = 0$, but we can use the principal value
$$\left(\mathcal{P}\frac{1}{x}\right)[f(x)] = \lim_{\epsilon \to 0^+} \left(\int_{-\infty}^{-\epsilon} \frac{f(x)}{x} dx + \int_{\epsilon}^{\infty} \frac{f(x)}{x} dx \right).$$
All the integrals here are real, but for many applications, $f(x)$ will be a meromorphic complex function. Then we can simply evaluate the principal value integral by taking a contour that goes around the pole at $x = 0$ by a semicircle, and closes at infinity.

Note. We may also regulate $1/x$ by adding an imaginary part to x. The Sokhotsky formula is
$$\lim_{\epsilon \to 0^+} \frac{1}{x + i\epsilon} = \mathcal{P}\frac{1}{x} - i\pi\delta(x)$$
where both sides do not converge as functions, but merely as distributions. This can be shown straightforwardly by integrating both sides against a test function and taking real and imaginary parts; note that we cannot assume the test function is analytic and use contour integration.

Example. A Kramers-Kronig relation. Suppose that our test function $f(x)$ is analytic in the lower-half plane and decays sufficiently quickly. Then applying $1/(x+i\epsilon)$ to $f(x)$ gives zero by contour integration, so we have

$$\mathcal{P}\int_{-\infty}^{\infty}\frac{f(x)}{x}dx = i\pi f(0)$$

by the Sokhotsky formula. In particular, this relates the real and imaginary parts of $f(x)$.

Note. One has to be careful with performing algebra with distributions. Suppose that $xa(x) = 1$ where $a(x)$ is a distribution, and both sides are regarded as distributions. Then dividing by x is not invertible; we instead have

$$a(x) = \mathcal{P}\frac{1}{x} + A\delta(x)$$

where A is not determined. This is important for Green's functions below.

9.4 Green's Functions

Next, we consider Green's functions for second-order ODEs. They are used to solve problems with forcing terms.

- We consider linear differential operators of the form

  $$\mathcal{L} = \alpha(x)\frac{d^2}{dx^2} + \beta(x)\frac{d}{dx} + \gamma(x)$$

  defined on $[a,b]$, and wish to solve the problem $\mathcal{L}y(x) = f(x)$ where $f(x)$ is a forcing term. For mechanical systems, such terms represent literal forces; for first-order systems such as heat, they represent sources.

- We define the Green's function $G(x,\xi)$ of $\mathcal{L}$ to satisfy

  $$\mathcal{L}G = \delta(x-\xi)$$

  where $\mathcal{L}$ always acts solely on x. To get a unique solution, we must also set boundary conditions; for concreteness we choose $G(a,\xi) = G(b,\xi) = 0$.

- The Green's function $G(x,\xi)$ is the response to a δ-function source at ξ. Regarding the equation above as a matrix equation, it is the inverse of $\mathcal{L}$, and the solution to the problem with general forcing is

  $$y(x) = \int_a^b G(x,\xi)f(\xi)\,d\xi.$$

  Here, the integral is just a continuous variant of matrix multiplication. The differential operator $\mathcal{L}$ can be thought of the same way; its matrix elements are derivatives of δ-functions.

- To construct the Green's function, take a basis of solutions $\{y_1, y_2\}$ to the homogeneous equation (i.e. no forcing term) such that $y_1(a) = 0$ and $y_2(b) = 0$. Then we must have

  $$G(x,\xi) = \begin{cases} A(\xi)y_1(x) & x < \xi, \\ B(\xi)y_2(x) & x > \xi. \end{cases}$$

- Next, we need to join these solutions together at $x = \xi$. We know that $\mathcal{L}G$ has only a δ-function singularity at $x = \xi$. Hence the singularity must be provided by the second derivative, or else we would get stronger singularities; then the first derivative has a discontinuity while the Green's function itself is continuous. Explicitly,

$$G(x = \xi^-, \xi) = G(x = \xi^+, \xi), \quad \left.\frac{\partial G}{\partial x}\right|_{x=\xi^-} - \left.\frac{\partial G}{\partial x}\right|_{x=\xi^+} = \frac{1}{\alpha(\xi)}.$$

- Solving the resulting equations gives

$$G(x, \xi) = \frac{1}{\alpha(\xi) W(\xi)} \times \begin{cases} y_1(x) y_2(\xi) & a \le x < \xi, \\ y_2(x) y_1(\xi) & \xi < x \le b. \end{cases}$$

  Here, $W = y_1 y_2' - y_2 y_1'$ is the Wronskian, and it is nonzero because the solutions form a basis.

- This reasoning fully generalizes to higher order ODEs. For an n^{th} order ODE, we have a basis of n solutions, a discontinuity in the $n - 1^{\text{th}}$ derivative, and $n - 1$ continuity conditions.

- If the boundary conditions are inhomogeneous, we use the linearity trick again: we solve the problem with inhomogeneous boundary conditions but no forcing (using our earlier methods), and with homogeneous boundary conditions with forcing.

- We can also compute the Green's function in terms of the eigenfunctions. Letting $G(x, \xi) = \sum_n \hat{G}_n(\xi) Y_n(x)$, and expanding $\mathcal{L}G = \delta(x - \xi)$ gives

$$w(x) \sum_n \hat{G}_n(\xi) \lambda_n Y_n(x) = w(x) \sum_n Y_n(x) Y_n^*(\xi)$$

  which implies $\hat{G}_n(\xi) = Y_n^*(\xi)/\lambda_n$. This is the same result we found several sections earlier.

- Note that the coefficients $\hat{G}_n(\xi)$ are singular if $\lambda_n = 0$. This is simply a manifestation of the fact that $A\mathbf{x} = \mathbf{b}$ has no unique solution if A has a zero eigenvalue.

- For example, consider $\mathcal{L}y = y'' - y$ on $[0, a]$ with boundary conditions $y(0) = y(a) = 0$. Generically, there are no zero eigenvalues, but in the case $a = n\pi$ we have $y = \sin(x)$. Thus, when we're dealing with boundary conditions it can be difficult to see whether a solution is unique; it must be treated on a case-by-case basis. Note that the invertibility of $\mathcal{L}$ depends on the boundary conditions; though the operator $\mathcal{L}$ is fixed, the space on which it acts is determined by the boundary conditions.

- Green's functions can be defined for a variety of boundary conditions. For example, when time is the independent variable with $t \in [t_0, \infty)$, then we might take $y(t_0) = y'(t_0) = 0$. Then the Green's function $G(t, \tau)$ must be zero until $t = \tau$, giving the retarded Green's function. Using a "final" condition instead would give the advanced Green's function.

Example. A driven harmonic oscillator is described by the differential equation

$$\ddot{y} + y = F(t).$$

Suppose the oscillator begins at rest and then experiences a Gaussian pulse, $F(t) = e^{-t^2/\tau^2}$. To find the motion after the driving, we use the retarded Green's function, which is $G(t,t') = \sin(t-t')\theta(t-t')$. After a long time, the complex amplitude of the oscillation is therefore

$$\int_{-\infty}^{\infty} e^{-t'^2/\tau^2} e^{-i\omega t'}\, dt' = \int_{-\infty}^{\infty} e^{-(t'/\tau + i\omega\tau/2)^2} e^{-\omega^2\tau^2/4}\, dt' = \sqrt{\pi}\tau e^{-\omega^2\tau^2/4}.$$

Notice that this goes to zero for both $\tau \to 0$, in which case the pulse is too short to do anything, and for $\tau \to \infty$, in which case it just moves the particle adiabatically, without transferring energy to it. The maximum amplitude occurs for $\tau = \sqrt{2}$.

9.5 Variational Principles

In this section, we consider some problems involving minimizing a functional

$$F[y] = \int_{\alpha}^{\beta} f(y, y', x)\, dx.$$

The Euler–Lagrange equation gives

$$\frac{\partial f}{\partial y} - \frac{d}{dx}\frac{\partial f}{\partial y'} = 0$$

for fixed endpoints. When f does not depend explicitly on x, Noether's theorem yields

$$f - \frac{\partial f}{\partial y'} y' = \text{const.}$$

This quantity is also called the first integral.

Example. The path of a light ray in the xz plane with $n(z) = \sqrt{a - bz}$. Here, the functional is the total time, and we parametrize the path by $z(x)$. Then

$$f = \frac{dt}{dx} = n(z)\sqrt{1 + z'^2}$$

which has no explicit x-dependence, giving the first integral $\sqrt{(a-bz)/(1+z'^2)}$. Separating and integrating shows that the path is a parabola; a linear $n(z)$ would give a circle.

Example. The brachistochrone. A bead slides on a frictionless wire from $(0,0)$ to (x,y) with y positive in the downward direction. We have

$$f = \frac{dt}{dx} \propto \sqrt{\frac{1 + (y')^2}{y}}$$

which yields the first integral $1/\sqrt{y(1+y'^2)}$. Separating and integrating, then parametrizing appropriately gives

$$x = c(\theta - \sin\theta), \quad y = c(1 - \cos\theta)$$

which is a cycloid.

Example. The isoperimetric problem: maximize the area enclosed by a curve with fixed perimeter. To handle this constrained variation, we use Lagrange multipliers. In general, if we have the constraint $P[y] = c$, then we extremize the functional

$$\Phi[y] = F[y] - \lambda(P[y] - c)$$

without constraint, then pick λ to satisfy the constraint. (For multiple constraints, we just add one term for each constraint, with a different λ_i.) In this case, the area and perimeter are

$$A[y] = \oint_C y(x)\,dx, \quad P[y] = \oint_C \sqrt{1 + (y')^2}\,dx$$

where x is integrated from α to β (for the top half), then back down from β to α (for the bottom half). We must extremize the functional

$$f[y] = y - \lambda\sqrt{1 - y'^2}$$

and the Euler–Lagrange equation applies because there are no endpoints. We thus have the first integral $y - \lambda/\sqrt{1 + (y')^2}$, which can be separated and integrated to show the solution is a circle.

Example. Suppose a charged particle starts at rest and ends at rest, traveling a distance d along a line in a time T. What is the path that minimizes the radiation emitted? This quantity is proportional to

$$U = \int_0^T P\,dt, \qquad P = \left(\frac{d^2x}{dt^2}\right)^2.$$

This is our only example in which the integrand depends on a second derivative, but it can be handled in a similar way. The Euler–Lagrange equation generalizes to

$$\frac{\partial P}{\partial x} - \frac{d}{dt}\frac{\partial P}{\partial \dot{x}} + \frac{d^2}{dt^2}\frac{\partial P}{\partial \ddot{x}} = 0.$$

Only the third term is nonzero, giving $d^2P/dt^2 = 0$, or equivalently $d^4x/dt^4 = 0$. The solution is therefore a cubic. Imposing the boundary conditions yields the solution

$$x(t) = \frac{3d}{T^2}\,t^2 - \frac{2d}{T^3}\,t^3$$

which is a path with constant jerk.

As an application, we consider Noether's theorem.

- We consider a one-parameter family of transformations parametrized by s. To first order,

  $$q \to q + s\delta q, \quad \dot{q} \to \dot{q} + s\delta\dot{q}.$$

  Note that $\delta\dot{q} = (\dot{\delta q})$ because we are varying along paths, on which $\dot{q}$ and q are related.

- For this transformation to be a symmetry, the Lagrangian must change by a total derivative, as this preserves stationary paths of the action,

  $$\delta L = s\left(\delta q\frac{\partial L}{\partial q} + \delta\dot{q}\frac{\partial L}{\partial \dot{q}}\right) = s\frac{dK}{dt}.$$

  Applying the Euler–Lagrange equations, on shell we have

  $$s\frac{dK}{dt} = s\frac{d}{dt}\left(\delta q\frac{\partial L}{\partial \dot{q}}\right) \quad \to \quad \frac{d}{dt}\left(\delta q\frac{\partial L}{\partial \dot{q}} - K\right) = 0.$$

  This is Noether's theorem.

- To get a shortcut for finding a conserved quantity, promote s to a function $s(t)$. Then we pick up an extra term,

$$\delta L = s\left(\delta q\frac{\partial L}{\partial q} + \delta\dot{q}\frac{\partial L}{\partial \dot{q}}\right) + \dot{s}\delta q\frac{\partial L}{\partial \dot{q}} = s\frac{dK}{dt} + \dot{s}\delta q\frac{\partial L}{\partial \dot{q}}$$

  where K is defined as above. Simplifying,

$$\delta L = \frac{d}{dt}(sK) + \dot{s}\left(\delta q\frac{\partial L}{\partial \dot{q}} - K\right)$$

  so that the conserved quantity is the coefficient of $\dot{s}$. This procedure can be done without knowing K beforehand; the point is to simplify the variation into the sum of a total derivative and a term proportional to $\dot{s}$, which is only possible when we are considering a real symmetry.

- We can also phrase the shortcut differently. Suppose we can get the variation in the form

$$\delta L = s\dot{K} + \dot{s}J.$$

  Applying the product rule and throwing away a total derivative,

$$\delta L \sim s(\dot{K} - \dot{J})$$

  and the variation of the action must vanish on-shell for any variation, including a variation from a general $s(t)$. Then we need $\dot{K} - \dot{J} = 0$, so $K - J$ is conserved. This is simply a rephrasing of the previous method. (Note that we can always write δL as linear in s and $\dot{s}$, but the coefficient of s will only be a total derivative when we are dealing with a symmetry.)

- The same setup can be done in Hamiltonian mechanics, where the action is

$$I[q,p] = \int p\dot{q} - H(q,p)\,dt$$

  and q and p are varied independently, with fixed endpoints for x. This is distinct from the Lagrangian picture where q and $\dot{q}$ cannot be varied independently on paths, even if they are off-shell. In the Hamiltonian picture, q and p are only on on-shell paths.

Example. Time translational symmetry. We perform a time shift $\delta q = \dot{q}$, giving

$$\frac{dK}{dt} = \dot{q}\frac{\partial L}{\partial q} + \ddot{q}\frac{\partial L}{\partial \dot{q}} = \frac{dL}{dt} - \frac{\partial L}{\partial t}.$$

If time translational symmetry holds, $\partial L/\partial t = 0$, giving $K = L$ and the conserved quantity

$$H = \dot{q}\frac{\partial L}{\partial \dot{q}} - L.$$

On the other hand, using our shortcut method in Hamiltonian mechanics,

$$q \to q + s\dot{q}, \quad \dot{q} \to \dot{q} + \dot{s}\dot{q} + s\ddot{q}, \quad p \to p + s\dot{p}$$

giving the variation

$$\delta I = \int s\dot{p}\dot{q} + sp\ddot{q} + \dot{s}p\dot{q} - \frac{\partial H}{\partial q}\dot{q} - \frac{\partial H}{\partial p}\dot{p}\,dt = \int \frac{d}{dt}(sp\dot{q} - sH) + \dot{s}H$$

where we used $\partial H/\partial t = 0$. We then directly read off the conserved quantity H.

We can also handle functionals of functions with multiple arguments, in which case the Euler–Lagrange equation gives partial differential equations. Note that this is different from functionals of multiple functions, in which case we get multiple Euler–Lagrange equations.

Example. A minimal surface is a surface of minimal area satisfying some boundary conditions. The functional is

$$F[y] = \int dx_1\, dx_2\, \sqrt{1 + y_1^2 + y_2^2}, \quad y_i = \frac{\partial y}{\partial x_i}$$

which can be seen by rotating into a coordinate system where $y_2 = 0$. Denoting the integrand as f, the Euler–Lagrange equation is

$$\frac{d}{dx_i}\frac{\partial f}{\partial y_i} = \frac{\partial f}{\partial y}$$

and the right-hand side is zero. Simplifying gives the minimal surface equation

$$(1 + y_1^2)y_{22} + (1 + y_2^2)y_{11} - 2y_1 y_2 y_{12} = 0.$$

If the first derivatives are small, this reduces to Laplace's equation $\nabla^2 y = 0$.

Example. Functionals like the one above are common in field theories. For example, the action for waves on a string is

$$S[y] = \frac{1}{2}\int dx\, dt\, (\rho \dot{y}^2 - T y'^2).$$

Using our Euler–Lagrange equation above, there is no dependence on y, giving

$$\frac{d}{dx}(-Ty') + \frac{d}{dt}(\rho \dot{y}) = 0$$

which yields the wave equation. It can be somewhat confusing to treat x and t on the same footing in this way, so sometimes it's easier to set the variation to zero directly.

Example. In geometrical optics, the path of light minimizes the travel time. In a vertically stratified medium, we can parametrize the speed of light by $c(z) = c_0/n(z)$ and the path by $x(z)$, giving

$$c_0 T = \int dz\, \sqrt{1 + (dx/dz)^2}\, n(z).$$

This has translational symmetry along the x direction, and the corresponding conserved quantity is

$$\frac{n(z)(dx/dz)}{\sqrt{1 + (dx/dz)^2}} = \frac{n(z)}{\sqrt{1 + (dz/dx)^2}} = n(z)\sin\theta,$$

where θ is the angle to the vertical. This is Snell's law. For a general $n(\mathbf{r})$, there is no symmetry, and thus no simple analogue of Snell's law; the only way to find the path is to integrate a second-order equation of motion. But for a rotationally invariant medium where n only depends on r, the ray path always lies in a plane, and we can parametrize the path in polar coordinates, giving

$$c_0 T = \int dr\, \sqrt{1 + r^2(d\theta/dr)^2}\, n(r).$$

This has rotational symmetry, and the corresponding conserved quantity is

$$\frac{n(r) r^2\, d\theta/dr}{\sqrt{1 + r^2(d\theta/dr)^2}} = \frac{rn(r)}{\sqrt{1 + (d\log r/d\theta)^2}}.$$

This doesn't have a simple geometric interpretation, since the surfaces of constant n are curved, but it's the closest analogue to Snell's law.

10 Methods for PDEs

10.1 Separation of Variables

We begin by studying Laplace's equation,

$$\nabla^2 \psi = 0.$$

Later, we will apply our results to the study of the heat, wave, and Schrodinger equations,

$$K\nabla^2 \psi = \frac{\partial \psi}{\partial t}, \quad c^2 \nabla^2 \psi = \frac{\partial^2 \psi}{\partial t^2}, \quad -\nabla^2 \psi + V(\mathbf{x})\psi = i\frac{\partial \psi}{\partial t}.$$

Separating the time dimension in these equations will often yield a Helmholtz equation in space,

$$\nabla^2 \psi + k^2 \psi = 0.$$

Finally, an important variant of the wave equation is the massive Klein-Gordan equation,

$$c^2 \nabla^2 \psi - m^2 \psi = \frac{\partial^2 \psi}{\partial t^2}.$$

As shown in electromagnetism, the solution to Laplace's equation is unique given Dirichlet or Neumann boundary conditions. We always work in a compact spatial domain Ω.

Example. In two dimensions, Laplace's equation is equivalent to

$$\frac{\partial^2 \psi}{\partial z \partial \overline{z}} = 0$$

where $z = x + iy$. Thus the general solution is $\psi(x, y) = \phi(z) + \chi(\overline{z})$ where ϕ and χ are holomorphic and antiholomorphic. For example, suppose we wish to solve Laplace's equation inside the unit disc subject to $\psi = f(\theta)$ on the boundary. We may write the boundary condition as a Fourier series,

$$f(\theta) = \sum_{n \in \mathbb{Z}} \hat{f}_n e^{in\theta}.$$

Now note that at $|z| = 1$, z^n and $\overline{z}^{-n}$ reduce to $e^{in\theta}$. Thus the solution inside the disc is

$$\psi(x, y) = \hat{f}_0 + \sum_{n=1}^{\infty} (\hat{f}_n z^n + \hat{f}_{-n} \overline{z}^n)$$

which is indeed the sum of a holomorphic and antiholomorphic function. Similarly, to get a bounded solution outisde the disc, we simply flip the powers.

Next, we introduce the technique of separation of variables.

- Suppose the boundary conditions are given in a three-dimensional rectangular region. Then it is convenient to separate in Cartesian coordinates. Writing

  $$\psi(x, y, z) = X(x)Y(y)Z(z)$$

  and plugging into Laplace's equation gives

  $$\frac{X''(x)}{X(x)} + \frac{Y''(y)}{Y(y)} + \frac{Z''(z)}{Z(z)} = 0.$$

- Thus every term must be independently constant, so

$$X'' = -\lambda X, \quad Y'' = -\mu Y, \quad Z'' = (\lambda + \mu)Z.$$

- Generally, we see that separation converts PDEs into individual Sturm-Liouville problems, with a specified relation between the eigenvalues (in this case, they must sum to zero). Each solution is a normal mode of the system – we've seen this vocabulary before, applied to eigenvalues in time. Homogeneous boundary conditions (e.g. 'zero on this surface') then give constraints on the allowed eigenvalues.

- Finally, we arrive at a set of allowed solutions and superpose them to satisfy a set of given inhomogeneous boundary conditions. This is often simplified by the orthogonality of the eigenfunctions; we project the inhomogeneous term onto each one.

We now apply the same principle, but in spherical polar coordinates.

- In spherical coordinates, the Laplacian is

$$\nabla^2 = \frac{1}{r^2}\partial_r(r^2\partial_r) + \frac{1}{r^2\sin\theta}\partial_\theta(\sin\theta\partial_\theta) + \frac{1}{r^2\sin^2\theta}\partial_\phi^2.$$

  For simplicity, we consider only axisymmetric solutions with no ϕ dependence.

- Separating $\psi(r,\theta) = R(r)\Theta(\theta)$ yields the equations

$$\frac{d}{d\theta}\left(\sin\theta\frac{d\Theta}{d\theta}\right) + \lambda\sin\theta\,\Theta = 0, \quad \frac{d}{dr}\left(r^2\frac{dR}{dr}\right) - \lambda R = 0.$$

- For the angular equation, we substitute $x = \cos\theta$, so that $x \in [-1,1]$, giving

$$\frac{d}{dx}\left((1-x^2)\frac{d\Theta}{dx}\right) = -\lambda\Theta.$$

  This is a Sturm-Liouville equation, which is self adjoint because $p(\pm1) = 0$, with weight function $w(x) = 1$. The solutions are hence orthogonal on $[-1,1]$.

- The solutions are the Legendre polynomials, obeying the Rodriguez formula

$$P_\ell(x) = \frac{1}{2^\ell\ell!}\frac{d^\ell}{dx^\ell}(x^2-1)^\ell, \quad \lambda = \ell(\ell+1), \quad \ell = 0,1,\ldots.$$

  They can be found by guessing a series solution and demanding the series truncates to a finite-degree polynomial. An explicit calculation shows that

$$\int_{-1}^{1} P_m(x)P_\ell(x)\,dx = \frac{2}{2\ell+1}\delta_{m\ell}.$$

  As in the previous example, any axisymmetric boundary condition on a sphere can be expanded in Legendre polynomials.

- Finally, the radial equation has solution

$$R_\ell(r) = A_\ell r^\ell + \frac{B_\ell}{r^{\ell+1}}.$$

  If we demand our solution to decay at $r \to \infty$, or to be regular at $r = 0$, then we can throw out the A_ℓ or B_ℓ.

- As an application, applying our results to the field of a point charge gives the multipole expansion, where $\ell = 0$ is the monopole, $\ell = 1$ is the dipole, and so on.
- Allowing for dependence on ϕ, the ϕ equation has solution $\Phi(\phi) = e^{im\phi}$ for integer m, while the θ equation yields an associated Legendre function; the radial equation remains the same.

In cylindrical coordinates, we encounter Bessel functions in the radial equation.

- Separating $\psi = R(r)\Theta(\theta)Z(z)$, we find that $\Theta(\theta) = e^{inz}$ and $Z(z) = e^{-z\sqrt{\mu}}$, while the radial equation becomes
$$r^2R'' + rR' + (\mu r^2 - \lambda)R = 0.$$
Converting to the Sturm-Liouville form gives
$$\frac{d}{dr}\left(r\frac{dR}{dr}\right) - \frac{n^2}{r}R = -\mu r R$$
which has the weight function $w(r) = r$.
- The eigenvalue μ doesn't matter because it simply sets the length scale. Eliminating it by setting $x = r\sqrt{\mu}$ gives Bessel's equation of order n,
$$x^2\frac{d^2R}{dx^2} + x\frac{dR}{dx} + (x^2 - n^2)R = 0.$$
The solutions are the Bessel functions $J_n(x)$ and $Y_n(x)$.
- The Bessel functions of the first kind, $J_n(x)$, are regular at the origin, but the $Y_n(x)$ are not; thus we can ignore them if we care about the region $x \to 0$.
- For small x, we have
$$J_n(x) \sim x^n, \quad Y_n(x) \sim x^{-n}$$
while for large x, we have
$$J_n(x) \sim \frac{\cos(x - n\pi/2 - \pi/4)}{\sqrt{x}}, \quad Y_n(x) \sim \frac{\sin(x - n\pi/2 - \pi/4)}{\sqrt{x}}.$$
The decrease $1/\sqrt{x}$ is consistent with our intuition for a cylindrical wave.
- We also encounter Bessel functions in two-dimensional problems in polar coordinates after separating out time; in that case time plays the same role that z does here.
- Solving the Helmholtz equation in three dimensions (again, often encountered by separating out time) yields the spherical Bessel functions $j_n(x)$ and $y_n(x)$. They behave somewhat like regular Bessel functions of order $n + 1/2$, but fall as $1/x$ for large x instead.

Next, we turn to the heat equation. Since it involves time, we write its solutions as Φ, while ψ is reserved for space only.

- For positive diffusion constant K, the heat equation 'spreads heat out', so it is only defined for $t \in [0, \infty)$. If we try to follow the time evolution backwards, we generically get singularities at finite time.

- The heat flux is $K\nabla\Phi$. Generally, we can show that the total heat $\int \Phi\, dV$ is conserved as long as no heat flux goes through the boundary.
- Another useful property is that if $\Phi(x,t)$ solves the heat equation, then so does $\Phi(\lambda x, \lambda^2 t)$, as can be checked explicitly. Then the time dependence of any solution can be written as a function of the similarity variable $\eta = x/\sqrt{Kt}$.
- For the one-dimensional heat equation, $\partial\Phi/\partial t = K\partial^2\Phi/\partial x^2$, we can write the solution as $\Phi(x,t) = F(\eta)/\sqrt{Kt}$. Then the equation reduces to

  $$2F' + \eta F = \text{const.}$$

  This shows that the normalized solution with $F'(0) = 0$ is

  $$G(x,t) = \frac{\exp(-x^2/4Kt)}{\sqrt{4\pi Kt}}.$$

  This is called the heat kernel, or the fundamental solution of the heat equation; at $t = 0$ it limits to $\delta(x)$. Convolving it with the state at time t_0 gives the state at time $t_0 + t$.
- Separating out time, $\Phi = T(t)\psi(\mathbf{r})$ gives the Helmholtz equation,

  $$\nabla^2\psi = -\lambda\psi, \quad T(t) = e^{-\lambda t}, \quad \lambda > 0.$$

  That is, high eigenvalues are quickly suppressed. For example, if we work on the line, where the spatial solutions are exponentials, and recall the decay properties of Fourier series, evolution under the heat equation for an infinitesimal time removes discontinuities!
- Since the heat equation involves time, we must also supply an initial condition along with standard spatial boundary conditions. We now prove uniqueness for Dirichlet conditions in time and space. Let Φ_1 and Φ_2 be solutions and let $\delta\Phi$ be their difference. Then

  $$\frac{d}{dt}\int_\Omega \delta\Phi^2\, dV \propto \int_\Omega (\delta\Phi)\nabla^2\delta\Phi\, dV = -\int_\Omega (\nabla\delta\Phi)^2 dV \leq 0$$

  where we integrated by parts and applied the boundary conditions to remove the surface term. Then the left-hand side is decreasing, but it starts at zero by the initial conditions, so it is always zero. (We can also show this by separating variables.)
- The spatial domain Ω must be compact for the integrals above to exist. For example, in an infinite domain we can have heat forever flowing in from infinity, giving a nonunique solution.

Example. The cooling of the Earth. We model the Earth as a sphere of radius R with an isotropic heat distribution and initial conditions

$$\Phi(r,0) = \Phi_0 \text{ for } r < R, \quad \Phi(R,t) = 0 \text{ for } t > 0$$

so that the Earth starts with a uniform temperature, with zero temperature at the surface (i.e. outer space). We separate variables by $\Phi(r,t) = R(r)T(t)$ giving

$$\frac{d}{dr}\left(r^2\frac{dR}{dr}\right) = -\lambda^2 r^2 R, \quad \frac{dT}{dt} = -\lambda^2 KT.$$

The radial equation has sinusoids decaying as $1/r$ for solutions,

$$R(r) = B_\lambda \frac{\sin(\lambda r)}{r} + C_\lambda \frac{\cos(\lambda r)}{r}.$$

For regularity at $r = 0$, we require $C_\lambda = 0$. To satisfy the homogeneous boundary condition, we set $\lambda = n\pi/R$, giving the solution

$$\Phi(r,t) = \frac{1}{r} \sum_{n\in\mathbb{Z}} A_n \sin\left(\frac{n\pi r}{R}\right) \exp\left(-\frac{n^2\pi^2}{r^2} Kt\right).$$

We then choose the coefficients A_n to fit the inhomogeneous initial condition. At time $t = 0$,

$$r\Theta_0 = \sum_{n\in\mathbb{Z}} A_n \sin\left(\frac{n\pi r}{R}\right) \quad \to \quad A_n = \Theta_0 \int_0^R \sin\left(\frac{n\pi r}{R}\right) r\, dr = (-1)^{n+1} \frac{\Theta_0 R}{n\pi}.$$

The solution is not valid for $r > R$ because the thermal diffusivity K changes, from the value for rock to the value for air.

Note. Solving problems involving the wave equation is rather similar; the only difference is that we get oscillation in time rather than exponential decay, and that we need both an initial position and velocity. To prove uniqueness, we use the energy functional

$$E = \frac{1}{2} \int_\Omega \ddot{\phi} + c^2(\nabla\phi)^2\, dV$$

which is positive definite and conserved. Then the difference of two solutions has zero initial energy, so it must be zero.

Note. There is no fundamental difference between initial conditions and (spatial) boundary conditions: they both are conditions on the boundary of the spacetime region where the PDE holds; Dirichlet and Neumann boundary conditions correspond exactly to initial positions and velocities. However, in practice they are treated differently because the time condition is 'one-sided': while we can specify that a rope is held at both of its ends, we usually can't specify where it'll be both now and in the future. As a result, while we only often need one (two-sided) boundary condition to get uniqueness, we need as many initial conditions as there are time derivatives.

Note. In our example above, the initial condition is inhomogeneous and the boundary condition is homogeneous. But if both were inhomogeneous, our method would fail because we wouldn't have any conditions to constrain the eigenvalues. In this case the trick is to use linearity, which turns the problem into the sum of two problems, each with one homogeneous condition.

10.2 The Fourier Transform

Fourier transforms extend Fourier series to nonperiodic functions $f : \mathbb{R} \to \mathbb{C}$.

- We define the Fourier transform $\tilde{f} = F[f]$ by

$$\tilde{f}(k) = \int e^{-ikx} f(x)\, dx.$$

All integrals in this section are over the real line. The Fourier transform is linear, and obeys

$$F[f(x-a)] = e^{-ika}\tilde{f}(k), \quad F[e^{i\ell x} f(x)] = \tilde{f}(k-\ell), \quad F[f(cx)] = \frac{\tilde{f}(k/c)}{|c|}.$$

- Defining the convolution of two functions as

$$(f * g)(x) = \int f(x-y)g(y)\,dy$$

the Fourier transform satisfies

$$F[f * g] = F[f]F[g].$$

- Finally, the Fourier transform converts differentiation to multiplication,

$$F[f'(x)] = ik\tilde{f}(k).$$

This allows differential equations with forcing to be rewritten nicely. If $\mathcal{L}(\partial)y(x) = f(x)$,

$$F[\mathcal{L}(\partial)y] = \mathcal{L}(ik)\tilde{y}(k), \quad \tilde{y}(k) = \tilde{f}(k)/\mathcal{L}(ik).$$

- The Fourier transform can be inverted by

$$f(x) = \frac{1}{2\pi}\int e^{ikx}\tilde{f}(k)\,dk.$$

This can be derived by taking the continuum limit of the Fourier series. In particular,

$$f(-x) = \frac{1}{2\pi}F[\tilde{f}(k)]$$

which implies that $F^4 = (2\pi)^2$. Intuitively, a Fourier transform is a rotation in (x, p) phase space by 90 degrees.

- Parseval's theorem carries over, as

$$(f, f) = \frac{1}{2\pi}(\tilde{f}, \tilde{f}).$$

This expression also holds replacing the second f with g, as unitary transformations preserve inner products.

- Defining the Fourier transform of a δ-function requires some more distribution theory, but naively we have $F[\delta(x)] = 1$, with the inverse Fourier transform implying the integral

$$\int e^{-ikx}\,dx = 2\pi\delta(k).$$

This result only makes sense in terms of distributions. As corollaries, we have

$$F[\delta(x-a)] = e^{-ika}, \quad F[e^{i\ell x}] = 2\pi\delta(k-\ell)$$

which imply

$$F[\cos(\ell x)] = \pi(\delta(k+\ell) + \delta(k-\ell)), \quad F[\sin(\ell x)] = i\pi(\delta(k+\ell) - \delta(k-\ell)).$$

Example. The Fourier transform of a step function $\Theta(x)$ is subtle. In general, the Fourier transforms of ordinary functions can be distributions, because functions in Fourier space are only linked to observable quantities in real space via integration. Naively, we would have $1/ik$ since δ is the derivative of Θ, but this is incorrect because dividing by k gives us extra $\delta(k)$ terms we haven't determined. Instead, we add an infinitesimal damping $\Theta(x) \to \Theta(x)e^{-\epsilon x}$ giving

$$\mathcal{F}\Theta = \lim_{\epsilon \to 0^+} \frac{1}{\epsilon + ik} = \mathcal{P}\frac{1}{ik} + \pi\delta(k)$$

by the Sokhotsky formula. As a consistency check, we have

$$\mathcal{F}[\Theta(-x)] = -\mathcal{P}\frac{1}{ik} + \pi\delta(k)$$

and the two sum to $2\pi\delta(k)$, which is indeed the Fourier transform of 1.

Note. There is an alternative way to think about the Fourier transform of the step function. For any function $f(x)$, split

$$f(x) = f_+(x) + f_-(x)$$

where the two terms have support for positive and negative x respectively. Then take the Fourier transform of each piece. The point of this split is that for nice functions, the Fourier integral

$$\tilde{f}_+(k) = \int_0^\infty f_+(x)e^{ikx}\,dx$$

will converge as long as $\operatorname{Im} k$ is sufficiently large; note we are now thinking of k as complex-valued. The Fourier transform can be inverted as long as we follow a contour across the complex k plane in this region of large $\operatorname{Im} k$. For the step function, we hence have

$$\mathcal{F}\Theta = \frac{1}{ik}, \quad \operatorname{Im} k > 0.$$

The expression is not valid at $\operatorname{Im} k = 0$, so we cannot integrate along this axis. This removes the ambiguity of whether we cross the pole above or below, at the cost of having to keep track of where in the complex plane $\mathcal{F}\Theta$ is defined. Often, as here, we can analytically continue $\tilde{f}_+$ and $\tilde{f}_-$ to a much greater region of the complex plane. A Fourier inversion contour is then valid as long as it passes above all the singularities of $\tilde{f}_+$ and below those of $\tilde{f}_-$. In a more general situation, there could also be branch cuts that obstruct the contour.

Example. Solving a differential equation by Fourier transform. Let $(\partial^2 + m^2)\phi(x) = -\rho(x)$. In the naive approach, we have

$$(k^2 - m^2)\tilde{\phi}(k) = \tilde{\rho}(k)$$

from which we conclude the Green's function is

$$\tilde{G}(k) = \frac{1}{k^2 - m^2}.$$

Then, to find the solution to the PDE, we perform the inverse Fourier transform for

$$\phi(\mathbf{x}) = \frac{1}{2\pi}\int \frac{e^{ikx}\tilde{\rho}(k)}{k^2 - m^2}\,dk.$$

However, this integral does not exist, so we must resort to performing a contour integral around the poles. This *ad hoc* procedure makes more sense using distribution theory. We can't really divide by k^2+m^2 since $\tilde{G}(k)$ is a distribution, so instead

$$\tilde{G}(k) = \mathcal{P}\frac{1}{k^2+m^2} + g_1\delta(k-m) + g_2\delta(k+m)$$

with g_1 and g_2 undetermined, reflecting the fact that the Green's function is not uniquely defined without boundary conditions. By the Sokhotsky formula, we can go back and forth between the principal value and the $i\epsilon$ regulator at the cost of modifying g_1 and g_2. This is extremely useful because of the link between causality and analyticity, as we saw for the Kramers-Kronig relations. In particular, the retarded and advanced Green's functions are just

$$\tilde{G}_{\text{ret}}(k) = \frac{1}{k^2-m^2-i\epsilon k}, \quad \tilde{G}_{\text{adv}}(k) = \frac{1}{k^2-m^2+i\epsilon k}$$

with no need for more delta function terms at all. Similarly, if we had a PDE instead, the general Green's function would be

$$\tilde{G}(\mathbf{k}) = \mathcal{P}\frac{1}{k^2+m^2} + g(\mathbf{k})\delta(k^2-m^2)$$

and the function $g(\mathbf{k})$ must be determined by boundary conditions.

Example. Solving another differential equation using a Fourier transform in the complex plane. We consider Airy's equation

$$\frac{d^2y}{dx^2} + xy = 0.$$

We write the solution as a generalized Fourier integral

$$y(x) = \int_\Gamma g(\zeta)e^{x\zeta}\,d\zeta.$$

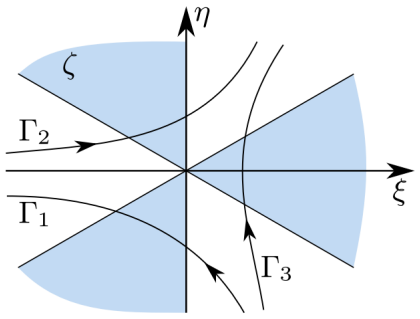
Plugging this in and integrating by parts, we have

$$g(\zeta)e^{x\zeta}\Big|_\Gamma \int_\Gamma (\zeta^2 g(\zeta) - g'(\zeta))e^{x\zeta}\,d\zeta = 0$$

which must vanish for all x. The first term is evaluated at the endpoints of the contour. For the second term to vanish for all x, we must have

$$g'(\zeta) = \zeta^2 g(\zeta), \quad g(\zeta) = Ce^{\zeta^3/3}.$$

At this point, this might seem strange, as we were supposed to have two independent solutions. But note that in order for $g(\zeta)e^{x\zeta}$ to vanish at the endpoints, the contour must go to infinity in one of the unshaded regions below.



If we take a contour that starts and ends in the same region, then we will get zero by Cauchy's theorem. Then there are two independent contours, starting in one region and ending in another, giving the two independent solutions; all others are related by summation or negation. Of course, the integrals cannot be performed in closed form, but for large x the integrals are amenable to saddle point approximation.

Note. The discrete Fourier transform applies to functions defined on $\mathbb{Z}_n$ and is useful for computing. It's independent of the Fourier series we considered earlier; their common property of a discrete spectrum comes from the compactness of the domains S^1 and $\mathbb{Z}_n$. More generally, we can perform Fourier analysis on any Abelian group, or even any compact, possibly non-Abelian group.

Example. Fourier transforms are useful for linear time-translation invariant (LTI) systems, $\mathcal{L}I = O$. These are more general than linear differential operators, as $\mathcal{L}$ might integrate I or impose a time delay. However, their response is local in frequency space, because if $\mathcal{L}(e^{i\omega t}) = O(t)$, then

$$\mathcal{L}(e^{i\omega(t-t_0)}) = O(t-t_0) = O(t)e^{-i\omega t_0}$$

which shows that $O(t) \propto e^{i\omega t}$. Thus we can write

$$\tilde{O}(\omega) = \tilde{I}(\omega)\tilde{R}(\omega)$$

where $\tilde{R}$ is called the transfer function or system function. Taking an inverse Fourier transform gives $O(t) = (I * R)(t)$, so R behaves like a Green's function; it is called the response function.

As an explicit example, consider the case

$$\sum_{i=0}^{n} a_i \frac{d^i O(t)}{dt^i} = I(t)$$

where R is simply a Green's function. In this case we have

$$\tilde{R}(\omega) = \frac{1}{a_0 + a_1 i\omega + \cdots + a_n (i\omega)^n} = \frac{1}{a_n} \prod_{j=1}^{J} \frac{1}{(i\omega - c_j)^{k_j}} = \sum_{j=1}^{J} \sum_{m=1}^{k_j} \frac{\Gamma_{mj}}{(i\omega - c_j)^m}$$

where the c_j are the roots of the polynomial and the k_j are their multiplicities, and we used partial fractions in the last step. In the case $m = 1$, we recall the result from the example above,

$$\mathcal{F}[e^{\alpha t}\Theta(t)] = \frac{1}{i\omega - \alpha}, \quad \mathrm{Re}(\alpha) < 0.$$

Therefore, using the differentiation rule, we have

$$\mathcal{F}[(t^m e^{\alpha t}/m!)\Theta(t)] = \frac{1}{(i\omega - \alpha)^{m+1}}, \quad \mathrm{Re}(\alpha) < 0$$

which provides the general solution for $R(t)$. We see that oscillatory/exponential solutions appear as poles in the complex plane, while higher-order singularities provide higher-order resonances.

Example. Stabilization by negative feedback. Consider a system function $\tilde{R}(\omega)$. We say the system is stable if it doesn't have exponentially growing modes; this corresponds to $\tilde{R}(\omega)$ having no poles in the upper half-plane. Now suppose we attempt to stabilize a system by adding negative feedback,

feeding the output scaled by $-r$ and time delayed by t_0 back into the input. Defining the feedback factor $k = re^{i\omega t_0}$, the new system function is

$$\tilde{R}(\omega)_{\text{loop}} = \frac{\tilde{R}(\omega)}{1 + k\tilde{R}(\omega)}$$

by the geometric series formula; this result is called Black's formula. Then the new poles are given by the zeroes of $1 + \alpha\tilde{R}(\omega)$.

The Nyquist criterion is a graphical method for determining whether the new system is stable. We consider a contour C along the real axis and closed along the upper half-plane, encompassing all poles and zeroes of $\tilde{R}(\omega)$. The Nyquist plot is a plot of $\tilde{R}(\omega)$ along C. By the argument principle, the number of times the Nyquist plot wraps around -1 is equal to the number of poles P of $\tilde{R}(\omega)$ in the upper-half plane minus the number of zeroes of $k\tilde{R}(\omega) + 1$ in the upper-half plane. Then the system is stable if the Nyquist plot wraps around -1 exactly P times. This is useful since we only need to know P, not the location of the poles or the number of zeroes.

Note. Causality is 'built in' to the Fourier transform. As we've seen in the above examples, damping that occurs forward in time (as required by $\text{Re}(\alpha) < 0$) automatically yields singularities only in the upper-half plane, and causal/retarded Green's functions that vanish for $t < 0$.

In general, the Green's functions returned by the Fourier transform are regular for $|t| \to \infty$, which serves as an extra implicit boundary condition. For example, for the damped harmonic oscillator we have

$$\tilde{G}(\omega) = \frac{1}{\omega_0^2 - \omega^2 - i\gamma\omega}$$

which yields a unique $G(t, \tau)$, because the advanced solution (which blows up at $t \to -\infty$) has been thrown out. On the other hand, for the undamped harmonic oscillator,

$$\tilde{G}(\omega) = \frac{1}{\omega_0^2 - \omega^2}$$

the Fourier inversion integral diverges, so $G(t, \tau)$ cannot be defined. We must specify a 'pole prescription', which corresponds to an infinitesimal damping. Forward damping gives the retarded Green's function, and reverse damping gives the advanced Green's function. Note that there's no analogue of the Feynman Green's function; that appears in field theory because there are both positive and negative-energy modes.

10.3 The Method of Characteristics

We begin by stepping back and reconsidering initial conditions and boundary conditions.

- Initial conditions and boundary conditions specify the value of a function ϕ and/or its derivatives, on a surface of codimension 1. In general, such information is called Cauchy data, and solving a PDE along with given Cauchy data is called a Cauchy problem.
- A Cauchy problem is well-posed if there exists a unique solution which depends continuously on the Cauchy data. We've seen that the existence and uniqueness problem can be subtle.
- We have already seen that the backwards heat equation is ill-posed. Another example is Laplace's equation on the upper-half plane with boundary conditions

$$\phi(x, 0) = 0, \quad \partial_y\phi(x, 0) = g(x), \quad g(x) = \frac{\sin(Ax)}{A}.$$

In this case the solution is

$$\phi(x,y) = \frac{\sin(Ax)\sinh(Ay)}{A^2}$$

which diverges in the limit $A \to \infty$, through the exponential dependence in $\sinh(Ay)$, even though $g(x)$ continuously approaches zero.

The method of characteristics helps us formalize how solutions depend on Cauchy data.

- We begin with the case of a first order PDE in $\mathbb{R}^2$,

$$\alpha(x,y)\partial_x\phi + \beta(x,y)\partial_y\phi = f(x,y).$$

Such a PDE is called quasi-linear, because it is linear in ϕ, but the functions α and β are not linear in x and y.

- Defining the vector field $\mathbf{u} = (\alpha, \beta)$, the PDE becomes

$$\mathbf{u} \cdot \nabla\phi = f.$$

The vector field $\mathbf{u}$ defines a family of integral curves, called characteristic curves,

$$C_t(s) = \{x(s,t), y(s,t)\}$$

where s is the parameter along the curve and t identifies the curve, satisfying

$$\left.\frac{\partial x}{\partial s}\right|_t = \alpha|_{C_t}, \quad \left.\frac{\partial y}{\partial s}\right|_t = \beta|_{C_t}.$$

- In the (s,t) coordinates, the PDE becomes a family of ODEs,

$$\left.\frac{\partial \phi}{\partial s}\right|_t = f|_{C_t}$$

Therefore, for a unique solution to exist, we must specify Cauchy data at exactly one point along each characteristic curve, i.e. along a curve B transverse to the characteristic curves. The value of the Cauchy data at that point determines the value of ϕ along the entire curve. Each curve is completely independent of the rest!

Example. The 1D wave equation is $(\partial_x^2 - \partial_t^2)\phi = 0$, which contains both right-moving and left-moving waves. The simpler equation $(\partial_x - \partial_t)\phi = 0$ only contains right-moving waves; the characteristic curves are $x - t = \text{const}$.

Example. We consider the explicit example

$$e^x\partial_x\phi + \partial_y\phi - 0, \quad \phi(x,0) = \cosh x.$$

The vector field $(e^x, 1)$ has characteristics satisfying

$$\frac{dx}{ds} = e^x, \quad \frac{dy}{ds} = 1$$

which imply

$$e^{-x} = -s + c, \quad y = s + d$$

where the constants c and d reflect freedom in the parametrizations of s and t. To fix s, we demand that the characteristic curves pass through B at $s = 0$. To fix t, we parametrize B itself by $(x, y) = (t, 0)$. This yields

$$e^{-x} = -s + e^{-t}, \quad y = s$$

and the solution is simply $\phi(s, t) = \cosh t$. Inverting gives the result

$$\phi(x, y) = \cosh \log(y + e^{-x}).$$

We could also add an inhomogeneous term on the right without much more effort.

Next, we generalize to the case of second-order PDEs, which yield new features.

- Consider a general second-order linear differential operator

  $$\mathcal{L} = a^{ij}(x)\partial_i\partial_j + b^i(x)\partial_i + c(x), \quad x \in \mathbb{R}^n$$

  where we choose a^{ij} to be symmetric. We define the symbol of $\mathcal{L}$ to be

  $$\sigma(x, k) = a^{ij}(x)k_ik_j + b^i(x)k_i + c(x).$$

  We similarly define the symbol of a PDE of general order.

- The principle part of the symbol, $\sigma^P(x, k)$, is the leading term. In the second-order case it is an x-dependent quadratic form,

  $$\sigma^P(x, k) = \mathbf{k}^T A\mathbf{k}.$$

- We classify $\mathcal{L}$ by the eigenvalues of A. The operator $\mathcal{L}$ is
  - elliptic if the eigenvalues all have the same sign (e.g. Laplace)
  - hyperbolic if all but one of the eigenvalues have the same sign (e.g. wave)
  - ultrahyperbolic if there is more than one eigenvalue with each sign (requires $d \geq 4$)
  - parabolic if there is a zero eigenvalue (i.e. the quadratic form is degenerate) (e.g. heat)

- We will focus on the two-dimensional case, where we have

  $$A = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$$

  and $\mathcal{L}$ is elliptic if $ac - b^2 > 0$, hyperbolic if $ac - b^2 < 0$, and parabolic if $ac - b^2 = 0$. The names come from the conic section $\mathcal{L}$ is in Fourier space.

- When the coefficients are constant, then the Fourier transform of $\mathcal{L}$ is the symbol $\sigma(ik)$. Another piece of intuition is that the principle part of the symbol dominates when the solution is rapidly varying.

- From our previous work, we've seen that typically we need:
  - Dirichlet or Neumann boundary conditions on a closed surface, for elliptic equations
  - Dirichlet and Neumann boundary conditions on an open surface, for hyperbolic equations
  - Dirichlet or Neumann boundary conditions on an open surface, for parabolic equations

Generically, stricter boundary conditions will not have solutions, or will have solutions that depend very sensitively on them.

Now we apply the method of characteristics for second-order PDEs.

- In this case, the Cauchy data consists of the value of ϕ on a surface Γ along with the normal derivative $\partial_{\mathbf{n}}\phi$. Let $\mathbf{t}_i$ denote the other directions. In order to propagate the Cauchy data to a neighboring surface, we need to know the normal second derivative $\partial_{\mathbf{n}}\partial_{\mathbf{n}}\phi$.
- Since we know ϕ on all of Γ, we know $\partial_{\mathbf{t}_i}\partial_{\mathbf{t}_j}\phi$ and $\partial_{\mathbf{n}}\partial_{\mathbf{t}_i}\phi$. To attempt to find $\partial_{\mathbf{n}}\partial_{\mathbf{n}}\phi$ we use the PDE, which is
$$a_{ij}\frac{\partial^2\phi}{\partial x_i\partial x_j} = \text{known}.$$
Therefore, we know the value of $a_{nn}\partial_{\mathbf{n}}\partial_{\mathbf{n}}\phi$, which gives the desired result unless a_{nn} is zero.
- We define a characteristic surface Σ to be one whose normal vector n^μ obeys $a_{\mu\nu}n^\mu n^\nu = 0$. Then we can propagate forward the Cauchy data on Γ as long as it is nowhere tangent to a characteristic surface.
- Generically, a characteristic surface has dimension one. In two dimensions, they are lines, and an equation is hyperbolic, parabolic, or elliptic at a point if it has two, one, or zero characteristic curves through that point.

Example. The wave equation is the archetypal hyperbolic equation. It's easiest to see its characteristic curves in 'light-cone' coordinates where $\xi_\pm = x \pm ct$, where it becomes
$$\frac{\partial^2\phi}{\partial\xi_+\partial\xi_-} = 0.$$
Then the characteristic curves are curves of constant $\xi_\pm$. Information is propagated along these curves in the sense that the general solution is $f(\xi_+) + g(\xi_-)$. On the other hand, the value of ϕ at a point depends on all the initial Cauchy data in its past light cone; the 'domain of dependence' is instead bounded by characteristic curves.

10.4 Green's Functions for PDEs

We now find Green's functions for PDEs, using the Fourier transform. We begin with the case of an unbounded spatial domain.

- We consider the Cauchy problem for the heat equation on $\mathbb{R}^n \times [0,\infty)$,
$$D\nabla^2\phi = \frac{\partial\phi}{\partial t}, \quad \phi(\mathbf{x}, t=0) = f(\mathbf{x}), \quad \lim_{\mathbf{x}\to\infty}\phi(\mathbf{x},t) = 0.$$
To do this, we find the solution for initial condition $\delta(\mathbf{x})$ (called the fundamental solution) by Fourier transform in space, giving
$$S_n(\mathbf{x},t) = \mathcal{F}^{-1}[e^{-Dk^2t}] = \frac{e^{-x^2/4Dt}}{(4\pi Dt)^{n/2}}.$$
The general solution is given by convolution with the fundamental solution. As expected, the position x only enters through the similarity variable x^2/t. We also note that the heat equation is nonlocal, as $S_n(\mathbf{x},t)$ is nonzero for arbitrarily large $\mathbf{x}$ at arbitrarily small t.

- We can also solve the heat equation with forcing and homogeneous initial conditions,

$$\frac{\partial \phi}{\partial t} - D\nabla^2 \phi = F(\mathbf{x}, t), \quad \phi(\mathbf{x}, t = 0) = 0.$$

  In this case, we want to find a Green's function $G(\mathbf{x}, t, \mathbf{y}, \tau)$ representing the response to a δ-function source at $(\mathbf{y}, t)$. Duhamel's principle states that it is simply related to the fundamental solution,

$$G(\mathbf{x}, t, \mathbf{y}, \tau) = \Theta(t - \tau) S_n(\mathbf{x} - \mathbf{y}, t - \tau).$$

  To understand this, note that we can imagine starting time at $t = \tau^+$. In this case, we don't see the δ-function driving; instead, we see its outcome, a δ-function initial condition at $\mathbf{y}$. The general solution is given by convolution with the Green's function.

- In both cases, a time direction is picked out by specifying $\phi(t = 0)$ and solving for ϕ at times $t > 0$. In particular, this forces us to get the retarded Green's function.

- As another example, we consider the forced wave equation on $\mathbb{R}^n \times (0, \infty)$ for $n = 3$,

$$\frac{\partial^2 \phi}{\partial t^2} - c^2 \nabla^2 \phi = F, \quad \phi(t = 0) = \partial_t \phi(t = 0) = 0.$$

  Taking the spatial Fourier transform, the Green's function satisfies

$$\left(\frac{\partial^2}{\partial t^2} + k^2 c^2 \right) \tilde{G}(\mathbf{k}, t, \mathbf{y}, \tau) = e^{-i\mathbf{k}\cdot\mathbf{y}} \delta(t - \tau).$$

  Applying the initial condition and integrating gives

$$\tilde{G}(\mathbf{k}, t, \mathbf{y}, \tau) = \Theta(t - \tau) e^{-i\mathbf{k}\cdot\mathbf{y}} \frac{\sin(kc(t - \tau))}{kc}.$$

  This result holds in all dimensions.

- To take the Fourier inverse, we perform the $\mathbf{k}$ integration in spherical coordinates, but the final angular integration is only nice in odd dimensions. In three dimensions, we find

$$G(\mathbf{x}, t, \mathbf{y}, \tau) = -\frac{\delta(|\mathbf{x} - \mathbf{y}| - c(t - \tau))}{4\pi c |\mathbf{x} - \mathbf{y}|}$$

  so that a force at the origin makes a shell that propagates at speed c. In one dimension, we instead have $G(\mathbf{x}, t, \mathbf{y}, \tau) \sim \theta(|\mathbf{x} - \mathbf{y}| - c(t - \tau))$, so we find a raised region whose boundary propagates at speed c. In even dimensions, we can't perform the $e^{ikr\cos\theta}\, d\theta$ integral. Instead, we find a boundary that propagates with speed c with a long tail behind it.

- Another way to phrase this is that in one dimension, the instantaneous force felt a long distance from the source is a delta function, just like the source. In three dimensions, it is the derivative. Then in two dimensions, it is the half-derivative, but this is not a local operation.

- The same result can be found by a temporal Fourier transform, or a spacetime Fourier transform. In the latter case, imposing the initial condition to get the retarded Green's function is a little more subtle, requiring a pole prescription.

- For the wave equation, Duhamel's principle relates the Green's function to the solution for an initial velocity but zero initial position.

The Green's function is simply related to the fundamental solution only on an unbounded domain. In the case of a bounded domain Ω, Green's functions must additionally satisfy boundary conditions on $\partial\Omega$. However, it is still possible to construct a Green's function using a fundamental solution.

Example. The method of images. Consider Laplace's equation defined on a half-space with homogeneous Dirichlet boundary conditions $\phi = 0$. The fundamental solution is the field of a point charge. The Green's function can be constructed by putting another point charge with opposite charge, 'reflected' in the plane; choosing the same charge would work for homogeneous Neumann boundary conditions.

The exact same reasoning works for the wave equation. Dirichlet boundary conditions correspond to a hard wall, and we imagine an upside-down 'ghost wave' propagating the other way. Similarly, for the heat equation, Neumann boundary conditions correspond to an insulating barrier, and we can imagine a reflected, symmetric source of heat.

For less symmetric domains, Green's functions require much more work to construct. We consider the Poisson equation as an extended example.

- We begin with finding the fundamental solution to Poisson's equation,

$$\nabla^2 G_n(\mathbf{x}) = \delta^n(\mathbf{x}).$$

Applying rotational symmetry and integrating over a ball of radius r,

$$1 = \int_{B_r} \nabla^2 G_n \, dV = \int_{\partial B_r} \nabla G_n \cdot d\mathbf{S} = r^{n-1} \frac{dG_n}{dr} \int_{S^{n-1}} d\Omega_n.$$

Denoting A_n as the area of the $(n-1)$-dimensional sphere, we have

$$G_n(\mathbf{x}) = \begin{cases} x + c_1 & n = 1, \\ \frac{\log x}{2\pi} + c_2 & n = 2, \\ -\frac{1}{A_n(n-2)} \frac{1}{x^{n-2}} + c_n & n \geq 3. \end{cases}$$

For $n \geq 3$ the constant can be set to zero if we require $G_n \to 0$ for $x \to \infty$. Otherwise, we need additional constraints. We then define $G_n(\mathbf{x}, \mathbf{y}) = G_n(\mathbf{x} - \mathbf{y})$, which is the response at $\mathbf{x}$ to a source at $\mathbf{y}$.

- Next, we turn to solving the Poisson equation on a compact domain Ω. We begin with deriving some useful identities. For any regular functions $\phi, \psi : \Omega \to \mathbb{R}$,

$$\int_{\partial\Omega} \phi \nabla\psi \cdot d\mathbf{S} = \int_\Omega \nabla \cdot (\phi \nabla\psi) \, dV = \int_\Omega \phi \nabla^2 \psi + (\nabla\phi) \cdot (\nabla\psi) \, dV$$

by the divergence theorem. This is Green's first identity. Antisymmetrizing gives

$$\int_\Omega \phi \nabla^2 \psi - \psi \nabla^2 \phi = \int_{\partial\Omega} (\phi \nabla\psi - \psi \nabla\phi) \cdot d\mathbf{S}$$

which is Green's second identity.

- Next, we set $\psi(\mathbf{x}) = G_n(\mathbf{x}, \mathbf{y})$ and $\nabla^2 \phi(\mathbf{x}) = -F(\mathbf{x})$, giving Green's third identity

$$\phi(\mathbf{y}) = -\int_\Omega G_n(\mathbf{x}, \mathbf{y}) F(\mathbf{x}) \, dV + \int_{\partial\Omega} (\phi(\mathbf{x}) \nabla G_n(\mathbf{x}, \mathbf{y}) - G_n(\mathbf{x}, \mathbf{y}) \nabla\phi(\mathbf{x})) \cdot d\mathbf{S}$$

where we used a delta function to do an integral, and all derivatives are with respect to $\mathbf{x}$.

- At this point it looks like we're done, but the problem is that generally we can only specify ϕ or $\nabla\phi \cdot \hat{\mathbf{n}}$ at the boundary, not both. Once one is specified, the other is determined by uniqueness, so the equation above is really an expression for ϕ in terms of itself, not a closed form for ϕ.

- For concreteness, suppose we take Dirichlet boundary conditions $\phi|_{\partial\Omega} = g$. We define a Dirichlet Green's function $G = G_n + H$ where H satisfies Laplace's equation throughout Ω and $G|_{\partial\Omega} = 0$. Then using Green's third identity gives

$$\phi(\mathbf{y}) = \int_{\partial\Omega} g(\mathbf{x}) \nabla G(\mathbf{x}, \mathbf{y}) \cdot d\mathbf{S} - \int_{\Omega} G(\mathbf{x}, \mathbf{y}) F(\mathbf{x})\, dV$$

  which is the desired closed-form expression! Of course, at this point the hard task is to construct H, but at the very least this problem has no source terms.

- As a concrete example, we can construct an explicit form for H whenever the method of images applies. For example, for a half-space it is the field of a reflected opposite charge.

- Similarly, we can construct a Neumann Green's function. There is a subtlety here, as the integral of $\nabla\phi \cdot d\mathbf{S}$ must be equal to the integral of the driving F, by Gauss's law. If this doesn't hold, no solution exists.

- The surface terms can be given a physical interpretation. Suppose we set $\phi|_{\partial\Omega} = 0$ in Green's third identity, corresponding to grounding the surface $\partial\Omega$. At the surface, we have

$$(\nabla\phi) \cdot \hat{\mathbf{n}} \propto \mathbf{E}_{\perp} \propto \rho$$

  which means that the surface term is just accounting for the field of the screening charges.

- Similarly, we can interpret the surface term in our final result, when we turn on a potential $\phi|_{\partial\Omega} = g$. To realize this, we make $\partial\Omega$ the inner surface of a very thin capacitor. The outer surface $\partial\Omega'$, just outside $\partial\Omega$, is grounded. The surfaces are split into parallel plates and hooked up to batteries with emf $g(\mathbf{x})$, giving locally opposite charge densities on $\partial\Omega'$ and $\partial\Omega$. Then the potential g can be thought of as coming from nearby opposite sheets of charge. The term ∇G describes such sources, by thinking of the derivative as a finite difference.

11 Approximation Methods

11.1 Asymptotic Series

We illustrate the ideas behind perturbation theory with some algebraic equations with a small parameter ϵ, before moving onto differential equations. We begin with some motivating examples which will bring us to asymptotic series.

Example. Solve the equation

$$x^2 + \epsilon x - 1 = 0.$$

The exact solution is

$$x = -\frac{\epsilon}{2} \pm \sqrt{1 + \frac{\epsilon^2}{4}} = \begin{cases} 1 - \frac{\epsilon}{2} + \frac{\epsilon^2}{8} + \dots \\ -1 - \frac{\epsilon}{2} + \frac{\epsilon^2}{8} + \dots \end{cases}.$$

This series converges for $|\epsilon| < 2$ and rapidly if ϵ is small; it is a model example of the perturbation method. Now we show two ways to find the series without already knowing the exact answer.

First, rearrange the equation to the form $x = f(x)$,

$$x = \pm\sqrt{1 - \epsilon x}.$$

Then we may use successive approximations,

$$x_{n+1} = \sqrt{1 - \epsilon x_n}.$$

The starting point x_0 can be chosen to be an exact solution when $\epsilon = 0$, in this case $x_0 = 1$. Then

$$x_1 = \sqrt{1 - \epsilon}, \quad x_2 = \sqrt{1 - \epsilon\left(1 - \frac{\epsilon}{2}\right)}$$

and so on. The x_n term matches the series up to the ϵ^n term. To see why, note that if the desired fixed point is x_*, then

$$x_{n+1} - x_* = f(x_n) - x_* = f(x_* + x_n - x_*) - x_* \approx (x_n - x_*)f'(x_*).$$

Near the fixed point we have $f'(x_*) \approx -\epsilon/2$, so the error decreases by a factor of ϵ every iteration. The most important part of this method is to choose f so that $f'(x_*)$ is small, ensuring rapid convergence. For instance, if we had $f'(x_*) \sim 1 - \epsilon$ instead, convergence could be very slow.

Second, expand about one of the roots when $\epsilon = 0$ in a series in ϵ,

$$x = 1 + \epsilon x_1 + \epsilon^2 x_2 + \dots.$$

By plugging this into the equation, expanding in powers of ϵ, and setting each coefficient to zero, we may determine the x_i iteratively. This tends to be easier when working to higher orders. In general, one might need to expand in a different variable than ϵ, but this works for regular problems.

Example. Solve the equation

$$\epsilon x^2 + x - 1 = 0.$$

This is more subtle because there are two roots for any $\epsilon > 0$, but only one root for $\epsilon = 0$. Problems where the $\epsilon \to 0$ limit differs in an important way from the $\epsilon = 0$ case are called singular. The exact solutions are

$$x = \frac{-1 \pm \sqrt{1 + 4\epsilon}}{2\epsilon} = \begin{cases} 1 - \epsilon + 2\epsilon^2 + \dots \\ -\frac{1}{\epsilon} - 1 + \epsilon - 2\epsilon^2 + \dots \end{cases}$$

where the series converges for $|\epsilon| < 1/4$. We see the issue is that one root diverges to infinity. We can capture it using the expansion method by starting the series with ϵ^{-1},

$$x = \frac{x_{-1}}{\epsilon} + x_0 + \epsilon x_1 + \dots.$$

This also captures the regular root in the case $x_{-1} = -1$. However, we again only knew to start the series at $1/\epsilon$ by using the exact solution.

We can arrive at the same conclusion by changing variables by a rescaling,

$$x = X/\epsilon, \quad X^2 + x - \epsilon = 0.$$

This is now a regular problem which can be handled as above. Again, the difficult part is choosing the right rescaling to accomplish this. Consider the general rescaling $x = \delta X$, which gives

$$\epsilon\delta^2 X^2 + \delta X - 1 = 0.$$

The rescaling is good if the formerly singular root becomes $O(1)$. We would thus like at least two of the quantities $(\epsilon\delta^2, \delta, 1)$ to be similar in size, with the rest much smaller. This gives a regular perturbation problem, where the similar terms give an $O(1)$ root, and the rest perturb it slightly. By casework, this only happens for $\delta \sim 1$ and $\delta \sim 1/\epsilon$, giving the regular and singular roots respectively. This method is called finding the "dominant balance" or "distinguished limit".

Example. Solve the equation

$$(1-\epsilon)x^2 - 2x + 1 = 0.$$

We see that when $\epsilon = 0$ we have a double root $x = 1$. Naively taking

$$x = 1 + \epsilon x_1 + \epsilon^2 x_2$$

we immediately find the equations

$$\epsilon^0 : 0 = 0, \quad \epsilon^1 : 0 = 1.$$

To see the problem, consider one of the exact solutions,

$$x = \frac{1}{1-\epsilon^{1/2}} = 1 + \epsilon^{1/2} + \epsilon + \epsilon^{3/2} + \dots.$$

Hence we should have expanded in powers of $\epsilon^{1/2}$,

$$x = 1 + \epsilon^{1/2} x_{1/2} + \epsilon x_1 + \dots.$$

Setting the coefficient of $\epsilon^{n/2}$ to zero determines $x_{(n-1)/2}$.

To find the expansion sequence in general, we suppose

$$x = 1 + \delta_1 x_1, \quad \delta_1(\epsilon) \ll 1$$

and substitute it in. Simplifying, we find

$$\delta_1^2 x_1^2 - \epsilon + 2\epsilon\delta_1 x_1 + \delta_1^2 \epsilon x_1^2 = 0.$$

We now apply dominant balance again. The last two terms are always subleading, so balancing the first two gives $\delta_1 = \epsilon^{1/2}$, from which we determine $x_1 = 1$. At this point we could guess the next term is $O(\epsilon)$, but to be safe we could repeat the procedure, setting

$$x = 1 + \epsilon^{1/2} + \delta_2 x_2, \quad \delta_2(\epsilon) \ll \epsilon^{1/2}.$$

However, this rapidly gets more complicated for higher orders.

Finally, we could use the iterative method. We choose

$$x_{n+1} = 1 \pm \epsilon^{1/2} x_n$$

which ensures rapid convergence. Taking the positive root and starting with $x_0 = 1$ gives

$$x_1 = 1 + \epsilon^{1/2}, \quad x_2 = 1 + \epsilon^{1/2} + \epsilon, \quad \ldots.$$

Example. Solve the equation

$$xe^{-x} = \epsilon.$$

One root is near $x = 0$ and is easy to approximate, as we may expand the exponential in a series; the other becomes large as $\epsilon \to 0$. The expansion series is not obvious, so we use the iterative procedure. We know that when $x = L \equiv \log 1/\epsilon$,

$$xe^{-x} = \epsilon L \gg \epsilon.$$

On the other hand, when $x = 2L$,

$$xe^{-x} = 2\epsilon^2 L \ll \epsilon.$$

Hence the desired solution is approximately L. The easiest way to proceed is with the iterative method. We rearrange the equation to

$$x_{n+1} = L + \log x_n$$

and choose $x_0 = L$. Then, omitting absolute value signs for brevity,

$$x_1 = L + \log L, \quad x_2 = L + \log(L + \log L) = L + \log L + \log\left(1 + \frac{\log L}{L}\right).$$

The final logarithm can be expanded in a series, and continuing gives us an expansion with terms of the form $(\log L)^m/L^n$. Even for tiny ϵ, L is not very large, and $\log L$ isn't either. Hence the series converges very slowly.

Since we are working with expansions more general than convergent power series, we formalize them as asymptotic expansions.

- We say $f = O(g)$ as $\epsilon \to 0$ if there exists K and ϵ_0 so that $|f| < K|g|$ for all $\epsilon < \epsilon_0$.
- We say $f = o(g)$ as $\epsilon \to 0$ if $f/g \to 0$ as $\epsilon \to 0$.
- A set of functions $\{\phi_n(\epsilon)\}$ is an asymptotic sequence as $\epsilon \to 0$ if, for each n and $i > 0$, $\phi_{n+i}(\epsilon) = o(\phi_n(\epsilon))$ as $\epsilon \to 0$.

- A function $f(\epsilon)$ has an asymptotic expansion with respect to the asymptotic sequence $\{\phi_n(\epsilon)\}$ as $\epsilon \to 0$ if there exists constants so that

$$f(\epsilon) \sim \sum_n a_n \phi_n(\epsilon)$$

which stands for

$$f(\epsilon) = \sum_{n=0}^{N} a_n \phi_n(\epsilon) + o(\phi_N(\epsilon))$$

for all N.

- Given $\{\phi_n\}$, the coefficients a_n of f are unique. This is easily proven by induction. However, the converse is not true: the coefficients a_n don't determine f. Just like ordinary power series, we may be missing terms that are smaller than any of the ϕ_n.

- The above definition of asymptotic expansion implies that as $\epsilon \to 0$, for all $N \geq 0$,

$$\lim_{\epsilon \to 0} \frac{f(\epsilon) - \sum_{n=0}^{N} f_n(\epsilon)}{f_N(\epsilon)} = 0.$$

That is, unlike the regular definition of convergence, we take $\epsilon \to 0$ rather than $N \to \infty$.

- Asymptotic series may be integrated term by term. However, they may not be differentiated term by term, because unlike power series, the functions $f_n(\epsilon)$ may be quite singular (e.g. $\epsilon \cos(1/\epsilon)$) and grow much larger than expected upon differentiating.

- Asymptotic series may be plugged into each other, but some care must be taken. For example, taking the exponential of only the leading terms of a series may give a completely wrong result; we must instead take all terms of order 1 or higher.

- As we've seen above, the terms in an asymptotic series can get quite complicated. However, it is at least true that functions obtained by a finite number of applications of $+$, $-$, $\times$, $\nabla\cdot$, exp, and log may always be ordered; these are called Hardy's logarithmico-exponential functions.

Example. Often an asymptotic expansion works better than a convergent power series. We have

$$\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2}\, dt = \frac{2}{\sqrt{\pi}} \int_0^z \sum_{n=0}^{\infty} \frac{(-t^2)^n}{n!}\, dt = \frac{2}{\sqrt{\pi}} \sum_{n=0}^{\infty} \frac{(-1)^n z^{2n+1}}{(2n+1)n!}$$

where all manipulations above are valid since the series has an infinite radius of convergence. However, for large z the series converges very slowly, and many terms in the series are much larger than the final result, so roundoff error affects the accuracy.

A better series can be constructed by noting

$$\operatorname{erf}(z) = 1 - \frac{2}{\sqrt{\pi}} \int_z^{\infty} e^{-t^2}\, dt.$$

We now integrate by parts using

$$\int_z^{\infty} e^{-t^2}\, dt = \int_z^{\infty} \frac{2te^{-t^2}}{2t}\, dt = \frac{e^{-z^2}}{2z} - \int_z^{\infty} \frac{e^{-t^2}}{2t^2}\, dt.$$

Iterating this procedure gives

$$\operatorname{erf}(z) = 1 - \frac{e^{-z^2}}{z\sqrt{\pi}}\left(1 - \frac{1}{2z^2} + \frac{3!!}{(2z^2)^2} - \frac{5!!}{(2z^2)^3} + \dots\right).$$

This series diverges for all z, with radius of convergence zero. However, it is an asymptotic series as $z \to \infty$. For large z, cutting off the series even at a few terms gives a very good approximation. For any fixed z, the series eventually diverges as more terms are included; generally the optimal truncation is to stop at the smallest term.

One might worry that asymptotic series don't give a guarantee of quality, since the series can get worse as more terms are used, but in practical terms, the usual definition of convergence doesn't guarantee quality either. In physics, our expansion parameters will usually be much closer to zero than the our number of terms will be to infinity, so using an asymptotic series will be more accurate. And in numerics, the roundoff errors due to the large terms in a convergent series can make the result inaccurate no matter how many terms we take.

11.2 Asymptotic Evaluation of Integrals

Now we turn to some techniques for asymptotic evaluation of integrals. As we've seen above, the simplest method is repeated integration by parts.

Example. If $f(\epsilon)$ is smooth near $\epsilon = 0$, then

$$f(\epsilon) = f(0) + \int_0^\epsilon f'(x)\,dx.$$

Integrating by parts gives

$$f(\epsilon) = f(0) + (x-\epsilon)f'(x)\Big|_0^\epsilon + \int_0^\epsilon (\epsilon - x)f''(x)\,dx.$$

It's not hard to see that by repeating this, we just recover the Taylor series.

Example. We would like to evaluate

$$I(x) = \int_x^\infty e^{-t^4}\,dt$$

in the limit $x \to \infty$. Integrating by parts,

$$I(x) = -\frac{1}{4}\int_x^\infty \frac{1}{t^3}\frac{d}{dt}(e^{-t^4})\,dt = \frac{e^{-x^4}}{4x^3} - \frac{3}{4}\int_x^\infty \frac{1}{t^4}e^{-t^4}\,dt.$$

This is the beginning of an asymptotic series because the remainder term is at most $I(x)/x^4$, and the ratio vanishes as $x \to \infty$. For large x, even the first term alone is a good approximation.

Example. As a trickier example, we evaluate

$$I(x) = \int_0^x t^{-1/2}e^{-t}\,dt$$

in the limit $x \to \infty$. However, the simplest approach

$$I(x) = -t^{-1/2}e^{-t}\Big|_0^x - \frac{1}{2}\int_0^x t^{-3/2}e^{-t}\,dt$$

gives a singular boundary term. Instead, we evaluate

$$I(x) = I(\infty) - \int_x^\infty t^{-1/2} e^{-t}\, dt, \quad I(\infty) = \Gamma(1/2) = \sqrt{\pi}.$$

The second term may be integrated by parts, giving

$$I(x) = \sqrt{\pi} - \frac{e^{-x}}{\sqrt{x}} + \frac{1}{2}\int_x^\infty t^{-3/2} e^{-t}\, dt$$

which is the start of an asymptotic series. In general, integration by parts fails if the endpoints yield contributions larger than the original integral itself. The reason such large contributions can appear is that every round of integration by parts makes the remaining integral more singular at $t = 0$ by differentiating the $t^{-1/2}$.

Example. We evaluate

$$I(x) = \int_0^\infty e^{-xt^2}\, dt$$

in the limit $x \to \infty$. Naive integration by parts yields a singular boundary term and an infinite remaining integral. In fact, integration by parts cannot possibly work here because the exact answer is $\sqrt{\pi}/2\sqrt{x}$, a fractional power. Integration by parts also doesn't work if the dominant contribution is from an interior point rather than an endpoint, which would have occurred if the lower bound were not 0.

Laplace's method may be used to find

$$I(x) = \int_a^b f(t) e^{x\phi(t)}\, dt$$

in the limit $x \to \infty$, where $f(t)$ and $\phi(t)$ are real and continuous.

Example. Find the asymptotic behavior of

$$I(x) = \int_0^{10} \frac{e^{-xt}}{1+t}\, dt$$

as $x \to \infty$. For high x the integrand is localized near $t = 0$. Hence we split

$$I(x) = \int_0^\epsilon \frac{e^{-xt}}{1+t}\, dt + O(e^{-\epsilon x}), \quad \frac{1}{x} \ll \epsilon \ll 1.$$

Concretely, we could take $\epsilon = 1/\sqrt{x}$. For the remaining integral, change variable to $s = xt$ to yield

$$I(x) \sim \frac{1}{x}\int_0^{x\epsilon} \frac{e^{-s}}{1+s/x}\, ds.$$

Since s/x is small in the entire integration range, we Taylor expand the denominator for

$$I(x) \sim \frac{1}{x}\int_0^{x\epsilon} e^{-s} \sum_n \frac{(-s)^n}{x^n}\, ds = \sum_{n=0}^\infty \frac{1}{x^{n+1}} \int_0^{x\epsilon} (-s)^n e^{-s}\, ds.$$

By extending the upper limit of integration to infinity, we pick up $O((\epsilon x)^n e^{-\epsilon x})$ error terms. Also, by interchanging the order of summation and integration, we have produced an asymptotic series,

$$I(x) \sim \sum_n \frac{1}{x^{n+1}} \int_0^\infty (-s)^n e^{-s}\, ds = \sum_n \frac{(-1)^n n!}{x^{n+1}}.$$

Note that we could have gotten an easier, better bound by extending the upper bound of integration to infinity at the start, but we do things in this order to show the general technique.

Now we develop Laplace's method formally.

- Laplace's method is justified by Watson's lemma: if $f(t)$ is continuous on $[0, b]$ and has the asymptotic expansion
$$f(t) \sim t^\alpha \sum_{n=0}^{\infty} a_n t^{\beta n}$$
as $t \to 0^+$, where $\alpha > -1$ and $\beta > 0$, then
$$I(x) = \int_0^b f(t) e^{-xt}\, dt \sim \sum_{n=0}^{\infty} \frac{a_n \Gamma(\alpha + \beta n + 1)}{x^{\alpha+\beta n+1}}$$
as $x \to +\infty$. The conditions $\alpha > -1$ and $\beta > 0$ ensure the integral converges, and in the case $b = \infty$ we also require $f(t) = O(e^{ct})$ for some constant c at infinity. Watson's lemma can also be used to justify the methods below.

- In the case where the asymptotic series for f is uniformly convergent in a neighborhood of the origin, then Watson's lemma may be established by interchanging the order of integration and summation. Otherwise, we cut off the sums at a finite number of terms and simply show the error terms are sufficiently small to have an asymptotic series.

- We now consider the general integral
$$I(x) = \int_a^b f(t) e^{x\phi(t)}\, dt.$$
The dominant contribution comes from the maximum of $\phi(t)$, which can occur at the endpoints or at an interior point. We'll find only the leading contribution in each case.

- First, suppose the maximum is at $t = a$, and set $a = 0$ for simplicity. As in the example,
$$I(x) = \int_0^\epsilon f(x) e^{x\phi(t)}\, dt + \int_\epsilon^b f(t) e^{x\phi(t)}\, dt, \quad x^{-1} \ll \epsilon \ll x^{-1/2}.$$
Then the second term is $O(e^{x\epsilon\phi'(0)})$ smaller than the first, and hence negligible if $x\epsilon \gg 1$.

- In the first term we assume we can expand $\phi(t)$ and $f(t)$ in the asymptotic series
$$\phi(t) \sim \phi(0) + t\phi'(0) + \ldots, \quad f(t) \sim f(0) + tf'(0) + \ldots$$
where generically $\phi'(0) \neq 0$. Changing variables to $s = xt$,
$$I(x) \sim \frac{e^{x\phi(0)}}{x} \int_0^{x\epsilon} \left(f(0) + \frac{s}{x} f'(0) \right) e^{s\phi'(0) + s^2\phi''(0)/2x + \ldots}\, ds.$$
Given that $s^2/x \ll 1$, which is equivalent to $\epsilon \ll x^{-1/2}$, the second-order term in the integral can be neglected. Similarly, the $(s/x)f'(0)$ term may be neglected.

- Now the upper bound of integration can be extended to ∞ with exponentially small error, for
$$I(x) \sim -\frac{f(a) e^{x\phi(a)}}{x\phi'(a)}.$$
There are also higher-order corrections which we can compute by taking higher-order terms in the series. The overall error once these corrections are taken care of is exponentially small.

- Maxima at interior points are a bit more subtle since ϕ' vanishes there. In this case suppose the maximum is at $c = 0$ for simplicity, and split the integral as

$$I(x) = \int_a^{-\epsilon} f(t)e^{x\phi(t)}\,dt + \int_{-\epsilon}^{\epsilon} f(t)e^{x\phi(t)}\,dt + \int_{\epsilon}^{b} f(t)e^{x\phi(t)}\,dt.$$

  As before the first and third terms are exponentially small, and negligible if $x\epsilon^2 \gg 1$, where the different scaling occurs because the linear term $\phi'(0)$ vanishes.

- Within the second integral we expand

$$\phi(t) \sim \phi(0) + \frac{t^2}{2}\phi''(0) = \frac{t^3}{6}\phi'''(0) + \ldots, \quad f(t) \sim f(0) + tf'(0) + \ldots$$

  where generically $\phi''(0) \neq 0$. Changing variables to $s = \sqrt{x}t$,

$$I(x) \sim \frac{e^{x\phi(0)}}{\sqrt{x}} \int_{-\sqrt{x}\epsilon}^{\sqrt{x}\epsilon} (f(0) + \frac{s}{x}f'(0) + \ldots)e^{s^2\phi''(0)/2 + s^3\phi'''(0)/6\sqrt{x} + \cdots}\,ds.$$

  For the leading term to dominate, we need $\sqrt{x}\epsilon/x \ll 1$ and $(\sqrt{x}\epsilon)^3/\sqrt{x} \ll 1$. The latter is more stringent, and putting together our constraints gives

$$x^{-1/2} \ll \epsilon \ll x^{-1/3}.$$

- Finally, incurring another exponentially small error by extending the integration bounds to $\pm\infty$, we conclude that

$$I(x) \sim \sqrt{\frac{2\pi}{-x\phi''(c)}} f(c)e^{x\phi(c)}.$$

Now we turn to the method of stationary phase.

- The method of stationary phase is used for integrals of the form

$$I(x) = \int_a^b f(t)e^{ix\psi(t)}\,dt$$

  where $\psi(t)$ is real.

- The rigorous foundation of the method is the Riemann-Lebesgue lemma: if the integral $\int_a^b f(t)\,dt$ is absolutely convergent and $\psi(t)$ is continuously differentiable on $[a, b]$ and not constant on any subinterval of $[a, b]$, then

$$\int_a^b f(t)e^{ix\psi(t)}\,dt \to 0$$

  as $x \to \infty$.

- The Riemann-Lebesgue lemma makes it easy to get leading endpoint contributions. For instance,

$$I(x) = \int_0^1 \frac{e^{ixt}}{1+t}\,dt = -\frac{ie^{ix}}{2x} + \frac{i}{x} - \frac{i}{x}\int_0^1 \frac{e^{ixt}}{(1+t)^2}\,dt$$

  and the Riemann-Lebesgue lemma ensures the remaining term is subleading.

- As in Laplace's method, it's more subtle to find contributions from interior points. We get a large contribution at every point ψ' vanishes, since we don't get rapid phase cancellation in that region. Concretely, suppose the only such point is $\psi'(c) = 0$. We split the integral as

$$I(x) = \int_a^{c-\epsilon} f(t)e^{ix\psi(t)}\,dt + \int_{c-\epsilon}^{c+\epsilon} f(t)e^{ix\psi(t)}\,dt + \int_{c+\epsilon}^{b} f(t)e^{ix\psi(t)}\,dt$$

  for $\epsilon \ll 1$. For the first term, we integrate by parts to find

$$\int_a^{c-\epsilon} f(t)e^{ix\psi(t)}\,dt = \frac{f(t)}{ix\psi'(t)}e^{ix\psi(t)}\Big|_a^{c-\epsilon} + \text{subleading} = O\left(\frac{1}{x\psi'(c-\epsilon)}\right) = O\left(\frac{1}{x\epsilon\psi''(c)}\right).$$

  We pick up a similar contribution from the second term. Note that unlike Laplace's method, these error terms are only algebraically small, not exponentially small.

- For the second term, we expand

$$f(t) \sim f(c) + (t-c)f'(c) + \ldots, \quad \psi(t) \sim \psi(c) + \frac{(t-c)^2}{2}\psi''(c) + \frac{(t-c)^3}{6}\psi'''(c) + \ldots.$$

  Plugging this in and changing variables to $s = x^{1/2}(t-c)$ we get

$$\frac{e^{ix\psi(c)}}{x^{1/2}} \int_{-x^{1/2}\epsilon}^{x^{1/2}\epsilon} \left(f(c) + \frac{s}{x^{1/2}}f'(c) + \ldots\right) e^{i\frac{s^2}{2}\psi''(c) + i\frac{s^3}{6x^{1/2}}\psi'''(c) + \ldots}\,ds.$$

  The third-derivative term in the exponent is smaller by a factor of $s^3/x^{1/2}$, so it is subleading if $\epsilon \ll x^{-1/3}$. Similarly, the $f'(c)$ term is smaller by a factor of $s/x^{1/2}$, so it is subleading if $\epsilon \ll 1$.

- Therefore, the leading term is

$$\frac{f(c)e^{ix\psi(c)}}{x^{1/2}} \int_{-x^{1/2}\epsilon}^{x^{1/2}\epsilon} e^{i\frac{s^2}{2}\psi''(c)}\,ds.$$

  When we extend the limits of integration to $\pm\infty$, we pick up $O(1/x\epsilon)$ error terms as before. The integral can then be done by contour integration, rotating the contour to yield a Gaussian integral, to conclude

$$I(x) = \frac{\sqrt{2\pi}f(c)e^{ix\psi(c)}e^{\pm i\pi/4}}{x^{1/2}|\psi''(c)|^{1/2}} + O(1/x\epsilon).$$

  In order for this to be the leading term, it must be greater than $O(1/x\epsilon)$ and hence $\epsilon \gg x^{-1/2}$.

- Putting our most stringent constraints together, we require

$$x^{-1/2} \ll \epsilon \ll x^{-1/3}$$

  just as for Laplace's method for an interior point. Unfortunately it's difficult to improve the approximation, because the next terms involve nonlocal contributions.

Finally, we consider the method of steepest descents.

- Laplace's method and the method of stationary phase are really just special cases of the method of steepest descents, which is for contour integrals of the form

$$I(x) = \int_C f(t)e^{x\phi(t)}\,dt.$$

  We might think naively that the greatest contribution comes from the maximum of $\operatorname{Re}\phi$, but this is incorrect due to the rapid phase oscillations. Similarly regions with zero stationary phase may have negligible magnitude.

- To get more insight, write $\phi = u + iv$. The Cauchy-Riemann equations tell us that u and v are harmonic functions with $(\nabla u)\cdot(\nabla v) = 0$. Hence the landscape of u consists of hills and valleys at infinity, along with saddle points. Assuming the contour goes to infinity, it must follow a path where $u \to -\infty$ at infinity.

- Now consider deforming the path so that v is constant. Then the path is parallel to ∇u, so it generically follows paths of steepest descent. Since u goes to $-\infty$ at infinity, there must be points where $u' = 0$ in the contour, with each point giving a contribution by Laplace's method. Note that if we instead took u constant we would use the method of stationary phase, but this is less useful because the higher-order terms are much harder to compute.

- In general we have some flexibility in the contour. Since the contribution is local, we only need to know which saddle points it passes through, and which poles we cross. This is also true computationally: switching to something close to the steepest descent contour makes numerical evaluation much easier, but we don't have to compute the exact contour for this to work.

- One might worry how to determine which saddle points are relevant. If all the zeroes of f are simple, there is no problem because each saddle point is only connected to one valley; the relevant saddle points are exactly those connected to the valley at the endpoints at infinity. We are free in principle to deform the contour to pass through other saddle points, but we'd pick up errors from the regions of high u that are much larger than the value of the integral.

Example. The gamma function for $x \gg 1$. We may define the gamma function by

$$\frac{1}{\Gamma(x)} = \frac{1}{2\pi i}\int_C e^t t^{-x}\,dt$$

where C is a contour which starts at $t = -\infty - ia$, encircles the branch cut which we take to lie along the negative real axis, and ends at $t = \infty + ib$. Rewriting the integrand at $e^{t - x\log t}$, there is a saddle at $t = x$. But since x is large, it's convenient to rescale,

$$\frac{1}{\Gamma(x)} = \frac{1}{2\pi i x^{x-1}}\int_C e^{x(s-\log s)}\,ds, \quad t = xs.$$

Defining $\phi(s) = s - \log s$, the saddle is now at $s = 1$. The steepest descent contour passes through $s = 1$ vertically. Near this point we have

$$\phi(s) \sim 1 + \frac{(s-1)^2}{2} - \frac{(s-1)^3}{3} + \dots.$$

Rescaling by $u = \sqrt{x}(s-1)$ we have

$$\frac{1}{\Gamma(x)} \sim \frac{e^x}{2\pi i x^{x-1}\sqrt{x}}\int e^{\frac{u^2}{2} - \frac{u^3}{3\sqrt{x}} + \dots}\,du.$$

As usual, we extend the range of integration to infinity, giving

$$\frac{1}{\Gamma(x)} \sim \frac{e^x}{2\pi i x^{x-1/2}} \int e^{u^2/2}\, du = \frac{e^x}{\sqrt{2\pi} x^{x-1/2}}$$

where the integral converges since u ranges from $-i\infty$ to $i\infty$. This is the usual Stirling's approximation, but we can get increasingly accurate approximations by going to higher order.

Example. The Airy function for $x \gg 1$. The Airy function is defined by

$$\mathrm{Ai}(x) = \frac{1}{2\pi} \int_C e^{i(t^3/3+xt)}\, dt.$$

Dividing the plane into six sextants like quadrants, the integrand only decays in the first, third, and fifth sextants, and the contour starts at infinity in the third sextant and ends at infinity in the first.

Differentiating the exponent shows the saddle points are at $t = \pm i x^{1/2}$. Rescaling $t = x^{1/2} z$,

$$\mathrm{Ai}(x) = \frac{x^{1/2}}{2\pi} \int_C e^{x^3/2\phi(z)}, \quad \phi(x) = i(z^3/3 + z).$$

The steepest descent contour goes through the saddle point $z = i$ but not $z = -i$, giving

$$\mathrm{Ai}(x) \sim \frac{e^{-2x^{3/2}/3}}{2\sqrt{\pi} x^{1/4}}.$$

Now consider $\mathrm{Ai}(-x)$ for $x \gg 1$. In this case the saddle points are at $t = \pm 1$ and both are relevant. Adding the two contributions gives

$$\mathrm{Ai}(-x) \sim \frac{1}{\sqrt{\pi} x^{1/4}} \cos\left(\frac{\pi}{4} - \frac{2x^{3/2}}{3}\right).$$

The fact that there are two different asymptotic expansions for different regimes is called the Stokes phenomenon. If we view $\mathrm{Ai}(z)$ as a function on the complex plane, these regions are separated by Stokes and anti-Stokes lines.

11.3 Matched Asymptotics

11.4 Multiple Scales

11.5 WKB Theory