Assorted notes

Sanjit Dandapanthula[1]
sanjitd@cmu.edu

November 30, 2025

[1]Carnegie Mellon University, Department of Statistics

Table of contents

Chapter 1

Probability theory

In this chapter, we review some fundamental theorems from measure-theoretic probability theory. The sources in this chapter are widely varied, and depend mostly on my style preference for each topic. I'll mostly assume the basics of measure-theoretic probability and only cover theorems that I find interesting or useful. I'll also state most of the theorems for real-valued random variables, but many of them can be easily generalized to $\mathbb{R}^d$-valued random variables (or even more general spaces).

1.1 Basic theorems

We start with Skorokhod's representation theorem, which says that random variables converging in distribution can be coupled in such a way that they converge almost surely.

Definition 1.1.1 (Convergence in distribution)**.** We say that $X_n \xrightarrow{d} X$ (X_n *converges in distribution* to X) if $F_{X_n}(x) \to F_X(x)$ for all continuity points x of F. Probabilists also call this *weak convergence*, but this actually corresponds to weak-* convergence in the sense of functional analysis.

Theorem 1.1.1 (Skorokhod representation)**.** *Let $X_n \xrightarrow{d} X$. Then, there exists a probability space $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$ and random variables Y_n and Y on this space such that $Y_n \overset{d}{=} X_n$, $Y \overset{d}{=} X$, and $Y_n \xrightarrow{a.s.} Y$.*

Proof. The proof is easy; define the quantile transformation $F^{-1}(x) = \inf\{x \in \mathbb{R} : F(x) \geq x\}$, draw a uniformly random variable $U \sim \text{Unif}(0,1)$, and define $Y_n = F_{X_n}^{-1}(U)$ and $Y = F_X^{-1}(U)$. □

Next, we state the portmanteau lemma, which characterizes convergence in distribution in a variety of ways.

Theorem 1.1.2 (Portmanteau lemma)**.** *The following are equivalent:*

1. $X_n \xrightarrow{d} X$.

2. $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$ *for all* $f \in C_b(\mathbb{R})$.

3. $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$ *for all Lipschitz* $f \in C_b(\mathbb{R})$.

4. $\liminf_{n\to\infty} \mathbb{E}[f(X_n)] \geq \mathbb{E}[f(X)]$ *for all lower semi-continuous* f *taking values in* $[0, \infty]$.

5. $\liminf_{n\to\infty} \mathbb{P}(X_n \in G) \geq \mathbb{P}(X \in G)$ *for all open* G.

6. $\mathbb{P}(X_n \in A) \to \mathbb{P}(X \in A)$ *for all* $A \in \mathcal{B}(\mathbb{R})$ *with* $\mathbb{P}(X \in \partial A) = 0$.

Proof. To show (1) implies (2), use Skorokhod's representation theorem (Theorem 1.1.1) and apply the dominated convergence theorem since f is bounded. Obviously, (2) implies (3). To show (3) implies (4), note that we can find Lipschitz functions $f_m \uparrow f$ pointwise since f is l.s.c. (Proposition 3.2.2); take the $\liminf$ in n and then the limit in m using the monotone convergence theorem. It's clear that (4) implies (5) since $\mathbf{1}_G$ is l.s.c. Then, (5) implies (6) by taking complements to get a similar statement for closed set and applying these results to $\text{int}(A)$ and $\overline{A}$. (6) implies (1) is obvious. □

Continuous mappings preserve convergence in distribution, in probability, and almost surely.

Theorem 1.1.3 (Continuous mapping)**.** *If* $X_n \to X$ *in distribution, in probability, or almost surely, then* $g(X_n) \to g(X)$ *in the same sense for any continuous function* g.

Proof. The proof is obvious for almost sure convergence and convergence in probability, and for convergence in distribution, use (2) in the portmanteau lemma (Theorem 1.1.2). □

Now, we state Prokhorov's theorem, which is a version of Bolzano-Weierstrass for probability measures.

Definition 1.1.2 (Tightness)**.** A set $\{X_\alpha\}_{\alpha\in A}$ of random variables is *tight* if for all $\epsilon > 0$ there exists a compact set K such that $\sup_{\alpha\in A} \mathbb{P}(X_\alpha \notin K) \leq \epsilon$.

Theorem 1.1.4 (Prokhorov)**.** *A set* $\{X_\alpha\}_{\alpha\in A}$ *of random variables is tight if and only if every sequence has a weakly convergent subsequence.*

Note that Prokhorov's theorem generalizes to any *Polish space* (complete separable metric space).

Proof. For the forward direction, pick a sequence $(X_n)_{n=1}^\infty \subseteq \{X_\alpha\}_{\alpha\in A}$ and enumerate $\mathbb{Q}$. By Cantor's diagonalization argument, extract a subsequence $(F_{X_{n_k}})_{k=1}^\infty$ such that $F_{X_{n_k}} \to F$ on all rationals. Define $F(x) = \inf_{q>x} F(q)$; it is easy to show that F is a cdf.

For the reverse direction, suppose every sequence in $\{X_\alpha\}_{\alpha\in A}$ has a weakly convergent subsequence but $\{X_\alpha\}_{\alpha\in A}$ isn't tight; in particular, there exists $\epsilon > 0$ such that for all compact K we have $\sup_{\alpha\in A} \mathbb{P}(X_\alpha \notin K) > \epsilon$. Then, pick $K_n = [-n, n]$ and choose X_n such that $\mathbb{P}(X_n \notin K_n) > \epsilon$. Extract a weakly convergent

subsequence $X_{n_k} \xrightarrow{d} X$ and pick $M > 0$ so that $\mathbb{P}(X \in (-M, M)) > 1 - \epsilon$. But we know by (5) in the portmanteau lemma (Theorem 1.1.2) that

$$1 - \epsilon < \mathbb{P}(X \in (-M, M)) \leq \liminf_{n \to \infty} \mathbb{P}(X_{n_k} \in (-M, M)) \leq 1 - \epsilon.$$

giving a contradiction. □

In the context of the Banach-Alaoglu theorem (Theorem 2.7.6) and the Riesz-Markov representation theorem, one can view Prokhorov's theorem as saying that a family of probability measures is relatively closed in the unit ball of the dual space of $C_0(\mathbb{R})$ (the space of finite Radon measures) if and only if it is tight and closed. Furthermore, any probability measure in a Polish space is tight.

Theorem 1.1.5 (Ulam's lemma)**.** *Any probability measure μ on a Polish space $\mathcal{X}$ is σ-finite and therefore tight.*

Proof. Let $\{x_n\}_{n=1}^{\infty}$ be a countable dense set and define the sets $K_m = \bigcup_{n=1}^{m} \overline{B(x_n, 1/m)}$ for each $m \in \mathbb{N}$, which are compact because they are each totally bounded and complete. Then $\mathcal{X} = \bigcup_{m=1}^{\infty} K_m$, so $\mathcal{X}$ is σ-finite. □

Note that a Borel probability measure on a Polish space is also automatically *regular*, meaning that the measure of any set can be approximated from above by open sets and from below by closed sets; this is a consequence of the Riesz-Markov representation theorem. Finally, we show that the L^p norms are ordered in probability spaces.

Proposition 1.1.6 (Ordering of L^p norms)**.** *If $1 \leq p < q \leq \infty$, then $\|X\|_p \leq \|X\|_q$ for all random variables X.*

Proof. Hölder's inequality with the conjugate exponents q/p and $q/(q-p)$ gives

$$\|X\|_p^p = \mathbb{E}[|X|^p] \leq \||X|^p\|_{q/p} \, \|1\|_{q/(q-p)} = \|X\|_q^p.$$

This only works in a probability space because we needed that $\|1\|_{q/(q-p)} = 1$. □

1.2 Lévy's continuity theorem and inversion formula

In this section, we give fundamental results about characteristic functions.

Definition 1.2.1 (Characteristic function)**.** The *characteristic function* of a random variable X is $\varphi_X(t) = \mathbb{E}[e^{itX}]$.

We start by proving the Lévy continuity theorem, which characterizes convergence in distribution in terms of the convergence of characteristic functions.

Theorem 1.2.1 (Lévy continuity)**.** *Suppose that $\varphi_{X_n} \to \varphi$ converges pointwise for all $t \in \mathbb{R}$. Then, there exists a random variable X such that $X_n \xrightarrow{d} X$ if and only if φ is continuous at 0.*

Proof. The forward direction is immediate from the dominated convergence theorem, so we'll focus on the reverse direction. If $(X_n)_{n=1}^{\infty}$ was tight, then every subsequence would have a further subsequence converging in distribution to some random variable X. But then by the portmanteau lemma, the characteristic functions would converge along all subsequences to φ_X and the result would immediately follow. So Lévy's continuity theorem is really about how the continuity of φ at 0 implies tightness.

The key idea in this proof is to use the Lebesgue differentiation theorem. Note that $\varphi_{X_n}(0) = 1$. Therefore, we study the following integral using the Lebesgue differentiation theorem and Fubini's theorem:

$$\begin{aligned}
0 &= \lim_{\delta \downarrow 0} \frac{1}{2\delta} \int_{-\delta}^{\delta} \Re(1 - \varphi_{X_n}(t))\, dt \\
&= \lim_{\delta \downarrow 0} \mathbb{E}\left[\frac{1}{2\delta} \int_{-\delta}^{\delta} (1 - \cos(tX_n))\, dt\right] \\
&= \lim_{\delta \downarrow 0} \mathbb{E}\left[1 - \frac{\sin(\delta X_n)}{\delta X_n}\right] \\
&\geq \lim_{\delta \downarrow 0} \mathbb{E}\left[\frac{\mathbf{1}_{|\delta X_n| \geq \pi}}{2}\right] \\
&= \lim_{\delta \downarrow 0} \frac{1}{2}\, \mathbb{P}(|X_n| \geq \pi/\delta).
\end{aligned}$$

The inequality follows because when $x \geq \pi$, we have

$$\left|\frac{\sin(x)}{x}\right| \leq \frac{1}{\pi} \implies 1 - \frac{\sin(x)}{x} \geq 1 - \frac{1}{\pi} \geq \frac{1}{2}.$$

Picking δ small, we deduce that $(X_n)_{n=1}^{\infty}$ is tight as desired. □

In fact, the proof shows that the continuity of $\Re(\varphi)$ at 0 is the only necessary condition for convergence in Lévy's continuity theorem. We can also use the Lévy continuity theorem to immediately get a converse to the weak law of large numbers.

Corollary 1.2.1.1 (Weak law of large numbers)**.** *If $X_1, \ldots, X_n$ are i.i.d. with characteristic function φ and mean μ, then $\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mu$ if and only if φ is differentiable at 0 and $\varphi'(0) = i\mu$.*

Another corollary of the Lévy continuity theorem is the Lindeberg-Lévy central limit theorem.

Theorem 1.2.2 (Lindeberg-Lévy CLT)**.** *If $X_1, \ldots, X_n$ are i.i.d. with mean μ and variance σ^2, then $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i \xrightarrow{d} \mathcal{N}(0, \sigma^2)$.*

An alternative proof of the Lindeberg-Lévy CLT is to use the idea of a *Lindeberg exchange*; we start with Rademacher random variables in the sum and exchange them one-by-one for the X_i. Next, we state Lévy's inversion formula, which shows that random variables are uniquely determined by their characteristic functions.

Theorem 1.2.3 (Lévy inversion formula)**.** *For continuity points $a < b$ of F_X, we have*

$$F_X(b) - F_X(a) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{-ita} - e^{-itb}}{it} \varphi_X(t)\, dt.$$

Proof. The characteristic function φ_X is essentially the Fourier-Stieltjes transform of the distribution of X (up to sign changes), so the result follows from the general inversion formula from Fourier analysis. □

1.3 Kolmogorov's three-series lemma and the SLLN

In this section, we give a proof of the strong law of large numbers using Kolmogorov's three-series lemma, which fully characterizes almost sure convergence of a series of random variables. We start by proving Kolmogorov's maximal inequality.

Theorem 1.3.1 (Kolmogorov's inequality)**.** *Suppose $X_1, \dots, X_n$ are independent with $\mathbb{E}[X_i] = 0$ and $\mathbb{E}[X_i^2] = \sigma_i^2$ and let $S_k = X_1 + \cdots + X_k$. Then, for all $t > 0$ we have*

$$\mathbb{P}\left(\max_{1 \le k \le n} |S_k| \ge t \right) \le \frac{1}{t^2} \sum_{i=1}^{n} \sigma_i^2.$$

Proof. Define the stopping time $\tau = \inf\{1 \le k \le n : |S_k| \ge t\} \wedge n$. Then, we compute the following L^2 estimate, since $\mathbf{1}_{i \le \tau}$ is independent of X_i:

$$\mathbb{E}[S_\tau^2] = \sum_{i=1}^{n} \mathbb{E}[X_i^2\, \mathbf{1}_{i \le \tau}] = \sum_{i=1}^{n} \mathbb{E}[X_i^2]\, \mathbb{P}(i \le \tau) \le \sum_{i=1}^{n} \sigma_i^2.$$

Now the result follows from Chebyshev's inequality applied to S_τ. □

From Kolmogorov's inequality we obtain the Khintchine-Kolmogorov two-series lemma.

Theorem 1.3.2 (Khintchine-Kolmogorov two-series lemma)**.** *Let $(X_n)_{n=1}^{\infty}$ be independent with $\mathbb{E}[X_n] = \mu_n$ and $\mathbb{E}[X_n^2] = \sigma_n^2$ and let $S_k = X_1 + \cdots + X_k$. Then, if $\sum_{n=1}^{\infty} \mu_n < \infty$ and $\sum_{n=1}^{\infty} \sigma_n^2 < \infty$, then S_k converges almost surely and in L^2.*

Proof. Suppose without loss of generality that $\mu_n = 0$ for all $n \in \mathbb{N}$ (consider $X_n \mapsto X_n - \mu_n$). We know that

$$\mathbb{P}\left(S_k \text{ converges}\right) = \mathbb{P}\left(\bigcap_{k=1}^{\infty} \bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} \{|S_n - S_m| \le 1/k\} \right).$$

By Kolmogorov's inequality (Theorem 1.3.1), we have

$$\mathbb{P}\left(\max_{m\le i\le n}|S_i - S_m| \ge 1/k\right) \le k^2 \sum_{i=m}^{n} \sigma_i^2 \le k^2 \sum_{i=m}^{\infty} \sigma_i^2.$$

Carefully letting $n \to \infty$ and then $m \to \infty$, it follows that for any $k \in \mathbb{N}$ we have

$$\mathbb{P}\left(\bigcap_{m=1}^{\infty}\bigcup_{n=m}^{\infty}\{|S_n - S_m| \ge 1/k\}\right) = 0 \implies \mathbb{P}\left(\bigcup_{m=1}^{\infty}\bigcap_{n=m}^{\infty}\{|S_n - S_m| \le 1/k\}\right) = 1.$$

Almost sure convergence now follows by taking an intersection over all $k \in \mathbb{N}$; denote the limit as S. By Fatou's lemma, we have

$$\mathbb{E}[(S_n - S)^2] = \mathbb{E}\left[\liminf_{k\to\infty}(S_n - S_k)^2\right] \le \liminf_{k\to\infty}\mathbb{E}[(S_n - S_k)^2] = \sum_{i=n}^{\infty}\sigma_i^2,$$

which tends to zero as $n \to \infty$; this gives the L^2 convergence. □

Now, we use the two-series lemma to prove the three-series lemma.

Theorem 1.3.3 (Kolmogorov's three-series lemma)**.** *Let $\lambda > 0$ be any constant (for instance, one could pick $\lambda = 1$). Suppose $(X_n)_{n=1}^{\infty}$ are independent and define $X_n^{(\lambda)} := X_n \mathbf{1}_{|X_n|\le\lambda}$. Then, $\sum_{n=1}^{\infty} X_n$ converges almost surely if and only if the following three series converge:*

1. $\sum_{n=1}^{\infty}\mathbb{P}(|X_n| \ge \lambda)$
2. $\sum_{n=1}^{\infty}\mathbb{E}[X_n^{(\lambda)}]$
3. $\sum_{n=1}^{\infty}\mathrm{Var}(X_n^{(\lambda)})$.

Proof. First, we'll show the reverse direction. By the first Borel-Cantelli lemma and convergence of series (1), we have that the probability that $\{|X_n| \ge \lambda\}$ infinitely often is zero; in particular, $X_n = Y_n$ eventually. By the Khintchine-Kolmogorov two-series lemma (Theorem 1.3.2) and convergence of series (2) and (3), we have that Y_n (and therefore X_n) converges almost surely.

Next, we show the forward direction. If series (1) didn't converge, then by the second Borel-Cantelli lemma (and independence of the X_n), we would have $|X_n| \ge \lambda$ infinitely often with probability 1, and $\sum_{n=1}^{\infty} X_n$ would diverge almost surely. Note that convergence of series (3) implies convergence of series (2): $\sum_{n=1}^{\infty}(X_n^{(\lambda)} - \mathbb{E}[X_n^{(\lambda)}])$ converges almost surely by the two-series lemma (Theorem 1.3.2) and for $\sum_{n=1}^{\infty} X_n$ to converge almost surely, we need $\sum_{n=1}^{\infty}\mathbb{E}[X_n^{(\lambda)}]$ to converge. If series (3) didn't converge, then by a slight generalization of the Lindeberg-Lévy CLT, we would find that

$$\frac{1}{\sqrt{\sum_{i=1}^{n}\mathrm{Var}(X_i^{(\lambda)})}}\sum_{i=1}^{n}(X_i^{(\lambda)} - \mathbb{E}[X_i^{(\lambda)}]) \xrightarrow{d} \mathcal{N}(0,1).$$

This simple generalization follows from the Lévy continuity theorem (Theorem 1.2.1) and holds whenever the sum of the variances of i.i.d. random variables diverges. Note that $\sum_{i=1}^n X_i^{(\lambda)}$ converges almost surely whenever $\sum_{i=1}^n X_i$ converges almost surely since the summands are almost surely eventually equal. But then this means that

$$\frac{1}{\sqrt{\sum_{i=1}^n \operatorname{Var}(X_i^{(\lambda)})}} \sum_{i=1}^n X_i^{(\lambda)} \xrightarrow{p} 0$$

since the denominator converges to 0. This is a contradiction (since $\mathcal{N}(0,1)$ is nondegenerate), so series (3) must converge. □

Now, we need one last lemma before we can prove the strong law of large numbers.

Lemma 1.3.4 (Kronecker's lemma)**.** *If $a_n \uparrow \infty$ and $\sum_{n=1}^\infty b_n/a_n$ converges then $\frac{1}{a_n}\sum_{m=1}^n b_m \to 0$.*

Proof. The proof follows from *summation by parts*. □

Kronecker's lemma is useful because it changes questions about averages into questions about sums, and we can use the three-series lemma to handle sums.

Theorem 1.3.5 (Strong law of large numbers)**.** *Let $(X_n)_{n=1}^\infty$ be i.i.d. with $\mathbb{E}[X_n] = \mu$. Then, we have $\frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{a.s.} \mu$.*

Proof. Apply Kolmogorov's three series lemma (Theorem 1.3.3) with $\tilde{X}_n = X_n/n$. Then, use Kronecker's lemma to upgrade almost sure convergence of the series to almost sure convergence of the average. □

We chose to prove the SLLN here using the Kolmogorov three series lemma in order to learn how to deal with almost sure convergence of series. However, there is a slightly simpler proof of the SLLN using Riesz's lemma of the rising sun, which generalizes to prove the strong ergodic theorem.

Alternate proof of Theorem 1.3.5. The intuition for this proof comes from thinking of S_n as a stochastic process which is asymptotically linear with slope μ. Let $E_\alpha = \{\limsup_{n\to\infty} S_n/n > \alpha\}$ and $F_\beta = \{\liminf_{n\to\infty} S_n/n < \beta\}$, with $G = \{\lim_{n\to\infty} S_n/n \text{ exists and is equal to } \mu\}$. In particular, we have

$$G^c = \bigcup_{m=1}^\infty (E_{\mu+1/m} \cup F_{\mu-1/m}),$$

so it suffices to show that $\mathbb{P}(E_\alpha) = 0$ for $\alpha > \mu$. Since E_α is tail-measurable, its probability is either 0 or 1 by Kolmogorov's 0-1 law. Suppose that there is $\alpha \in \mathbb{R}$ such that $\mathbb{P}(E_\alpha) = 1$; we want to show that this forces $\alpha \le \mu$. By stationarity of $(X_n)_{n=1}^\infty$, we have

$$\mathbb{P}\left(\sup_{n>k} \frac{S_n - S_k}{n-k} > \alpha\right) = 1.$$

Now, we need a lemma before we continue the proof of the SLLN.

Lemma 1.3.6 (Riesz's lemma of the rising sun)**.** *Fix $\{X_1, \dots, X_M\}$ and $n \in [M]$. Then, we say $n \in L$ if $\max_{n<t\le M} \frac{S_t - S_n}{t-n} \le \alpha$ and $n \in D$ otherwise. Intuitively, imagine we're plotting S_n over time and there is light shining from the right with slope α. Here, L is the set of points which are in the light and D is the set of points in the dark. Then, we have*

$$\frac{1}{|D|} \sum_{n-1 \in D} X_n \ge \alpha.$$

This means that jumps after dark points contribute an average of at least α to the sum.

Proof. The trick of this proof is mostly visual. We can decompose D into disjoint intervals and it suffices to obtain a bound on each dark interval $I = [i, j]$. Then, it is easy to show that S_{j+1} keeps all of I in the dark; in particular, this means that

$$S_{j+1} - S_i = \sum_{n=i+1}^{j+1} X_n \ge \alpha(j-i) = \alpha|I|. \qquad \square$$

We now complete the proof of the SLLN using Riesz's lemma of the rising sun. Fix $\epsilon > 0$ and choose N large enough that $\mathbb{P}\left(\max_{1\le n\le N} S_n/n > \alpha\right) > 1 - \epsilon$ and

$$\mathbb{E}\left[|X_1| \, \mathbf{1}_{\max_{1\le n\le N} S_n/n \le \alpha}\right] < \epsilon,$$

by the dominated convergence theorem. Then if $M \ge N$, we know that

$$\mu M = \mathbb{E}[S_M] = \mathbb{E}[X_1] + \mathbb{E}\left[\sum_{n-1\in D} X_n\right] + \mathbb{E}\left[\sum_{\substack{n-1\in L \\ n\le M-N}} X_n\right] + \mathbb{E}\left[\sum_{\substack{n-1\in L \\ n> M-N}} X_n\right].$$

To deal with the first term, note that $\mathbb{P}(n \in D) > 1 - \epsilon$ for $n \le M - N$. By Riesz's rising sun lemma, we have the estimate

$$\mathbb{E}\left[\sum_{n-1\in D} X_n\right] \ge \alpha \mathbb{E}[|D|] = \alpha \sum_{n=1}^{M} \mathbb{P}(n \in D) \ge \alpha(1-\epsilon)(M-N).$$

We also obtain

$$\mathbb{E}\left[\sum_{n=1}^{M-N} X_n \, \mathbf{1}_{n-1\in L}\right] \ge -\sum_{n=1}^{M-N} \mathbb{E}[|X_n| \, \mathbf{1}_{n-1\in L}] \ge -\epsilon(M-N)$$

and

$$\mathbb{E}\left[\sum_{\substack{n-1\in L \\ n> M-N}} X_n\right] \ge -\sum_{n> M-N} \mathbb{E}[|X_n|] \ge -E[|X_1|]\, N.$$

Putting all the bounds together and taking $M \to \infty$, we obtain $\mu \ge \alpha(1-\epsilon) - \epsilon$. Letting $\epsilon \downarrow 0$ gives the result. $\square$

1.4 Disintegration of measure and regular conditional probability

In this section, we formalize the notion of conditioning on a set of measure zero using *regular conditional probabilities*. In the following, denote by $\mathcal{P}(\mathcal{X})$ the set of probability measures over $\mathcal{X}$. Intuitively, the disintegration of measure theorem says that we can write any measure as an average of measures for each point $x \in \mathcal{X}$, and this disintegration is essentially unique.

Theorem 1.4.1 (Disintegration of measure)**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces (complete and separable metric spaces), let $\mu \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$, and denote by $\mu_{\mathcal{X}}$ the marginal of μ on $\mathcal{X}$. Then, there exists a measurable map $x \to \pi_x$ from $\mathcal{X}$ into $\mathcal{P}(\mathcal{Y})$ (uniquely determined $\mu_{\mathcal{X}}$-almost everywhere) such that*

$$\mu = \int_{\mathcal{X}} (\delta_x \times \pi_x)\, d\mu_{\mathcal{X}}(x)$$

Here, the topology on $\mathcal{P}(\mathcal{Y})$ is the weak- topology induced by weak convergence.*

The proof is lengthy so we outline it here but leave some technical details to the reader.

Proof. In a Polish space, there is always a countable family of bounded continuous functions $\{f_n\}_{n=1}^{\infty}$ that separates point; the distance function is continuous, so we can consider the sequence of distance functions from elements of a countable dense subset by separability (potentially capping them off at rationals). For each $f \in C_b(\mathcal{Y})$, we can define a measure ν_f on $\mathcal{X}$ by

$$\nu_f(A) = \int_{A \times \mathcal{Y}} f(y)\, d\mu(x, y).$$

Since $\nu_f \ll \mu_{\mathcal{X}}$, we can define the Radon-Nikodym derivative $h_f : \mathcal{X} \to \mathbb{R}$ such that

$$\nu_f(A) = \int_A h_f(x)\, d\mu_{\mathcal{X}}(x).$$

Since the h_f are defined $\mu_{\mathcal{X}}$-almost everywhere, we can define the linear functional $L_x(f_n) = h_{f_n}(x)$ for each $n \in \mathbb{N}$, for $\mu_{\mathcal{X}}$-almost every $x \in \mathcal{X}$. Since L_x is linear and continuous, we can extend the definition to all of $C_b(\mathcal{Y})$ by the Stone-Weierstrass theorem and the fact that $\{f_n\}_{n=1}^{\infty}$ separates points. In particular, linear combinations of the f_n are dense in $C_b(\mathcal{Y})$ with respect to uniform convergence on compact sets and we can easily define L_x on these linear combinations. Note that $L_x(f) \geq 0$ when $f \geq 0$ and $L_x(1) = 1$, so by the Riesz-Markov representation theorem, there exists a unique probability measure π_x on $\mathcal{Y}$ such that

$$h_f(x) = L_x(f) = \int_{\mathcal{Y}} f(y)\, d\pi_x(y).$$

It can be verified that any bounded continuous function $f : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$, the map $x \mapsto h_f(x)$ is measurable so $x \mapsto \pi_x$ is measurable as well. It's now clear that $\mu_{\mathcal{X}}$-almost every $x \in \mathcal{X}$, the measure π_x concentrates

on the *fiber* $\pi_{\mathcal{X}}^{-1}(x)$ (where $\pi_{\mathcal{X}}$ is the projection from $\mathcal{X} \times \mathcal{Y}$ into the first coordinate). For bounded measurable functions of the form $g(x)f(y)$, one can verify that

$$\int_{\mathcal{X}\times\mathcal{Y}} g(x)f(y)\,d\mu(x,y) = \int_{\mathcal{X}} g(x)\left(\int_{\mathcal{Y}} f(y)\,d\pi_x(y)\right)d\mu_{\mathcal{X}}(x),$$

so we can extend this to all bounded measurable functions by Dynkin's π-λ theorem, showing that

$$\mu = \int_{\mathcal{X}} (\delta_x \times \pi_x)\,d\mu_{\mathcal{X}}(x).$$

Essential uniqueness follows immediately because for any bounded measurable function f,

$$\int_{\mathcal{X}}\int_{\mathcal{Y}} f(x,y)\,d\pi_x(y)\,d\mu_{\mathcal{X}}(x) = \int_{\mathcal{X}\times\mathcal{Y}} f(x,y)\,d\mu(x,y)$$

forces π_x to be unique $\mu_{\mathcal{X}}$-almost everywhere. □

The function $(x, A) \mapsto \pi_x(A)$ is called a *regular conditional probability* and is denoted by $\mathbb{P}(A \mid X = x)$. The disintegration of measure theorem shows that we can think of π_x as a conditional probability given x, but that such a thing is only uniquely defined for $\mu_{\mathcal{X}}$-almost every $x \in \mathcal{X}$. In particular, if we let $(X, Y) \sim \mu$ we can think of π_x as a conditional probability given x in the sense that

$$\int_A\int_{\mathcal{Y}} f(y)\,d\pi_x(y)\,d\mu_{\mathcal{X}}(x) = \int_A \mathbb{E}[f(Y) \mid X = x]\,d\mu_{\mathcal{X}}(x) = \int_{A\times\mathcal{Y}} f(y)\,d\mu(x,y)$$

for any bounded measurable function $f : \mathcal{Y} \to \mathbb{R}$ and any measurable set $A \in \mathcal{X}$.

Chapter 2

Functional analysis

The study of functional analysis is a generalization of linear algebra to infinite-dimensional spaces. The key algebraic structures are normed spaces, inner product spaces, Banach spaces, and Hilbert spaces. In this chapter, we cover several important theorems of functional analysis to provide intuition for these spaces. We assume an undergraduate background in linear algebra and real analysis. A lot of this chapter is taken from Kreyszig (1991), although some proofs come from other sources and the content has been reorganized so that it makes the most sense to me.

2.1 Key algebraic structures

First, we begin with a few basic results about normed spaces. Note that all theorems for arbitrary normed spaces or inner product spaces can be specialized to Banach spaces or Hilbert spaces respectively.

2.1.1 Normed spaces and Banach spaces

Definition 2.1.1 (Banach space)**.** A *Banach space* is a complete normed space.

Here are a few examples of normed spaces and Banach spaces to keep in mind.

- $\mathbb{Q}$ with the absolute value is not a Banach space.
- $C([0,1])$ with the 2-norm is not a Banach space.
- $\mathbb{R}^n$ with the Euclidean norm is a Banach space.
- $C([0,1])$ with the sup-norm is a Banach space.
- $L^p(\mathbb{R})$ with the L^p norm is a Banach space.
- $L^p(\mathbb{N})$ (also called ℓ^p) with the p-norm is a Banach space.

We have two notions of basis in infinite dimensions.

Definition 2.1.2 (Hamel basis)**.** A *Hamel basis* for a vector space X is a set of linearly independent vectors that span X.

Note that every vector space has a Hamel basis due to Zorn's lemma (Axiom A.1.1). However, Hamel bases are not very useful in infinite dimensions because they are not unique and may not be countable.

Definition 2.1.3 (Schauder basis)**.** A *Schauder basis* for a normed space X is a sequence of vectors $(e_n)_{n\in\mathbb{N}}$ such that every $x \in X$ can be written as a unique series $\sum_{n=1}^{\infty} \alpha_n e_n$ where the series converges in the norm of X.

A Hamel basis corresponds to a basis in the sense of linear algebra. A Schauder basis allows us to write any vector as an infinite series and not just a finite linear combination. As an example, note that the standard basis is a Schauder basis for $L^p(\mathbb{N})$ for $1 \leq p < \infty$ but not for $L^\infty(\mathbb{N})$ (e.g., take $x = (1, 1, 1, \dots)$).

Proposition 2.1.1. *If a normed space X has a Schauder basis, then it is separable.*

Proof. Pick rational coefficients in the series representation. □

Definition 2.1.4 (Absolute convergence)**.** A series $\sum_{n=1}^{\infty} x_n$ (where the x_n are in a normed space) is said to *converge absolutely* if the series $\sum_{n=1}^{\infty} \|x_n\|$ converges.

The key lemma that we use to get basic results in finite-dimensional normed spaces is the following.

Lemma 2.1.2 (Quantitative independence bound)**.** *If $(x_i)_{i=1}^n$ is independent, then there exists $c > 0$ such that for all scalars $(\alpha_i)_{i=1}^n$, we have*

$$\left\| \sum_{i=1}^{n} \alpha_i x_i \right\| \geq c \sum_{i=1}^{n} |\alpha_i|.$$

Proof. Suppose w.l.o.g. that $\sum_{i=1}^{n} |\alpha_i| = 1$ and proceed by contradiction. Now pick a sequence of vectors $\alpha^{(k)}$ such that

$$\left\| \sum_{i=1}^{n} \alpha_i^{(k)} x_i \right\| \leq \frac{1}{k}.$$

The $\alpha^{(k)}$ are bounded so extract a convergent subsequence by Bolzano-Weierstrass and deduce a contradiction by independence and continuity of the norm. □

There are a few immediate consequences of the quantitative independence bound for finite-dimensional normed spaces. The proofs are easy so we omit them.

Corollary 2.1.2.1. *Finite-dimensional normed spaces are complete (and therefore closed).*

Corollary 2.1.2.2. *All norms on finite-dimensional spaces are equivalent (i.e., there are constants* $0 < c_1 < c_2$ *such that* $c_1\|x\|_0 \leq \|x\|_1 \leq c_2\|x\|_0$ *for all* x*).*

Using only the key lemma, we can in fact show the Riesz's lemma.

Lemma 2.1.3 (Riesz)**.** *If* Y *is a strict closed subspace of a normed space* X*, then for all* $\theta \in (0,1)$ *there exists* $u \in X$ *with* $\|u\| = 1$ *such that*

$$\inf_{y \in Y} \|u - y\| \geq \theta.$$

Riesz's lemma says that you can find unit vectors far from a closed space in any normed space. The closedness assumption is important because c_0 (the set of sequences with finitely many nonzero terms) is a non-closed subspace of ℓ^2 but there is no unit vector in ℓ^2 far from c_0.

Proof. Start by picking a nonzero $x_0 \in X \setminus Y$. Then $a = \inf_{y \in Y} \|x_0 - y\| > 0$ since Y is closed. Find $y_0 \in Y$ which is pretty close to x_0:

$$a \leq \|x_0 - y_0\| \leq \frac{a}{\theta}.$$

Letting $u = (x_0 - y_0)/\|x_0 - y_0\|$, it's easy to see that u satisfies the conditions in Riesz's lemma. □

Using Riesz's lemma, we can prove the following fundamental result.

Theorem 2.1.4 (Compactness of the unit ball)**.** *The closed unit ball in a normed space is compact if and only if the space is finite-dimensional.*

Proof. The reverse direction is obvious by Heine-Borel. For the forward direction, one can inductively extract a sequence of points in the unit ball without any convergent subsequence using Riesz's lemma (Lemma 2.1.3). □

2.1.2 Linear operators

Definition 2.1.5 (Bounded operator)**.** A linear operator $T : X \to Y$ between normed spaces is *bounded* if

$$\|T\| := \sup_{x \neq 0} \frac{\|T(x)\|}{\|x\|} < \infty.$$

Proposition 2.1.5 (Boundedness is equivalent to continuity)**.** *A linear operator* $T : X \to Y$ *between normed spaces is bounded if and only if it is continuous (or even continuous at a point).*

Proof. Assuming boundedness, $\|T(x - y)\| \leq c\|x - y\|$ implies continuity. For the converse, suppose T is continuous at $x_0 \in X$. Then, for all $\epsilon > 0$, there exists $\delta > 0$ such that $\|x - x_0\| \leq \delta$ implies $\|T(x) - T(x_0)\| \leq \epsilon$. The rest of the proof proceeds by picking a small vector pointing from x_0 in the

direction of x, which is a common technique in functional analysis. For any $x \neq 0 \in X$ associate with it the vector $\tilde{x} = x + \delta x/\|x\|$ so that $\|\tilde{x} - x_0\| = \delta \implies (\delta/\|x\|)\, T(x) = \|T(\tilde{x}) - T(x_0)\| \leq \epsilon$. Rearrange to obtain the result. □

The set of bounded linear operators between two normed spaces X and Y (denoted $B(X, Y)$) is itself a normed space with the operator norm.

Proposition 2.1.6. *$B(X, Y)$ is a normed space with the operator norm and is complete if Y is complete.*

Note that only the output space needs to be complete for $B(X, Y)$ to be complete.

Proof. It's obvious that $B(X, Y)$ is a normed space. If $(T_n)_{n=1}^{\infty}$ is Cauchy then $(T_n x)_{n=1}^{\infty}$ is Cauchy for all x and converges in Y to y_x. Define $Tx = y_x$ and show it's in $B(X, Y)$. □

2.1.3 Hilbert spaces

Definition 2.1.6 (Hilbert space)**.** A *Hilbert space* is a complete inner product space.

Easy algebra gives the following lemma.

Lemma 2.1.7 (Parallelogram equation)**.** *For all x, y in any inner product space X, we have*

$$\|x + y\|^2 + \|x - y\|^2 = 2(\|x\|^2 + \|y\|^2).$$

Notably, not all norms come from inner products (e.g., the sup-norm on $C([0, 1])$ or the p-norm on ℓ^p spaces). The proofs all rely on the fact that given a norm in an inner product space, we can recover the inner product.

Lemma 2.1.8 (Polarization identity)**.** *If X is an inner product space over $\mathbb{R}$, then*

$$\langle x, y \rangle = \frac{1}{4} \left(\|x + y\|^2 - \|x - y\|^2 \right).$$

If X is an inner product space over $\mathbb{C}$, then

$$\langle x, y \rangle = \frac{1}{4} \left(\|x + y\|^2 - \|x - y\|^2 + i(\|x + iy\|^2 - \|x - iy\|^2) \right).$$

Interestingly, "rotations" don't exist in complex inner product spaces.

Proposition 2.1.9. *Let $Q : X \to X$ be a bounded linear operator.*

- *First, $\langle Qx, y \rangle = 0$ for all $x \in X$ and $y \in Y$ if and only if $Q = 0$.*
- *In fact, if X is complex and $Q : X \to X$ is a bounded linear operator then $\langle Qx, x \rangle = 0$ for all $x \in X$ suffices to force $Q = 0$.*

Proof. The first statement is obvious by setting $y = Qx$. For the second statement, notice that for all $\alpha \in \mathbb{C}$ and all $x, y \in X$, we have

$$\begin{aligned} 0 &= \langle Q(x + \alpha y), x + \alpha y\rangle \\ &= \langle Qx, x\rangle + \alpha\langle Qy, x\rangle + \overline{\alpha}\langle Qx, y\rangle + |\alpha|^2\langle Qy, y\rangle \\ &= \alpha\langle Qy, x\rangle + \overline{\alpha}\langle Qx, y\rangle. \end{aligned}$$

We conclude by setting $\alpha = i$ and $\alpha = 1$ and using the first statement. □

Now we prove the Cauchy-Schwarz inequality.

Theorem 2.1.10 (Cauchy-Schwarz)**.** *For all x, y in an inner product space X, we have*

$$|\langle x, y\rangle| \leq \|x\|\|y\|,$$

with equality if and only if x and y are linearly dependent.

Proof. We study $0 \leq \|x - \alpha y\|^2$ for $\alpha \in \mathbb{C}$. Standard geometry in $\mathbb{R}^n$ endowed with the usual inner product suggests that putting $\alpha = \frac{\langle x,y\rangle}{\|y\|^2}$ should intuitively minimize the expression on the RHS. Now just expand and rearrange. The derivation then shows that equality holds if and only if $x - \alpha y = 0$ or $y = 0$. □

The Cauchy-Schwarz inequality implies the triangle inequality and continuity of the inner product. One of the most important properties of a Hilbert space is the existence of orthogonal projections, which we now show.

Theorem 2.1.11 (Projection onto a convex set)**.** *If $M \subseteq X$ is convex and complete for any inner product space X, then for all x there exists a unique $y \in M$ such that*

$$\delta = \inf_{\tilde{y} \in M} \|x - \tilde{y}\| = \|x - y\|.$$

Proof. Pick $y_n \in M$ such that $\|x - y_n\| \downarrow \delta$. Then if $v_n := y_n - x$, the parallelogram equality gives

$$\|y_n - y_m\|^2 = 2\|v_n\|^2 + 2\|v_m\|^2 - \|v_n + v_m\|^2.$$

The first two terms go to δ, and convexity of M gives

$$\|v_n + v_m\| = \|y_n + y_m - 2x\| = 2\left\|\frac{y_n + y_m}{2} - x\right\| \geq 2\delta.$$

So $(y_n)_{n=1}^{\infty}$ is Cauchy, and completeness of M gives the result. Another application of the parallelogram equality shows that the projection is unique. □

Theorem 2.1.11 leads to the following theorem.

Theorem 2.1.12 (Orthogonal decomposition)**.** *If Y is a closed subspace of a Hilbert space X then $X = Y \oplus Y^{\perp}$ (everything in X is a unique sum of something in Y and something in $Y^{\perp}$).*

2.1.4 Orthogonality and Fourier series

Suppose we have an orthonormal set $(e_n)_{n=1}^{\infty}$ in an inner product space X.

Theorem 2.1.13 (Bessel's inequality)**.** *For all $x \in X$, we have*

$$\sum_{n=1}^{\infty} |\langle x, e_n \rangle|^2 \leq \|x\|^2.$$

Proof. Let $y = \sum_{k=1}^{n} \langle x, e_k \rangle e_k$ and let $z = x - y$ so that $\langle z, e_k \rangle = 0$ for all $1 \leq k \leq n$. So $\|x\|^2 = \|y\|^2 + \|z\|^2$ and the result follows from $\|z\| \geq 0$. □

The terms $\langle x, e_n \rangle$ are called the *Fourier coefficients* of x.

Theorem 2.1.14 (Convergence of orthonormal series)**.** *The series*

$$\sum_{n=1}^{\infty} \alpha_n e_n$$

converges if and only if

$$\sum_{n=1}^{\infty} |\alpha_n|^2 < \infty.$$

Also if the limit is x then $\alpha_n = \langle x, e_n \rangle$ are the Fourier coefficients.

Proof. Use the Pythagorean identity $\left\| \sum_{n=1}^{N} \alpha_n e_n \right\|^2 = \sum_{n=1}^{N} |\alpha_n|^2$ to show the first statement. The second statement follows from continuity of the inner product, since the partial sums are assumed to converge to x. □

Remarkably, there can only be countably many non-zero Fourier coefficients, no matter how large the orthonormal set is.

Theorem 2.1.15 (Riesz-Fischer)**.** *If $(e_i)_{i \in I}$ is an orthonormal set in a Hilbert space X, then for all $x \in X$ the set $\{i : \langle x, e_i \rangle \neq 0\}$ is countable.*

Proof. By Bessel's inequality, the set $\{i : |\langle x, e_i \rangle| \geq 1/k\}$ is finite for all $k \in \mathbb{N}$. □

A Schauder basis may not be spanning, so we define the following generalization of a spanning set.

Definition 2.1.7 (Total set)**.** A *total set* $M \subseteq X$ has $\overline{\text{span}(M)} = X$.

All nontrivial Hilbert spaces have total orthonormal sets by Gram-Schmidt and Zorn's lemma (Axiom A.1.1). All total orthonormal sets have the same cardinality, which is called the *Hilbert dimension* of X. The proof is set-theoretic and not difficult, so we omit it. We can also equivalently characterize total sets in Hilbert space.

Proposition 2.1.16. *If X is a Hilbert space, then M is total if and only if $x \perp M$ implies $x = 0$.*

Proof. The forward direction is true in any inner product space by continuity of the inner product. The reverse direction follows from the orthogonal decomposition $X = \overline{\text{span}(M)} \oplus \overline{\text{span}(M)}^{\perp}$. □

We have another characterization of totality for orthonormal sets.

Theorem 2.1.17 (Parseval)**.** *An orthonormal set M in a Hilbert space X is total if and only if for all $x \in X$, we have*

$$\|x\|^2 = \sum_{m \in M} |\langle x, m \rangle|^2.$$

This is called the Parseval relation.

Proof. For the reverse direction, if there existed a nonzero $x \perp M$ (applying Proposition 2.1.16), then x cannot satisfy the Parseval relation. For the forward direction, we can define $y = \sum_{m \in M} |\langle x, m \rangle|^2 m$ and show that $x - y \perp M$. Then $x - y = 0$ by Proposition 2.1.16. □

Orthonormal sets in a separable Hilbert space behave nicely.

Theorem 2.1.18. *If X is a separable Hilbert space then every orthonormal set is countable. Also, if there is a total orthonormal set then X is separable.*

Proof. For the first statement, let $(e_n)_{n=1}^{\infty}$ be an orthonormal set. Since $\|e_i - e_j\|^2 = 2$ for all $i \neq j$, neighborhoods of size $\sqrt{2}/2$ around the e_i are disjoint, but we need a countable dense set. The second statement follows from picking rational coefficients. □

Finally, we have the following result about isomorphism, which is analogous to the usual one for finite-dimensional spaces.

Theorem 2.1.19. *Hilbert spaces are isomorphic if and only if they have the same Hilbert dimension.*

Proof. The forward direction is obvious and the reverse direction follows from letting $T(x) = \sum_{k \in K} \langle x, e_k \rangle f_k$ where $(e_k)_{k \in K}$ and $(f_k)_{k \in K}$ are orthonormal bases for the two Hilbert spaces respectively. □

2.2 The dual space and the Riesz-Fréchet representation theorem

We call the space $X^* := B(X, \mathbb{F})$ the *dual space* of X where $\mathbb{F}$ is $\mathbb{R}$ or $\mathbb{C}$; Proposition 2.1.6 gives us some intuition about the dual space.

Proposition 2.2.1. *The dual space X^* is always a Banach space.*

Proof. $\mathbb{R}$ and $\mathbb{C}$ are complete, so use Proposition 2.1.6 (the space of bounded operators between X and Y is complete whenever Y is complete). □

Now, we have the Riesz-Fréchet representation theorem.

Theorem 2.2.2 (Riesz-Fréchet representation)**.** *If X is a Hilbert space, then for all $f \in X^*$ there exists a unique $y \in X$ such that*

$$f(x) = \langle x, y \rangle$$

for all $x \in X$ and $\|y\| = \|f\|$.

The Riesz-Fréchet representation theorem isn't surprising; $\mathbb{R}$ or $\mathbb{C}$ are one-dimensional Hilbert spaces over themselves, so f has to squash everything into one dimension. The only way to do that is by taking the inner product with a fixed vector.

Proof. The proof is trivial when $f = 0$ so assume $f \neq 0$. Now we want to show $\ker(f)^\perp$ is one-dimensional. For any nonzero $z_1, z_2 \in \ker(f)^\perp$, we have

$$f\left(z_1 - \frac{f(z_1)}{f(z_2)}\, z_2\right) = 0$$

So $z_1 - \frac{f(z_1)}{f(z_2)}\, z_2$ is in $\ker(f) \cap \ker(f)^\perp$, which shows that z_1 and z_2 are dependent. Now pick any $z_0 \in \ker(f)^\perp$ with norm 1 and define $y = \overline{f(z_0)}\, z_0$ so that $f(x) = \langle x, y \rangle$ for all x. Essentially, we are projecting onto $\operatorname{span}(z_0)$ and applying f. Uniqueness is immediate since $0 = \langle y_1 - y_2, y_1 \rangle - \langle y_1 - y_2, y_2 \rangle = \|y_1 - y_2\|$ implies $y_1 = y_2$ for alternatives y_1 and y_2. Finally, we have $\|y\|^2 = f(y) \leq \|f\|\|y\|$ so $\|y\| \leq \|f\|$, and the reverse inequality follows from Cauchy-Schwarz. □

2.3 The four pillars of functional analysis

In this section, we prove several important theorems in functional analysis, which are often called the "four pillars" of the subject. The Hahn-Banach theorem works in general normed spaces, but the other three theorems are specific to Banach spaces. The latter three will follow mostly from the Baire category theorem (Theorem A.2.1).

2.3.1 The Hahn-Banach theorem

The Hahn-Banach theorem allows us to extend bounded linear functionals from subspaces of any vector space to the entire space, and shows that the dual contains a lot of things. We begin with a related definition.

Definition 2.3.1 (Sublinear functional)**.** A *sublinear functional* on a vector space X is a function $p : X \to \mathbb{R}$ such that

$$p(\alpha x) = \alpha p(x)$$

for $\alpha \geq 0$ and

$$p(x + y) \leq p(x) + p(y)$$

for all $x, y \in X$.

An important example of a sublinear functional is the norm on a normed space.

Theorem 2.3.1 (Hahn-Banach)**.** *Suppose Z is a subspace of a vector space X and $p : X \to \mathbb{R}$ is a sublinear functional. If $f : Z \to \mathbb{R}$ is a linear functional such that $f(x) \leq p(x)$ for all $x \in Z$, then there exists a linear functional $\tilde{f} : X \to \mathbb{R}$ such that $\tilde{f}|_Z = f$ and $\tilde{f}(x) \leq p(x)$ for all $x \in X$.*

Proof. Apply Zorn's lemma (Axiom A.1.1) with $f_1 \leq f_2$ if f_2 extends f_1 on the poset $M = \{g : g \text{ linear}, g \text{ extends } f, g(x) \leq p(x)\}$. Now we have the existence of a maximal element $\tilde{f} : Y \to \mathbb{R}$. Suppose for a contradiction that $\tilde{f}$ isn't defined at some $z \in X$. Then, define $g(y + \alpha z) = \tilde{f}(y) + \alpha \tilde{f}(z)$ for all $y \in Y$ and $\alpha \in \mathbb{R}$; we need to pick $\tilde{f}(z)$ such that $g(x) \leq p(x)$. Rephrasing: for all $y \in Y$ and $\alpha > 0$, we need

$$\tilde{f}(y) + \alpha \tilde{f}(z) \leq p(y + \alpha z),$$
$$\tilde{f}(y) - \alpha \tilde{f}(z) \leq p(y - \alpha z).$$

Solving for the constraint on $\tilde{f}(z)$ and setting $y \mapsto y/\alpha$, we find:

$$\sup_{y \in Y} \{f(y) - p(y - z)\} \leq \tilde{f}(z) \leq \inf_{y \in Y} \{p(y + z) - f(y)\}.$$

As long as the left-hand side is not larger than the right-hand side, we will have a valid choice for $\tilde{f}(z)$. But we have

$$f(y_2) - p(y_2 - z) \leq p(y_1 + z) - f(y_1) \iff f(y_1 + y_2) \leq p(y_1 + z) + p(y_2 - z),$$

and the latter condition is implied by sublinearity of p. □

We can slightly generalize the Hahn-Banach theorem to complex vector spaces.

Theorem 2.3.2 (Generalized Hahn-Banach)**.** *Suppose Z is a subspace of a vector space X and $p : X \to \mathbb{R}$ is a subadditive functional; for all $x, y \in X$ and $\alpha \in \mathbb{C}$, we have*

$$p(\alpha x) = |\alpha| p(x)$$

and

$$p(x+y) \leq p(x) + p(y).$$

If $f : Z \to \mathbb{C}$ is a linear functional such that $|f(x)| \leq p(x)$ for all $x \in Z$, then there exists a linear functional $\tilde{f} : X \to \mathbb{C}$ such that $\tilde{f}|_Z = f$ and $|\tilde{f}(x)| \leq p(x)$ for all $x \in X$.

Proof. Use Hahn-Banach on the real and imaginary parts of f separately and set $\tilde{f}(x) = \tilde{f}_{\Re}(x) - i\tilde{f}_{\Im}(x)$. □

Note that Hahn-Banach is a generalization of the Riesz-Fréchet representation theorem (Theorem 2.2.2) to arbitrary vector spaces; if Z is a closed subspace of a Hilbert space X then the Riesz-Fréchet representation gives a linear extension $\tilde{f}(x) = \langle x, z \rangle$ for some $z \in X$. We now state the most important corollary of the Hahn-Banach theorem, which shows that there are lots of bounded linear functionals on a normed space.

Corollary 2.3.2.1. *If X is a normed space and $x_0 \in X$ is nonzero, then there exists a bounded linear functional $\tilde{f} : X \to \mathbb{F}$ such that $\tilde{f}(x_0) = \|x_0\|$ and $\|\tilde{f}\| = 1$.*

Proof. Apply the Hahn-Banach theorem with $Z = \text{span}(x_0)$ and $f(z) = \|z\|$ defined from Z to $\mathbb{F}$. □

This implies that the set of bounded linear functionals separates points in a normed space, and X^* therefore has a rich structure. This becomes useful later in the characterization of weak convergence and the adjoint operator. For applications of Hahn-Banach, see Sections 2.5 and 2.6.

2.3.2 The uniform boundedness principle (Banach-Steinhaus)

The uniform boundedness principle allows us to upgrade pointwise convergence of operators to uniform convergence, and follows from the Baire category theorem (Theorem A.2.1).

Theorem 2.3.3 (Banach-Steinhaus)**.** *If X is a Banach space and Y is a normed space, then for all $T_n : X \to Y$ such that*

$$\sup_{n \in \mathbb{N}} \|T_n x\| < \infty$$

for all $x \in X$, we have

$$\sup_{n \in \mathbb{N}} \|T_n\| < \infty.$$

This is sometimes called the uniform boundedness principle.

Proof. Define $A_k = \{x \in X : \|Tx\| \leq k\}$ so that A_k is closed and $X = \bigcup_{k=1}^{\infty} A_k$. By the Baire category theorem (since X is complete), there exists A_{k_0} containing a ball $B(x_0, r)$. We are almost done, since $B(x_0, r) - x_0$ contains vectors at a fixed length in all directions and $\|Tx\|$ is uniformly bounded in the ball.

To formalize this argument, for any $x \in X$ we can set $z = x_0 + \gamma x$ where $\gamma = r/(2\|x\|)$ so that $z \in B(x_0, r)$. So then we have the inequality

$$\|T_n x\| = \left\| \frac{1}{\gamma} T_n(z - x_0) \right\| \leq \frac{1}{\gamma}(\|T_n z\| + \|T_n x_0\|) \leq \frac{2k_0}{\gamma} = \frac{4k_0}{r}\|x\|.$$

This estimate shows that $\|T_n\|$ is uniformly bounded by $4k_0/r$, and we are done. □

A common use of the uniform boundedness principle is to construct operators which are pointwise bounded but not uniformly bounded, thereby showing that a space X is not complete. The uniform boundedness principle can also reveal interesting structure in many spaces, as we show in the following example.

Example 2.3.1 (Most continuous functions aren't locally differentiable)**.** There are lots of famous examples of continuous and nowhere differentiable functions (sample paths of Brownian motion, the Weierstrass function, etc.) but the uniform boundedness principle can be used to show that *most* continuous functions are actually nowhere locally differentiable. Consider $X = C([0, 1])$ endowed with the supremum norm and for each $x \in [0, 1]$ and $n \in \mathbb{N}$ we define the linear functional:

$$T_{n,x}(f) = \begin{cases} n\left(f\left(x + \frac{1}{n}\right) - f(x)\right) & x + \frac{1}{n} \leq 1 \\ n\left(f(x) - f\left(x - \frac{1}{n}\right)\right) & x + \frac{1}{n} > 1. \end{cases}$$

Essentially, $T_{n,x}$ are the difference quotients with $\Delta x = 1/n$; if f is differentiable at x, then $\sup_n |T_{n,x}(f)| < \infty$. Formally, define $D_x = \{f \in X : f \text{ is differentiable at } x\}$ so that $T_{n,x}$ is pointwise bounded on D_x. It is easy to see that $\sup_n \|T_{n,x}\| = \infty$ by constructing a sequence of functions that rise more and more sharply. In particular, by a slight generalization of the uniform boundedness principle to any nonmeager subset of a Banach space, we deduce that D_x must be meager. The set of anywhere locally differentiable functions is exactly the set of functions which are differentiable at any rational, so we deduce that the set of anywhere locally differentiable functions is meager in $C([0, 1])$.

For further example applications of the uniform boundedness principle, see Section 2.7.

2.3.3 The open mapping theorem

The open mapping theorem says that a surjective bounded linear operator between Banach spaces sends open sets to open sets. In particular, a bijective bounded linear operator between Banach spaces has a bounded inverse.

Definition 2.3.2 (Open mapping)**.** The map $T : X \to Y$ between metric spaces is an *open mapping* if $T(U)$ is open in Y whenever U is open in X.

Theorem 2.3.4 (Open mapping)**.** *If X and Y are Banach spaces and $T : X \to Y$ is a surjective bounded linear operator, then T is an open mapping.*

Proof. Let $B_r = B(0, r) \subseteq X$. It will suffice to show that $T(B_1)$ contains an open ball around 0 in Y. We know that

$$Y = \bigcup_{n=1}^{\infty} \overline{T(B_n)},$$

so the Baire category theorem implies that we can fit a ball in $\overline{T(B_n)}$ for some n. Shrinking this ball by a factor of n, we can find a ball $B(y_0, r_0) \subseteq \overline{T(B_1)}$. It is easy to show that $\overline{T(B_1)} - y_0 \subseteq \overline{T(B_2)}$, so that we can center the ball. Our goal is now to show that there is a ball in $T(B_1)$, thereby removing the closure.

Dilating the previous result, there exists $r > 0$ such that as long as $\|y\| < r/2^n$, then $y \in \overline{T(B_{2^{-n}})}$. We're going to show that $B_{r/2} \subseteq T(B_1)$, thereby removing the closure. As long as $\|y\| < r/2$, we can find $x_1 \in B_{1/2}$ such that $\|y - Tx_1\| < r/4$, and inductively, $x_n \in B_{2^{-n}}$ such that $\|y - \sum_{k=1}^{n} Tx_k\| < r/2^{n+1}$. Since X is complete, $\sum_{k=1}^{\infty} x_k$ converges to some $x \in X$. Now $\|x\| < 1$ and $y = Tx$, so the result follows. □

The open mapping theorem immediately has a useful corollary, called the bounded inverse theorem.

Corollary 2.3.4.1 (Bounded inverse)**.** *If X and Y are Banach spaces and $T : X \to Y$ is a bijective bounded linear operator, then T^{-1} is bounded.*

The bounded inverse theorem is used to show that solutions to linear equations in Banach spaces are continuous with respect to the desired output. For instance, if we are trying to solve $Tx = y$ for x and T is a bijective bounded linear operator between Banach spaces, then the solution $x = T^{-1}y$ is continuous in y; as an example, T might be a differential operator. Even if the map is not bijective, we have the following corollary.

Corollary 2.3.4.2. *Suppose X, Y, and T are as in the open mapping theorem (Theorem 2.3.4). Then there exists a constant $c > 0$ such that for any $y \in Y$ there exists $x \in X$ with $Tx = y$ and $\|x\| \leq c\|y\|$.*

Proof. By the open mapping theorem (Theorem 2.3.4), $T(B_1)$ contains an open ball $B(0, r)$. For any $y \in Y$, there exists $x \in B_1$ with $Tx = ry/(2\|y\|)$. In particular, we have $T(x(2\|y\|)/r) = y$ with $\|x(2\|y\|)/r\| \leq 2\|y\|/r$. Let $c = 2/r$. □

So the solution to a linear equation between Banach spaces is always bounded in norm by the desired output, as long as the linear map is surjective, which has important implications for numerical stability. If y has a small measurement error δ, then the solution x will have an error of at most $c\delta$, where c didn't depend on y.

2.3.4 The closed graph theorem

The closed graph theorem gives a sufficient condition for a linear operator between Banach spaces to be bounded, and follows from the open mapping theorem.

Definition 2.3.3 (Closed operator)**.** A linear operator $T : \mathrm{dom}(T) \to Y$ between normed spaces is *closed* if the graph of T is closed in $X \times Y$ under the norm $\|(x, y)\| = \|x\| + \|y\|$. Here, the graph of T is the set $\mathcal{G}(T) = \{(x, Tx) : x \in \mathrm{dom}(T)\}$.

We can characterize closed operators in another equivalent way.

Proposition 2.3.5. *A linear operator $T : \mathrm{dom}(T) \to Y$ between Banach spaces is closed if and only if for all sequences $(x_n)_{n=1}^{\infty}$ in $\mathrm{dom}(T)$ such that $x_n \to x$ and $Tx_n \to y$, we have $x \in \mathrm{dom}(T)$ and $Tx = y$.*

Example 2.3.2 (Differentiation operator)**.** The differentiation operator from $C^1([0, 1])$ to $C([0, 1])$ is closed but not bounded.

Now, we state the closed graph theorem, which gives sufficient conditions for a linear operator between Banach spaces to be bounded.

Theorem 2.3.6 (Closed graph)**.** *If X and Y are Banach spaces and $T : \mathrm{dom}(T) \to Y$ is a closed linear operator, then T is bounded.*

Proof. $X \times Y$ is complete, so consider the mapping $p(x, Tx) = x$ from $\mathcal{G}(T)$ to $\mathrm{dom}(T)$. This mapping is bounded and bijective, so by the open mapping theorem Theorem 2.3.4, p^{-1} is bounded (since $\mathcal{G}(T)$ and $\mathrm{dom}(T)$ are complete). In particular, this shows that T is bounded. □

We give sufficient conditions for the converse of the closed graph theorem to hold.

Proposition 2.3.7. *If $T : \mathrm{dom}(T) \to Y$ is a linear operator between any normed spaces then if $\mathrm{dom}(T)$ is complete and T is bounded, then T is closed. Also, if T is closed and Y is complete, then $\mathrm{dom}(T)$ is closed.*

Proof. The first statement is immediate from continuity of T. For the second statement, take a sequence $x_n \to x$ with $x_n \in \mathrm{dom}(T)$. Then, we have

$$\|T(x_n - x_m)\| \leq \|T\|\|x_n - x_m\| \to 0,$$

so $(Tx_n)_{n=1}^{\infty}$ is Cauchy. Now $Tx_n \to y$ for some $y \in Y$ by completeness of Y and $y = Tx$ by closedness of T. □

The closed graph theorem is used to show that an operator is bounded, since in many cases closedness easier to verify directly than boundedness.

2.4 Hilbert adjoint operators

First, we slightly generalize the Riesz-Fréchet representation theorem to sesquilinear forms.

Definition 2.4.1 (Sesquilinear form)**.** A *sesquilinear form* is $h : X \times Y \to \mathbb{F}$ which is linear in the first coordinate and conjugate linear in the second coordinate (where X and Y are vector fields over $\mathbb{F}$). Define the norm of h as

$$\|h\| = \sup_{\substack{x \neq 0 \\ y \neq 0}} \frac{|h(x,y)|}{\|x\|\|y\|}.$$

Theorem 2.4.1 (Generalized Riesz-Fréchet representation)**.** *If X and Y are Hilbert spaces and h is a bounded sesquilinear form on $X \times Y$ then we can write*

$$h(x,y) = \langle Sx, y \rangle,$$

where $S : X \to X$ is a unique linear operator with $\|S\| = \|h\|$.

Proof. $\overline{h(x,y)}$ is a bounded linear functional in y for fixed x, so the Riesz-Fréchet representation theorem gives $\overline{h(x,y)} = \langle y, z \rangle$ for some $z \in Y$; this means that $h(x,y) = \langle z, y \rangle$. Define $Sx = z$ for all x; it's easy to show S is linear and unique. By Cauchy-Schwarz, we have

$$\|h\| \leq \sup_{\|x\|=1} \frac{\|Sx\|}{\|x\|} = \|S\|.$$

Similarly, we get the reverse inequality by

$$\|h\| = \sup_{\substack{x \neq 0 \\ y \neq 0}} \frac{|h(x,y)|}{\|x\|\|y\|} \geq \sup_{\substack{x \neq 0 \\ Sx \neq 0}} \frac{|\langle Sx, Sx \rangle|}{\|x\|\|Sx\|} = \|S\|. \qquad \square$$

One main point of this generalization is to prove that the Hilbert adjoint operator exists.

Definition 2.4.2 (Hilbert adjoint)**.** If $T : X \to Y$ is bounded linear where X and Y are Hilbert spaces, then the *Hilbert adjoint* of T is the unique operator $T^* : Y \to X$ such that

$$\langle Tx, y \rangle = \langle x, T^* y \rangle$$

for all $x \in X$ and $y \in Y$.

We now verify that the Hilbert adjoint is well-defined.

Theorem 2.4.2. *T^* exists, is unique, and $\|T^*\| = \|T\|$.*

Proof. The function $h(x,y) = \langle y, Tx \rangle$ is a bounded sesquilinear form on $X \times Y$, so use the generalized Riesz-Fréchet representation theorem (Theorem 2.4.1). $\square$

The adjoint operator is a generalization of the conjugate transpose for finite-dimensional spaces. If $\langle x, y\rangle = x^\top \overline{y}$ then $\langle Bx, y\rangle = (Bx)^\top \overline{y} = x^\top \overline{B^* y}$, so $B^* = \overline{B}^\top$. The properties of the adjoint operator all follow immediately from the definition, except perhaps for the following.

Proposition 2.4.3. *The adjoint satisfies $\|T^*T\| = \|T\|^2$.*

Proof. We immediately have $\|T^*T\| \le \|T^*\|\|T\| = \|T\|^2$. To show the reverse inequality, we have for all $x \in X$ that

$$\|Tx\|^2 = \langle Tx, Tx\rangle = \langle x, T^*Tx\rangle \le \|x\|\|T^*Tx\|.$$

Taking the supremum over all x with $\|x\| = 1$ gives the result. □

We now have the following important definitions related to the adjoint operator.

Definition 2.4.3 (Self-adjoint operator)**.** An operator $T : X \to X$ is *self-adjoint* or *Hermitian* if $T = T^*$.

Definition 2.4.4 (Unitary operator)**.** An operator $T : X \to X$ is *unitary* if $T^* = T^{-1}$.

Self-adjoint operators are a generalization of conjugate symmetric matrices, and unitary operators are a generalization of orthogonal matrices. The adjoint operator has a few nice properties.

Proposition 2.4.4. *If X is a complex Hilbert space then T is self-adjoint if and only if $\langle Tx, x\rangle$ is real for all x.*

Proof. This statement is obvious. □

Proposition 2.4.5. *The limit of bounded self-adjoint linear operators is bounded and self-adjoint.*

Proof. We have by the triangle inequality

$$\|T - T^*\| \le \|T - T_n\| + \|T_n - T_n^*\| + \|T_n^* - T^*\| = 2\|T - T_n\| \to 0.$$ □

Unitary operators behave nicely in Hilbert spaces too.

Proposition 2.4.6. *Unitary operators are isometries ($\|Ux\| = \|x\|$), and a bounded linear operator on a complex Hilbert space is unitary if and only if it is isometric and surjective.*

Proof. The first statement and the forward implication of the second statement are easy to show. For the reverse implication, isometries are injective so T is bijective. Now $\langle x, x\rangle = \langle T^*Tx, x\rangle$ implies that $\langle (T^*T - I)x, x\rangle = 0$ for all x; conclude by Proposition 2.1.9 that $T^*T = I$. □

2.5 The adjoint operator

Suppose we have a linear functional g on Y. Then, if $T \in B(X, Y)$ then define $f(x) = g(Tx)$ for all $x \in X$.

Definition 2.5.1 (Adjoint operator)**.** The *adjoint operator* $T^\times : Y^* \to X^*$ is defined by $T^\times(g) = g \circ T$ for all $g \in Y^*$.

Here, $T : X \to Y$ but $T^\times : Y^* \to X^*$.

Proposition 2.5.1. *The adjoint operator is linear with* $\|T^\times\| = \|T\|$.

Proof. Linearity is clear and $\|T^\times g\| \leq \|T\|\|g\|$ implies that $\|T^\times\| \leq \|T\|$. Now, for all $x_0 \in X$, Corollary 2.3.2.1 gives $g_0 \in Y^*$ such that $g_0(Tx_0) = \|Tx_0\|$ and $\|g_0\| = 1$. So, we obtain

$$\|Tx_0\| = g_0(Tx_0) = T^\times(g_0)(x_0) \leq \|T^\times\|\|g_0\|\|x_0\| = \|T^\times\|\|x_0\|.$$

This implies that $\|T\| \leq \|T^\times\|$ and the result follows. □

Notice that for the adjoint operator to have the a large enough norm, we needed the dual space to be large enough. Also, if $T : \mathbb{R}^n \to \mathbb{R}^n$ has matrix T_E with respect to some basis E, then the adjoint operator has matrix $T_E^\top$.

2.6 Reflexive spaces and separability

Proposition 2.6.1. *The* canonical mapping $x \mapsto g_x$ *where* $g_x(f) = f(x)$ *is an isomorphism from* X *to a subset of* X^{**} *with* $\|g_x\| = \|x\|$.

Proof. The proof is immediate from definitions. □

Definition 2.6.1 (Reflexive space)**.** A normed space X is *reflexive* if the canonical mapping is surjective.

Note that if X is reflexive then X is complete because the double dual is complete. Furthermore, every Hilbert space is reflexive due to the Riesz-Fréchet representation theorem. Now, we would like to show that if X^* is separable in a normed space then X is separable.

Lemma 2.6.2. *Suppose that* Y *is a proper closed subspace of a normed space* X *and fix* $x_0 \in X \setminus Y$ *with* $\delta := \operatorname{dist}(x_0, Y)$. *Then there exists* $\tilde{f} \in X^*$ *such that* $\tilde{f}(x_0) = \delta$, *and* $\tilde{f}(y) = 0$ *for all* $y \in Y$, *and* $\|\tilde{f}\| = 1$.

Proof. Apply the Hahn-Banach theorem (Theorem 2.3.2). □

Theorem 2.6.3. *If* X^* *is separable then* X *is separable for any normed space* X.

Proof. The unit sphere $U^* = \{f : \|f\| = 1\} \subseteq X^*$ is closed so it contains a countable dense subset $(f_n)_{n=1}^\infty$. Find $(x_n)_{n=1}^\infty$ such that $|f_n(x_n)| \geq 1/2$ since $\|f_n\| = 1$. Let $Y = \overline{\text{span}(\{x_n\}_{n=1}^\infty)}$. If $Y \neq X$, then by Lemma 2.6.2 there exists $\tilde{f} \in X^*$ such that $\tilde{f}(y) = 0$ for all $y \in Y$ and $\|f\| = 1$. In particular, we have (since $\tilde{f}(x_n) = 0$ for all n) that

$$\frac{1}{2} \leq |f_n(x_n)| \leq |f_n(x_n) - \tilde{f}(x_n)| \leq \|f_n - \tilde{f}\|\|x_n\| = \|f_n - \tilde{f}\|.$$

This contradicts the density of $(f_n)_{n=1}^\infty$ in U^*, so $Y = X$. Now pick rational coefficients in the span defining Y and conclude. □

2.7 Weak convergence and weak-* convergence

Definition 2.7.1 (Weak convergence)**.** A sequence $(x_n)_{n=1}^\infty$ *converges weakly* to x (written $x_n \xrightarrow{w} x$) if $f(x_n) \to f(x)$ for all $f \in X^*$.

Weak convergence behaves like strong convergence in the following ways.

Proposition 2.7.1. *If $x_n \xrightarrow{w} x$ then all subsequences converge weakly and the limit is unique. Also, $(\|x_n\|)_{n=1}^\infty$ is bounded.*

Proof. The first two statements are easy to show. For the third statement, consider the canonical mapping $g_n(f) = f(x_n)$. Since $(f(x_n))_{n=1}^\infty$ converges, it is bounded for all $f \in X^*$. Because X^* is complete, the uniform boundedness principle (Theorem 2.3.3) implies that $(\|g_n\|)_{n=1}^\infty$ is bounded. Conclude using $\|g_n\| = \|x_n\|$. □

Weak convergence is a less strict form of strong convergence when $\dim(X) = \infty$.

Proposition 2.7.2. *Strong convergence implies weak convergence, and the converse is true when* $\dim(X) < \infty$.

Proof. To show strong convergence implies weak convergence, we know that $|f(x_n) - f(x)| \leq \|f\|\|x_n - x\| \to 0$. To show the converse, suppose that $(e_i)_{i=1}^n$ is a basis for X and use the canonical basis for the dual (the one composed of functionals that send e_i to 1 and all other vectors to 0). □

We can easily write down an equivalent characterization of weak convergence using a total subset of the dual (a set $M \subseteq X^*$ such that $\overline{\text{span}(M)} = X^*$).

Proposition 2.7.3. *Weak convergence $x_n \xrightarrow{w} x$ is equivalent to $(\|x_n\|)_{n=1}^\infty$ being bounded and $f(x_n) \to f(x)$ for all f in a total set $M \subseteq X^*$.*

Proof. The forward implication is obvious. For the reverse implication, pick any $f \in X^*$ and $(f_j)_{j=1}^\infty$ in $\text{span}(M)$ such that $f_j \to f$. Then use a 3ϵ argument, since $f(x_n)$ is close to $f_j(x_n)$, $f_j(x_n)$ is close to $f_j(x)$, and $f_j(x)$ is close to $f(x)$. We needed boundedness of $(\|x_n\|)_{n=1}^\infty$ to ensure bounds on the first and third terms. □

Weak convergence intuitively captures many forms of convergence not captured by strong convergence; for instance, the sequence $(e_n)_{n=1}^\infty$ in ℓ^2 converges weakly to 0 but not strongly. We can also define notions of convergence for operators.

Definition 2.7.2 (Uniform operator convergence)**.** A sequence of operators $(T_n)_{n=1}^\infty$ *converges uniformly* to T if $\|T_n - T\| \to 0$.

Definition 2.7.3 (Strong operator convergence)**.** A sequence of operators $(T_n)_{n=1}^\infty$ is *strongly operator convergent* to T if $(T_n x)_{n=1}^\infty$ converges to $T(x)$ for all $x \in X$.

Definition 2.7.4 (Weak operator convergence)**.** A sequence of operators $(T_n)_{n=1}^\infty$ *converges weakly* to T if $(T_n x)_{n=1}^\infty$ converges weakly to $T(x)$ for all $x \in X$.

For example, $T_n =$ (set first n coordinates to 0) in ℓ^2 converges strongly but not uniformly to 0. Similarly, $T_n =$ (shift forward by n) in ℓ^2 converges weakly but not strongly to 0.

Note that for bounded linear functionals, strong operator convergence is equivalent to weak operator convergence. In this case, we call uniform operator convergence *strong convergence* and we call weak operator convergence *weak-* convergence.*

Example 2.7.1 (Convergence in distribution)**.** Convergence of random variables in distribution is an example of weak-* convergence. For a compact set X, the dual of $C(X)$ is the space of regular signed Borel measures on X (sometimes denoted $\mathcal{M}(X)$); this is called the Riesz-Markov representation theorem. The set of probability measures $\mathcal{P}(X)$ is a subset of $\mathcal{M}(X)$, so weak-* convergence of probability measures exactly means that $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$ for all $f \in C(X) = C_b(X)$. Of course, this is the characterization of convergence in distribution given by the portmanteau lemma. Similar analysis works over unbounded sets X by considering the space of continuous functions vanishing at infinity (called $C_0(X)$), whose dual is the space of Radon measures on X.

Now, strong operator convergence implies that the limit operator is bounded, assuming that the input comes from a Banach space.

Lemma 2.7.4. *If $T_n \to T$ is strongly operator convergent then $T \in B(X, Y)$ as long as X is a Banach space.*

Proof. Linearity is immediate, and boundedness follows from the uniform boundedness principle (Theorem 2.3.3). □

We have the following characterization of strong operator convergence.

Proposition 2.7.5. *A sequence of operators $(T_n)_{n=1}^\infty$ converges strongly to T if and only if $(\|T_n\|)_{n=1}^\infty$ is bounded and $(T_n x)_{n=1}^\infty$ is Cauchy for all $x \in M$, where $M \subseteq X$ is total.*

Proof. The proof is more or less identical to the proof of Proposition 2.7.3. □

Since strong operator convergence is equivalent to weak-* convergence for bounded operators, the same characterization holds for weak-* convergence. We now prove the Banach-Alaoglu theorem, which gives a way to show weak-* convergence.

Theorem 2.7.6 (Banach-Alaoglu)**.** *If X is a normed space and $(f_n)_{n=1}^\infty$ is a sequence in X^* with $\|f_n\| \leq 1$ for all $n \in N$, then there exists a subsequence such that $(f_{n_k})_{k=1}^\infty$ converges in weak-* to some $f \in X^*$.*

This theorem can be stated more generally in the context of topological vector spaces: the closed unit ball of the dual of a normed space is compact in the weak-* topology. However, we state the result here (equivalently) in terms of subsequential weak-* convergence.

Proof. The trick of this proof is to embed our space into a compact product space, such that convergence in the product space automatically gives our desired convergence. We associate with each point $x \in X$ the compact set $K_x = [-\|x\|, \|x\|] \subseteq \mathbb{R}$. By Tychonoff's theorem (Theorem A.3.2), the product $\prod_{x \in X} K_x$ is compact. Define the map Φ on X^* by $\Phi(f) = (f(x))_{x \in X}$ so that $\Phi(f) \in \prod_{x \in X} K_x$ when $\|f\| \leq 1$. Extract a convergent subsequence $\Phi(f_{n_k}) \to \Phi(f)$ in the product topology on $\prod_{x \in X} K_x$. Now, for any $x \in X$, we have $f_{n_k}(x) \to f(x)$ so that $f_{n_k} \xrightarrow{w} f$ as desired. □

Chapter 3

Convex analysis

In this chapter, we discuss many classical concepts in convex analysis which appear in many other areas of applied math. We will assume that the reader knows some basic definitions, and most of the material will come from Rockafellar (1970).

3.1 Convex sets and functions

Most of this section is intended as review.

Proposition 3.1.1. *Intersections of convex sets are convex.*

Definition 3.1.1 (Epigraph)**.** If $f : \mathbb{R}^n \to [-\infty, \infty]$ is a function, then the *epigraph* of f is the set

$$\text{epi}(f) = \{(x, \mu) \in \mathbb{R}^n \times \mathbb{R} : f(x) \leq \mu\}.$$

Essentially, if you draw the graph of f, the epigraph is the set of points above the graph (including the graph itself).

Definition 3.1.2 (Proper function)**.** A function f is *proper* if it maps into $(\infty, \infty]$ and is not identically ∞.

Definition 3.1.3 (Convex function)**.** A function f is *convex* if its epigraph is a convex set. If f is proper then this is equivalent to Jensen's inequality.

Proposition 3.1.2. *If f is convex then the* sublevel sets $\{x : f(x) \leq \alpha\}$ *are convex for all $\alpha \in \mathbb{R}$. So are the strict sublevel sets $\{x : f(x) < \alpha\}$.*

Proof. The α-sublevel set is the intersection of $\text{epi}(f)$ with the hyperplane $\mu = \alpha$ (restricted to the first coordinate), and the convexity of the strict sublevel sets follows from Jensen's inequality. □

Proposition 3.1.3. *If f and φ are proper and convex with φ nondecreasing and $\varphi(\infty) = \infty$, then $\varphi \circ f$ is convex.*

The proof is easy so we omit it. For example, this implies that $e^{f(x)}$ is convex if $f > 0$ is convex.

Proposition 3.1.4. *$f(x) = \inf\{\mu : (x, \mu) \in K\}$ is convex if K is convex.*

Again, we omit the proof. We know that we can associate a convex set to every convex function by $\text{epi}(f)$, but it now follows that we can almost treat any convex set like $\text{epi}(g)$ for some convex function g.

3.2 Lower semi-continuous functions

To motivate this section, notice that convex functions need not have consistent limiting behavior. For example, take the function $f(x) = x^2 + \mathbf{1}_{x \neq 0}$ with $\text{dom}(f) = [0, \infty)$, which is convex. One could try to computationally optimize this function and obtain $x^* = 0$ as an output, but this is totally meaningless. In some sense, this is the only problem that convex functions can have. They can have points where they are unexpectedly high, but they can't have points where they are unexpectedly low. We need to impose some sort of continuity condition to ensure that the function behaves as expected, and the right condition is lower semi-continuity.

Definition 3.2.1 (Lower semi-continuity)**.** A function f is *lower semi-continuous* (l.s.c.) if $f(x) \leq \liminf_{y \to x} f(y)$ for all $x \in \text{dom}(f)$.

In particular, if we tried to minimize the function only using surrounding values, there would not be an unwelcome surprise at the limit no matter how we approach. We can equivalently characterize lower semi-continuity as follows, highlighting the deep connection between lower semi-continuity and convexity.

Proposition 3.2.1. *A function f is l.s.c. if and only if the sublevel sets $\{x : f(x) \leq \alpha\}$ are closed for all $\alpha \in \mathbb{R}$. Also, f is l.s.c. if and only if $\text{epi}(f)$ is closed.*

The proof is easy so we omit it. We will use this characterization later.

Proposition 3.2.2. *If f takes values in $[0, \infty]$ and is l.s.c. then it can be written as the pointwise supremum of Lipschitz functions.*

Note that the proof can be adapted so that this holds in any metric space.

Proof. For $\lambda \geq 0$, define $f_\lambda(x) = \inf\{f(y) + \lambda\|x - y\|_2 : y \in \mathbb{R}^n\}$; this is called the *Moreau-Yosida trick*. Here, f_λ is the infimal convolution of f with $\lambda\|\cdot\|_2$. It is easy to check that f_λ is λ-Lipschitz and $f_\lambda \uparrow f$ pointwise as $\lambda \to \infty$. □

3.3 Separation theorems

In this section, we cover the supporting hyperplane theorem and the separating hyperplane theorem, which are fundamental results in convex analysis.

Definition 3.3.1 (Gauge function)**.** The *gauge function* of a set C is the function $\gamma_C(x) = \inf\{\lambda > 0 : x \in \lambda C\}$, and is convex if C is convex. The gauge function is sometimes called the *Minkowski functional* of C.

Proposition 3.3.1. *The gauge function γ_C is a sublinear functional (as defined in Section 2.3.1) when C is convex.*

Proof. Positive homogeneity is clear. For subadditivity, fix any $\alpha > \gamma_C(x)$ and $\beta > \gamma_C(y)$ so that $x \in \alpha C$ and $y \in \beta C$. Then $x + y \in (\alpha + \beta)C$ so that $\gamma_C(x + y) \leq \alpha + \beta$. Take an infimum over all such α, β to get the result. □

Lemma 3.3.2 (Separating a point from a set)**.** *If C is a nonempty convex set then for all $x_0 \notin C$, there exists a nonzero $w \in \mathbb{R}^n$ such that $\langle w, x\rangle \leq \langle w, x_0\rangle$ for all $x \in C$.*

Proof. Suppose without loss of generality that $0 \in C$. Define $L = \text{span}(\{x_0\})$ and define a linear functional on L by $f(\alpha x_0) = \alpha$ for $\alpha \in R$. Now, we see that $f(v) \leq \gamma_C(v)$ for all $v \in L$ since $\gamma_C(\alpha x_0) \geq \alpha = f(\alpha x_0)$ for $\alpha > 0$ and $f(\alpha x_0) \leq 0 \leq \gamma_C(\alpha x_0)$ for $\alpha \leq 0$.

By the Hahn-Banach theorem (Theorem 2.3.2), we can extend f to a linear functional $\tilde{f}$ on all of $\mathbb{R}^n$ satisfying $\tilde{f}(v) \leq \gamma_C(v)$ for all $v \in \mathbb{R}^n$. By the Riesz-Fréchet representation theorem (Theorem 2.2.2), there exists a nonzero vector $w \in \mathbb{R}^n$ such that $\tilde{f}(x) = \langle w, x\rangle$ for all $x \in \mathbb{R}^n$. For any $x \in C$, we have $\langle w, x\rangle = \tilde{f}(x) \leq \gamma_C(x) \leq 1 = \tilde{f}(x_0)$. □

Note that the Hahn-Banach theorem wasn't strictly needed in $\mathbb{R}^n$, but the point is that the same proof works in arbitrary normed spaces. From this lemma, we derive the separating hyperplane theorem.

Theorem 3.3.3 (Separating hyperplane)**.** *If A and B are nonempty disjoint convex sets, then there exists a nonzero $w \in \mathbb{R}^n$ such that $\sup_{x\in A}\langle w, x\rangle \leq \inf_{x\in B}\langle w, x\rangle$.*

Proof. Let $C = A - B = \{x - y : x \in A, y \in B\}$. Then C is convex and does not contain 0, so by Lemma 3.3.2 there exists a nonzero w such that $\langle w, x\rangle \leq \langle w, 0\rangle = 0$ for all $x \in C$. This implies that $\langle w, a - b\rangle \leq 0$ for all $a \in A$ and $b \in B$, so $\langle w, a\rangle \leq \langle w, b\rangle$ for all $a \in A$ and $b \in B$. □

Sometimes, the separating hyperplane theorem is also called the *Hahn-Banach separation theorem*. Now, we are ready to show the supporting hyperplane theorem.

Theorem 3.3.4 (Supporting hyperplane)**.** *If C is a nonempty convex set and x_0 is a boundary point of C, then there exists a nonzero $w \in \mathbb{R}^n$ such that $\langle w, x\rangle \leq \langle w, x_0\rangle$ for all $x \in C$.*

Proof. Pick a sequence $(y_n)_{n=1}^{\infty}$ in C^c converging to x_0. By the separating hyperplane theorem (Theorem 3.3.3), there exists a nonzero continuous linear functional f_n such that $f_n(y) \leq f_n(y_n)$ for all $y \in C$ for all $n \in \mathbb{N}$. Renormalize the f_n to have norm 1 (preserving the separation); by the Banach-Alaoglu theorem (Theorem 2.7.6), there exists a subsequence $(f_{n_k})_{k=1}^{\infty}$ converging in weak-* to some continuous linear functional f with $\|f\| = 1$. For all $k \in \mathbb{N}$, we know that $f_{n_k}(x) \leq f_{n_k}(y_{n_k})$ for all $x \in C$, so $f_{n_k}(x - x_0) \leq f_{n_k}(y_{n_k} - x_0)$ for all $x \in C$. As $k \to \infty$, the left-hand side converges to $f(x - x_0)$ by weak-* convergence and the right-hand side converges to zero since $\|f_{n_k}\| = 1$ and $y_{n_k} \to x_0$. Hence, we deduce that $f(x) \leq f(x_0)$ for all $x \in C$, and we conclude by the Riesz-Fréchet representation theorem (Theorem 2.2.2). □

We could have given a purely elementary proof here as well, but the proof above is more general and gives a simple example of the usefulness of the Banach-Alaoglu theorem. The previous theorems have an immediate corollary.

Corollary 3.3.4.1. *A closed and convex set C is the intersection of all closed half-spaces containing it.*

In fact, we can give a useful property of convex l.s.c. functions using the supporting hyperplane theorem.

Theorem 3.3.5. *If f is convex and l.s.c. then f is the supremum of affine functions lying below f.*

Proof. Since f is l.s.c., we know that $\text{epi}(f)$ is closed and convex. By Corollary 3.3.4.1, $\text{epi}(f)$ is the intersection of all closed half-spaces containing it, each represented as $\{x \in \mathbb{R}^n : \langle a, x\rangle + b\mu \leq c\}$ for some $a \in \mathbb{R}^n$ and $b, c \in \mathbb{R}$. Not all of these half-spaces can have $b = 0$, since otherwise f is trivial and the theorem follows immediately. We can't have $b > 0$ at all, since the epigraph of f extends upwards infinitely. When $b < 0$, we can normalize to $b = -1$ so that $\langle a, x\rangle - c \leq \mu$; this defines an affine function $\langle a, x\rangle - c$ lying below f. The constraints induced by half-spaces with $b < 0$ are the only nontrivial ones, so the result follows. □

3.4 Subgradients and subdifferentials

The idea of subdifferentials is to use a supporting hyperplane to generalize the notion of a derivative for non-differentiable functions.

Definition 3.4.1 (Subgradient)**.** A vector x^* is a *subgradient* of f at x if $f(z) \geq f(x) + \langle x^*, z - x\rangle$ for all $z \in \mathbb{R}^n$.

Definition 3.4.2 (Subdifferential)**.** The subdifferential of f at x is the set of all subgradients at x, denoted $\partial f(x)$.

Although one can theoretically define a subdifferential for any function, we almost always restrict ourselves to convex functions so that the subdifferential enjoys several nice properties.

Proposition 3.4.1. *If f is convex then $\partial f(x)$ is a closed convex set for all $x \in \operatorname{dom}(f)$. Also, the subdifferential is nonempty for all $x \in \operatorname{int}(\operatorname{dom}(f))$, and for all $x \in \operatorname{dom}(f)$ if f is l.s.c..*

Proof. It follows from the supporting hyperplane theorem (Theorem 3.3.4) that the subdifferential is nonempty under the given conditions. Also, closedness and convexity are immediate from the definition. □

Example 3.4.1 (Subdifferential of the norm)**.** The subdifferential of the Euclidean norm $\|\cdot\|_2$ at x is $\{x/\|x\|_2\}$ at every $x \neq 0$ and $B(0,1)$ at $x = 0$, by Cauchy-Schwarz.

Note that if the gradient exists at a point x, then the subdifferential at that point is a singleton containing only the gradient; this can be seen by using Jensen's inequality when taking the directional derivative at x in the direction of $z - x$.

3.5 The Legendre-Fenchel transform

The Legendre-Fenchel transform gives a rich duality theory for convex functions.

Definition 3.5.1 (Legendre-Fenchel transform)**.** The *Legendre-Fenchel transform* of f is the function f^* defined by

$$f^*(y) = \sup_{x \in \mathbb{R}^n} \{\langle y, x\rangle - f(x)\}.$$

The function f^* is also called the *convex conjugate* of f.

In a general normed space X, the Legendre-Fenchel transform can be thought of as a transformation from X^* to $\mathbb{R}$, defined by

$$f^*(g) = \sup_{x \in X} \{g(x) - f(x)\}.$$

The Legendre-Fenchel transform represents the maximal difference between a linear function $\langle y, x\rangle$ and $f(x)$. If f is convex, the Legendre-Fenchel transform is immediately convex as it is the supremum of affine functions. Intuitively, if $f(x)$ is the cost to buy a portfolio x of products, then $f^*(g)$ can be thought of as the maximum profit that you can make by setting linear prices $g(x)$ (under any portfolio). The following is a simple result, but is fundamental in the theory.

Theorem 3.5.1 (Fenchel-Young inequality)**.** *If f is convex then $f(x) + f^*(y) \geq \langle y, x\rangle$ for all $x, y \in \mathbb{R}^n$. Equality holds if and only if $y \in \partial f(x)$.*

Proof. The inequality is obvious from the previous interpretation of the Legendre-Fenchel transform, and the equality condition follows from the definition of the subdifferential. □

Theorem 3.5.2 (Fenchel-Moreau)**.** *If f is l.s.c. and convex, then $f^{**} = f$.*

Proof. The inequality $f \geq f^{**}$ is immediate from definitions. Now, note that for every affine function $a \leq f$, we have $a^{**} = a \leq f^{**}$; in particular, the affine functions lying below f are the same ones below f^{**}. Taking a supremum over all affine functions $a \leq f$ and applying Theorem 3.3.5, it follows that $f^{**} \geq f$. □

Further, we have the following useful strong duality theorem, which gives a dual problem for the minimization of a function subject to constraints.

Theorem 3.5.3 (Fenchel-Rockafellar)**.** *Suppose f_1 and f_2 are convex functions on a normed space X taking values in $[0, \infty]$. Also, assume that there exists $x_0 \in X$ such that $f_1(x_0) + f_2(x_0) < \infty$ and f_1 is continuous at x_0. Then we have the minimax principle*

$$\inf_{x\in X}\{f_1(x) + f_2(x)\} = \sup_{g\in X^*}\{-f_1^*(-g) - f_2^*(g)\} = \sup_{g\in X^*}\inf_{x,y\in X}\{f_1(x) + f_2(y) + g(x - y)\}.$$

Intuitively, suppose x denotes a portfolio of different products. Suppose $f_1(x)$ is the cost of producing x, and $f_2(x)$ is a convex penalty describing whether producing x is feasible or not. Then, the primal problem is to minimize the cost of producing x subject to feasibility constraints. On the other hand, suppose g denotes a vector of market prices for the products. Then, $-f_1^*(-g) = \inf_{x\in X}\{g(x)+f_1(x)\}$ is the minimum net cost under prices g and $-f_2^*(g) = \inf_{x\in X}\{f_2(x) - g(x)\}$ is the minimum gross profit under prices g and feasibility constraints. The dual problem is to price in a way that maximizes the worst-case profit for the producer. By the Fenchel-Rockafellar duality, minimizing the cost of producing x subject to feasibility constraints is equivalent to pricing in a way that maximizes the worst-case profit for the producer.

Proof. By choosing $x = y$, we see that the right hand side is less than or equal to the left hand side. The reverse inequality will now follow from the separating hyperplane theorem. Define $m = \inf_{x\in X}\{f_1(x) + f_2(x)\}$, which is finite because we assumed $f_1(x_0) + f_2(x_0) < \infty$. Let $C = \mathrm{epi}(f_1)$ and $C' = \{(x, \mu) \in X \times \mathbb{R} : \mu < m - f_2(x)\}$; these are convex sets and disjoint by definition of m. Furthermore, C has nonempty interior since $(x_0, f_1(x_0) + 1) \in \mathrm{int}(C)$ by continuity of f_1 at x_0, so C has nonempty interior. Note that in an infinite-dimensional space, we need one of the sets to have nonempty interior in order to apply the separating hyperplane theorem.

By the separating hyperplane theorem (Theorem 3.3.3), there exists a nonzero $g \in X^*$ such that $g(x) + \alpha\lambda \geq g(y) + \alpha\mu$ for all $(x, \lambda) \in C$ and $(y, \mu) \in C'$. It is easy to see that $\alpha > 0$, so assume w.l.o.g. that $\alpha = 1$. Then, we can rearrange

$$g(x) + f_1(x) \geq g(x) + \lambda \geq g(y) + \mu \geq g(y) + m - f_2(y)$$

to obtain the reverse inequality. □

3.6 Cyclical monotonicity

Cyclical monotonicity is a characterizing property of the subdifferential of convex functions, and is used in optimal transport theory to prove Brenier's theorem.

Definition 3.6.1 (Cyclical monotonicity)**.** A subset $\Gamma \subseteq \mathbb{R}^n \times \mathbb{R}^n$ is *cyclically monotone* if for all finite collections $(x_1, y_1), \ldots, (x_m, y_m) \in \Gamma$, we have

$$\sum_{i=1}^{m} \langle y_i,\, x_{i+1} - x_i \rangle \leq 0,$$

under the convention $x_{m+1} = x_1$.

In some sense, cyclical monotonicity is the generalization of monotonicity to $\mathbb{R}^n$, since monotone functions on $\mathbb{R}$ are the derivatives of convex functions. The motivation for this definition is Rockafellar's theorem.

Theorem 3.6.1 (Rockafellar)**.** *A nonempty $\Gamma \subseteq \mathbb{R}^n \times \mathbb{R}^n$ is cyclically monotone if and only if it is included in the graph of $\partial\varphi$ for some proper l.s.c. convex function $\varphi : \mathbb{R}^n \to (-\infty, \infty]$.*

Proof. For the reverse direction, suppose that Γ is included in the graph of $\partial\varphi$ for a l.s.c. convex function φ. If $(x_1, y_1), \ldots, (x_m, y_m) \in \Gamma$, then for $1 \leq i \leq m$, we have

$$\varphi(x_{i+1}) \geq \varphi(x_i) + \langle y_i,\, x_{i+1} - x_i \rangle$$

by definition of the subdifferential. Summing over i shows that Γ is cyclically monotone. For the forward direction, suppose that Γ is cyclically monotone and nonempty. Pick $(x_0, y_0) \in \Gamma$ and define the function

$$\varphi(x) = \sup\{\langle y_m,\, x - x_m \rangle + \langle y_{m-1},\, x_m - x_{m-1} \rangle + \cdots + \langle y_0,\, x_1 - x_0 \rangle : (x_1, y_1), \ldots, (x_m, y_m) \in \Gamma\}.$$

Since φ is the supremum of affine functions it is l.s.c. and convex. Also, φ is proper since $\varphi(x_0) \leq 0$ by cyclical monotonicity. It is now easy to verify by the definition of φ that Γ is contained in the graph of $\partial\varphi$. □

Chapter 4

Optimal transport

In this chapter, we write about optimal transport theory, primarily using material from Villani (2003) and Chewi et al. (2025). Some material may also come from Santambrogio (2015) or Ambrosio et al. (2005).

4.1 The Monge-Kantorovich problem

The Monge problem is that of finding a deterministic transport map between measures. We let $\mathcal{P}(\mathcal{X})$ denote the set of probability measures over $\mathcal{X}$ and $T_\sharp$ denote the pushforward operator. In everything that follows, we will assume $\mathcal{X}$ and $\mathcal{Y}$ are Polish spaces (complete and separable metric spaces). The main reason to assume a Polish space setting is so that we can use Prokhorov's theorem (Theorem 1.1.4) to extract weakly convergent subsequences from tight sets, as well as for disintegration of measure (Theorem 1.4.1).

Definition 4.1.1 (Monge problem)**.** If $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$ and $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is a cost function, the *Monge problem* is

$$\inf_{T_\sharp \mu = \nu} \int c(x, T(x))\, d\mu(x).$$

The Kantorovich problem is the problem of finding a stochastic transport plan; one motivation for this is that there is literally no deterministic transport map from δ_0 to $\frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1$ and another is that the Monge problem is highly nonconvex and hard to solve. Let $\Pi(\mu, \nu)$ denote the set of couplings between μ and ν (joint distributions with marginals μ and ν).

Definition 4.1.2 (Kantorovich problem)**.** If $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$ and $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is a cost function, the *Kantorovich problem* is

$$\inf_{\pi \in \Pi(\mu,\nu)} \int c(x, y)\, d\pi(x, y).$$

Equivalently, if X and Y are random variables, the Kantorovich problem is to minimize $\mathbb{E}_\pi[c(X,Y)]$ over all joint distributions π of X and Y. Note that the Kantorovich problem is a linear program since the objective and constraints are linear. Here are a few simple facts about the set of couplings.

Proposition 4.1.1. *The set of couplings $\Pi(\mu,\nu)$ is a nonempty and convex. Also, if $(\pi_n)_{n=1}^\infty$ is a sequence in $\Pi(\mu,\nu)$, then there exists a weakly convergent subsequence with a limit also in $\Pi(\mu,\nu)$.*

Proof. Nonemptiness follows from $\mu \times \nu \in \Pi(\mu,\nu)$ and convexity is obvious. By Prokhorov's theorem (Theorem 1.1.4), it suffices to show that $\Pi(\mu,\nu)$ is tight and closed. Tightness follows because μ and ν are simultaneously concentrated on a compact set K, and $\pi((K\times K)^c) \le \mu(K^c) + \nu(K^c)$. Note that π is a coupling if and only if $\int f(x)\,d\pi(x,y) = \int f(x)\,d\mu(x)$ for all $f \in C_b(X)$ and $\int f(y)\,d\pi(x,y) = \int f(y)\,d\nu(y)$ for all $f \in C_b(Y)$, by the Riesz-Markov representation theorem. Therefore, $\Pi(\mu,\nu)$ is closed by the portmanteau lemma (Theorem 1.1.2). □

Proposition 4.1.2. *If the cost function c is l.s.c., then the Kantorovich problem has a solution.*

Proof. Let $\pi_n \in \Pi(\mu,\nu)$ be such that

$$\int c(x,y)\,d\pi_n(x,y) \to \inf_{\pi\in\Pi(\mu,\nu)} \int c(x,y)\,d\pi(x,y).$$

and extract a convergent subsequence $\pi_{n_k} \to \pi \in \Pi(\mu,\nu)$ by Proposition 4.1.1. By (4) in the portmanteau lemma (Theorem 1.1.2) and because c is l.s.c., we have

$$\int c(x,y)\,d\pi(x,y) \le \liminf_{k\to\infty} \int c(x,y)\,d\pi_{n_k}(x,y) = \inf_{\pi\in\Pi(\mu,\nu)} \int c(x,y)\,d\pi(x,y).$$

In particular, π solves the Kantorovich problem. □

Note that this statement is a special case of the *extreme value theorem*, which states that l.s.c. functions must attain their infimum on compact sets. Even though the Kantorovich problem often has a solution, it is not always unique; for example, consider the case $\mu = \frac{1}{2}\delta_{(-1,0)} + \frac{1}{2}\delta_{(1,0)}$ and $\nu = \frac{1}{2}\delta_{(0,-1)} + \frac{1}{2}\delta_{(0,1)}$ with the quadratic cost function.

4.2 Transport maps between empirical averages

In this section, we show that we can fully characterize optimal transport maps between measures associated to empirical averages. We start by proving the Krein-Milman theorem, which states that points in a compact convex set in a Banach space can be written as limits of convex combinations of extreme points.

Theorem 4.2.1 (Krein-Milman)**.** *Let K be a nonempty, compact, and convex subset of a Banach space and let $\mathcal{E}(K)$ denote the set of extremal points of K (points that cannot be written as the convex combination*

of any two other points). Then for all $x \in K$ there exists a probability measure ρ_x on $\mathcal{E}(X)$ such that $x = \int_{\mathcal{E}(X)} y \, d\rho_x(y)$.

Proof. We would first like to show that $C := \overline{\text{conv}(\mathcal{E}(K))} = K$. Suppose not for a contradiction; then there exists $x \in K \setminus C$. By the separating hyperplane theorem (Theorem 3.3.3), there exists a continuous linear functional f and a constant c such that $f(x) > \max_{y \in C} f(y) \geq \max_{y \in \mathcal{E}(K)} f(y)$ (using compactness of C and closedness of $\mathcal{E}(K)$). Let $M = f^{-1}(\max_{x \in C} f(x))$, which is a nonempty, closed, and compact; this fact is sometimes called the *Krein-Milman lemma* and is easy to show. So M has an extreme point z (which is also in $\mathcal{E}(K)$) and we have the inequality

$$f(z) \geq f(x) > \max_{y \in C} f(y) \geq \max_{y \in \mathcal{E}(K)} f(y) \geq f(z),$$

giving a contradiction. Then for each $x \in K$, we can extract a sequence of discrete probability measures $\rho_x^{(n)}$ such that $\int y \, d\rho_x^{(n)} \to x$ as $n \to \infty$. By the Banach-Alaoglu theorem (Theorem 2.7.6), we can extract a weakly convergent subsequence $\rho_x^{(n_k)}$ with limit ρ_x, and the result follows since $\int y \, d\rho_x = x$. □

Using the Krein-Milman theorem, we can show Choquet's theorem, which states that continuous linear functionals on a compact convex set in a Banach space achieve their minimum at an extreme point.

Theorem 4.2.2 (Choquet)**.** *Let K be a nonempty, compact, and convex subset of a Banach space and let $f : K \to \mathbb{R}$ be the restriction of a continuous linear functional. Then the minimum of f over K is achieved at an extreme point of K.*

Proof. Note that f has a minimizer $x \in K$ since K is compact. By the Krein-Milman theorem (Theorem 4.2.1), there exists a probability measure ρ_x on $\mathcal{E}(K)$ such that $x = \int_{\mathcal{E}(K)} y \, d\rho_x(y)$. Then, by continuity and linearity of f, we have

$$f(x) = f\left(\int_{\mathcal{E}(K)} y \, d\rho_x(y)\right) = \int_{\mathcal{E}(K)} f(y) \, d\rho_x(y).$$

It is easy to formalize this exchange by approximating $y \mapsto y$ uniformly on K by simple functions and applying the dominated convergence theorem. Clearly, this means that ρ_x cannot give positive mass to any point $y \in \mathcal{E}(K)$ such that $f(y) > f(x)$, so there must exist $y \in \mathcal{E}(K)$ such that $f(y) = f(x)$, which is the result we wanted. □

Using Choquet's theorem, we can show that the solution to the Kantorovich problem between discrete measures must be a deterministic permutation; in particular, the Kantorovich problem is equivalent to the Monge problem in this case.

Theorem 4.2.3 (Birkhoff)**.** *Let $\mathcal{B}_n$ denote the set of bistochastic $n \times n$ matrices (matrices with nonnegative entries such that each row and column sums to 1). Then the set of extremal points of $\mathcal{B}_n$ is exactly the set of permutation matrices.*

Proof. It's clear that all permutation matrices are extremal points, so we show the converse. It suffices to show that whenever M is an extremal point of $\mathcal{B}_n$, all entries of M are either 0 or 1. Suppose that $M \in \mathcal{B}_n$ with $M_{ij} \in (0,1)$ for some $i, j \in [n]$. Then make a list of indices in the matrix, first finding another entry in the same column in $(0,1)$ and then an entry in the same row from that new point in $(0,1)$, and so on. Continue this process until we either end up at the same row or the same column of M. The list will look something like $((i,j), \dots, (i_f, j_f))$; if we ended up at the same column, then we delete (i,j) from the beginning of our list.

The list is now of even length and only contains entries in $(0,1)$, so pick $\epsilon > 0$ smaller than the gap $\min\{\min\{M_{ij}\}_{ij}, \min\{1 - M_{ij}\}_{ij}\}$. Then we can create a matrix $M^{(1)}$ by bumping the entry of M at the first index in the list up by ϵ, the entry at the second index down by ϵ, and so on. Similarly, we create a matrix $M^{(2)}$ by bumping the entry of M at the first index in the list down by ϵ, the entry at the second index up by ϵ, and so on. Then $M = \frac{1}{2}(M^{(1)} + M^{(2)})$, and because $M^{(1)}, M^{(2)} \in \mathcal{B}_n$, M cannot be an extremal point of $\mathcal{B}_n$. □

Proposition 4.2.4. *The solution to the Kantorovich problem between discrete spaces $\mathcal{X}$ and $\mathcal{Y}$ where $\mu = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and $\nu = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$ is a deterministic transport map $T : \mathcal{X} \to \mathcal{Y}$ such that $T(x_i) = y_{\sigma(i)}$ for some permutation $\sigma \in S_n$.*

Proof. Let $\mathcal{B}_n$ denote the set of bistochastic $n \times n$ matrices, so that the Kantorovich problem is

$$\inf_{\pi \in \mathcal{B}_n} \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n \pi_{ij}\, c(x_i, y_j).$$

By Choquet's theorem (Theorem 4.2.2), the minimum is achieved at an extremal point of $\mathcal{B}_n$, which by Birkhoff's theorem (Theorem 4.2.3) is a permutation matrix. □

4.3 Wasserstein distances

Suppose $(\mathcal{X}, d)$ is a Polish space (a complete and separable metric space). Let $\mathcal{P}_p(\mathcal{X})$ denote the set of probability measures over $\mathcal{X}$ with finite pth moment, meaning that

$$\int d(x_0, x)^p \, d\mu(x) < \infty$$

for all $x_0 \in \mathcal{X}$. Let $\mathcal{T}_p(\mu, \nu)$ denote the optimal transportation cost between measures μ and ν if $c(x,y) = d(x,y)^p$; we can now define the Wasserstein distances on $\mathcal{P}_p(\mathcal{X})$ as follows.

Definition 4.3.1 (Wasserstein distances)**.** For $p \geq 0$, the p-Wasserstein distance between measures $\mu, \nu \in \mathcal{P}_p$ is defined as $W_p(\mu, \nu) = \mathcal{T}_p(\mu, \nu)^{\min\{1, 1/p\}}$.

Theorem 4.3.1. *The p-Wasserstein distance is actually a metric on $\mathcal{P}_p(\mathcal{X})$.*

Proof. We may assume $p \geq 1$, since otherwise d^p is topologically equivalent to d (although it may not be a metric) and is subadditive. It's obvious that W_p is finite, symmetric, and nonnegative. If $\mu = \nu$, then the measure $d\mu(x) \times d\delta_x(y)$ is a coupling, showing that $W_p(\mu, \nu) = 0$. On the other hand, if $W_p(\mu, \nu) = 0$, there exists a coupling π such that $\mathbb{E}_{(X,Y)\sim\pi}[d(X, Y)^p] = 0$ by Proposition 4.1.2. But now this forces $X = Y$ almost surely so $\mu = \nu$.

Finally, we prove the triangle inequality. First, we need the following lemma, which allows us to put together measures with a common marginal; essentially, this is a technical lemma that will allow us to put together optimal transport plans to make a (possibly suboptimal) transport plan.

Lemma 4.3.2 (Gluing)**.** *Let μ_1, μ_2, μ_3 be measures on Polish spaces $\mathcal{X}_1, \mathcal{X}_2, \mathcal{X}_3$ with $\pi_{12} \in \Pi(\mu_1, \mu_2)$ and $\pi_{23} \in \Pi(\mu_2, \mu_3)$. Then there exists a measure $\pi \in \mathcal{P}(\mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3)$ such that π has marginals π_{12} and π_{23} in its restriction to the first two or last two coordinates respectively.*

Proof. Applying the disintegration theorem (Theorem 1.4.1) along the second coordinate of π_{12} and the first coordinate of π_{23}, we have

$$\pi_{12} = \int_{\mathcal{X}_2} (\pi_{12})_y \otimes \delta_y \, d\mu_2(y)$$

and

$$\pi_{23} = \int_{\mathcal{X}_2} \delta_y \otimes (\pi_{23})_y \, d\mu_2(y),$$

such that $y \mapsto (\pi_{12})_y$ and $y \mapsto (\pi_{23})_y$ are measurable. Then we can define π by

$$\pi = \int_{\mathcal{X}_2} (\pi_{12})_y \otimes \delta_y \otimes (\pi_{23})_y \, d\mu_2(y),$$

which satisfies the properties of the gluing lemma. □

Now, we are ready to complete the proof of the triangle inequality. Consider $\mu_1, \mu_2, \mu_3 \in \mathcal{P}_p(\mathcal{X})$ and let $\pi_{12} \in \Pi(\mu_1, \mu_2)$ and $\pi_{23} \in \Pi(\mu_2, \mu_3)$ be optimal transport plans. Glue these together using the gluing lemma (Lemma 4.3.2) to get $\pi \in \mathcal{P}(\mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3)$ and let $(X, Y, Z) \sim \pi$. Then, we have by Minkowski's inequality that

$$\begin{aligned} W_p(\mu_1, \mu_3) &\leq \|d(X, Z)\|_p \\ &\leq \|d(X, Y) + d(Y, Z)\|_p \\ &\leq \|d(X, Y)\|_p + \|d(Y, Z)\|_p \\ &= W_p(\mu_1, \mu_2) + W_p(\mu_2, \mu_3). \end{aligned}$$

□

Example 4.3.1 (Point masses)**.** Note that $W_p(\delta_x, \delta_y) = d(x, y)$, so $(\mathcal{X}, d)$ is isometrically embedded in $(\mathcal{P}_p(\mathcal{X}), W_p)$ by the map $x \mapsto \delta_x$.

By the ordering of the L^p norms in a probability space (Proposition 1.1.6), we have $W_p(\mu, \nu) \leq W_q(\mu, \nu)$ for $1 \leq p \leq q$. Note that the Wasserstein distances metrize weak convergence.

Proposition 4.3.3. *For $p \geq 1$, $W_p(\mu_n, \mu) \to 0$ if and only if $\mu_n \xrightarrow{d} \mu$.*

Now, we give a few techniques to bound the Wasserstein distance. First, one can bound the Wasserstein distance from below by coming up with an inequality that holds uniformly over couplings, pulling out an $X - Y$ using Hölder's inequality, and then choosing the optimal coupling to get a tight bound.

Proposition 4.3.4. *If $X \sim \mu$ and $Y \sim \nu$ are sub-exponential in the sense that $\mathbb{E}[e^{|X|}] \leq 2$ and $\mathbb{E}[e^{|Y|}] \leq 2$, then for any $p > 1$ there exists a universal constant $c_p > 0$ so that for all $k \in \mathbb{N}$, we have*

$$\mathbb{E}[||X|^k - |Y|^k|] \leq (c_p k)^k \, W_p(\mu, \nu).$$

Hence, a bound on the Wasserstein distance between μ and ν simultaneously implies that all of their moments are close, as long as the tails decay sufficiently fast.

Proof. Note that if $f(x) = |x|^k$, then $k|x|^{k-1} \in \partial f(x)$ for all $x \in \mathbb{R}$ and $k \in \mathbb{N}$. In particular, we know that:

$$|X|^k - |Y|^k \leq |X - Y| \cdot k \, |X|^{k-1}.$$

A symmetric bound holds if we swap the roles of X and Y, so we can take the expectation on both sides to get that for any coupling of X and Y,

$$\mathbb{E}[||X|^k - |Y|^k|] \leq k \, \mathbb{E}[|X - Y| \, (|X| \vee |Y|)^{k-1}].$$

Applying Hölder's inequality with the conjugate exponents p and $q = p/(p-1)$, we have

$$k \, \mathbb{E}[|X - Y| \, (|X| \vee |Y|)^{k-1}] \leq k \, \|X - Y\|_p \, \|(|X| \vee |Y|)^{k-1}\|_q.$$

In particular, this inequality holds for the optimal coupling, so we can replace the right-hand side with

$$k \, W_p(\mu, \nu) \, \|(|X| \vee |Y|)^{k-1}\|_q.$$

But now it is easy to show that $|X| \vee |Y|$ is sub-exponential and there exists a constant c_r such that $\|Z\|_r \leq c_r r$ for sub-exponential Z and $r \geq 1$. Therefore, we have that

$$\|(|X| \vee |Y|)^{k-1}\|_q \leq (c_p(k-1))^{k-1} \leq (c_p k)^k,$$

which concludes the proof. □

Bounding Wasserstein distances from above is actually easier, since we only need to come up with a (potentially suboptimal) coupling and find the resulting cost.

Proposition 4.3.5. *Suppose $\mu, \nu \in \mathcal{P}_p(\mathcal{X})$ and fix $x_0 \in \mathcal{X}$. Then, for all $p \geq 0$, we have*

$$W_p(\mu,\nu)^{(p\vee 1)} \leq (2^{p-1} \vee 1) \int d(x_0,x)^p \, d|\mu-\nu|(x) = (2^{p-1} \vee 1) \, \|d(x_0,\cdot)^p(\mu-\nu)\|_{\mathrm{TV}}.$$

This proposition says that the Wasserstein distance is loosely bounded above by a weighted total variation distance.

Proof. The proof is to consider the coupling where we keep all the shared mass between μ and ν fixed in place and then distribute the rest uniformly. More formally, let

$$\pi = (\mathrm{Id} \times \mathrm{Id})_\sharp \, (\mu \wedge \nu) + \frac{1}{\alpha}(\mu-\nu)_+ \times (\mu-\nu)_-,$$

where $\alpha = (\mu-\nu)_+(\mathcal{X}) = (\mu-\nu)_-(\mathcal{X})$ is the total excess mass. Hence, we have the bound

$$W_p(\mu,\nu)^{(p\vee 1)} \leq \int d(x,y)^p \, d\pi(x,y) = \frac{1}{\alpha} \int d(x,y)^p \, d(\mu-\nu)_+(x) \, d(\mu-\nu)_-(y).$$

When $p \geq 1$, we can use Jensen's inequality on the convex function $x \mapsto |x|^p$ to get

$$\begin{aligned} d(x,y)^p &\leq (d(x_0,x) + d(x_0,y))^p \\ &= \left(\frac{1}{2}(2d(x_0,x)) + \frac{1}{2}(2d(x_0,y)) \right)^p \\ &\leq \frac{1}{2}(2d(x_0,x))^p + \frac{1}{2}(2d(x_0,y))^p \\ &= 2^{p-1}(d(x_0,x)^p + d(x_0,y)^p). \end{aligned}$$

If $0 \leq p < 1$, we immediately have $d(x,y)^p \leq d(x_0,x)^p + d(x_0,y)^p$ by subadditivity of $x \mapsto |x|^p$. Hence, we deduce

$$\begin{aligned} &\frac{1}{\alpha} \int d(x,y)^p \, d(\mu-\nu)_+(x) \, d(\mu-\nu)_-(y) \\ &\quad \leq \frac{(2^{p-1} \vee 1)}{\alpha} \left(\int d(x_0,x)^p \, d(\mu-\nu)_+(x) \, d(\mu-\nu)_-(y) + \int d(x_0,y)^p \, d(\mu-\nu)_+(x) \, d(\mu-\nu)_-(y) \right) \\ &\quad = (2^{p-1} \vee 1) \left(\int d(x_0,x)^p \, d(\mu-\nu)_+(x) + \int d(x_0,y)^p \, d(\mu-\nu)_-(y) \right) \\ &\quad = (2^{p-1} \vee 1) \int d(x_0,x)^p \, d|\mu-\nu|(x). \end{aligned}$$

□

4.4 The Kantorovich duality

In this section, we develop the dual formulation of the Kantorovich problem, and show that strong duality holds under very general conditions.

4.4.1 Lower semi-continuous cost functions

Theorem 4.4.1 (Kantorovich duality)**.** *Suppose $\mathcal{X}$ and $\mathcal{Y}$ are Polish spaces with $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$, and suppose $c \geq 0$ is an l.s.c. cost function. We define the set of dual variables by*

$$\Phi_c = \{(\phi, \psi) \in L^1(\mu) \times L^1(\nu) : \phi(x) + \psi(y) \leq c(x,y) \text{ for } \mu\text{-a.e. } x \in \mathcal{X} \text{ and } \nu\text{-a.e. } y \in \mathcal{Y}\}.$$

Then the Kantorovich problem enjoys the following strong duality:

$$\inf_{\pi \in \Pi(\mu,\nu)} \int c(x,y)\, d\pi(x,y) = \sup_{(\phi,\psi)\in\Phi_c} \left\{ \int \phi(x)\, d\mu(x) + \int \psi(y)\, d\nu(y) \right\}.$$

Intuitively, suppose that a shipper offers to pick up goods from location x at cost $\phi(x)$ and deliver them to location y at cost $\psi(y)$ in a way which is beneficial to you; namely, so that your total cost through this service $(\phi(x)+\psi(y))$ is at most your cost $(c(x,y))$ of carrying the goods directly from x to y. Then, strong duality says that a clever shipper can price their service such that you pay them almost as much as you would have paid anyways.

Proof. We first show the result in the case where $\mathcal{X}$ and $\mathcal{Y}$ are compact, and c is continuous. Define the functional f_1 on $C_b(\mathcal{X}, \mathcal{Y})$ by

$$f_1(u) = \begin{cases} 0 & u(x,y) \geq -c(x,y) \\ \infty & \text{otherwise.} \end{cases}$$

Similarly, we define the functional f_2 on $C_b(\mathcal{X} \times \mathcal{Y})$ by

$$f_2(u) = \begin{cases} \int \phi(x)\, d\mu(x) + \int \psi(y)\, d\nu(y) & u(x,y) = \phi(x) + \psi(y) \\ \infty & \text{otherwise.} \end{cases}$$

Let $\mathcal{M}(\mathcal{X} \times \mathcal{Y})$ denote the set of Radon measures on $\mathcal{X} \times \mathcal{Y}$, which is the dual of $C_b(\mathcal{X}, \mathcal{Y})$ by the Riesz-Markov representation theorem; similarly, let $\mathcal{M}_+(\mathcal{X} \times \mathcal{Y})$ denote the set of nonnegative Radon measures on $\mathcal{X} \times \mathcal{Y}$. Then, f_1 and f_2 are both convex with $f_1(1) + f_2(1) < \infty$ and f_1 is continuous at 1. By the Fenchel-Rockafellar duality (Theorem 3.5.3), we have

$$\inf_{u \in C_b(\mathcal{X}\times\mathcal{Y})} \{f_1(u) + f_2(u)\} = \sup_{\pi \in \mathcal{M}(\mathcal{X}\times\mathcal{Y})} \{-f_1^*(-\pi) - f_2^*(\pi)\}.$$

The left-hand side is

$$\begin{aligned} &\inf \left\{ \int \phi(x)\, d\mu(x) + \int \psi(y)\, d\nu(y) : \phi(x) + \psi(y) \geq -c(x,y) \right\} \\ &\quad = - \sup_{(\phi,\psi)\in\Phi_c} \left\{ \int \phi(x)\, d\mu(x) + \int \psi(y)\, d\nu(y) \right\}. \end{aligned}$$

Next, we compute the Legendre-Fenchel transform of f_1:

$$f_1^*(-\pi) = \sup_{u \in C_b(\mathcal{X},\mathcal{Y})} \left\{ \int u(x,y)\, d\pi(x,y) : u(x,y) \leq c(x,y) \right\} = \begin{cases} \int c(x,y)\, d\pi(x,y) & \pi \in \mathcal{M}_+(\mathcal{X},\mathcal{Y}) \\ \infty & \text{otherwise.} \end{cases}$$

Similarly, the Legendre-Fenchel transform of f_2 is

$$f_2^*(\pi) = \begin{cases} 0 & \pi \in \Pi(\mu, \nu) \\ \infty & \text{otherwise.} \end{cases}$$

Since we can approximate functions in $L^1(\mu)$ and $L^1(\nu)$ by continuous functions, the proof is complete in the case when $\mathcal{X}$ and $\mathcal{Y}$ are compact and c is continuous. The rest of the proof is technical, since we have to carefully relax the assumptions of compactness and continuity.

First, we relax the assumption of compactness but keep the assumption that c is uniformly continuous and bounded. Suppose π_* is an optimal coupling of μ and ν with respect to c, which exists by Proposition 4.1.2. By tightness of π_* in the Polish space $\mathcal{X} \times \mathcal{Y}$ due to Ulam's lemma (Theorem 1.1.5), there is a compact set $\mathcal{X}_0 \times \mathcal{Y}_0$ such that $\pi_*(\mathcal{X}_0 \times \mathcal{Y}_0) = 1 - \delta$. Pick $\tilde{\pi}_0$ which is optimal in $\mathcal{X}_0 \times \mathcal{Y}_0$ and construct $\tilde{\pi} = \pi_*(\mathcal{X}_0 \times \mathcal{Y}_0)\, \tilde{\pi}_0 + \mathbf{1}_{(\mathcal{X}_0 \times \mathcal{Y}_0)^c}\, \pi_* \in \Pi(\mu, \nu)$, which is close to optimal. Also, by our previous result we may pick $\tilde{\phi}_0$ and $\tilde{\psi}_0$ which are dual-optimal for $\tilde{\pi}_0$. The goal is to upgrade these to functions in Φ_c which are close to optimal for the primal problem.

Note that there exists $(x_0, y_0) \in \mathcal{X} \times \mathcal{Y}$ such that $\tilde{\phi}_0(x_0) + \tilde{\psi}_0(y_0) \geq -1$ since the choice $\phi = \psi = 0$ is always feasible. Then, since replacing $(\tilde{\phi}_0, \tilde{\psi}_0) \mapsto (\tilde{\phi}_0 + \epsilon, \tilde{\psi}_0 - \epsilon)$ for $\epsilon \in \mathbb{R}$ maintains feasibility, we may assume $\tilde{\phi}_0(x_0) \geq -1/2$ and $\tilde{\psi}_0(y_0) \geq -1/2$. So for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, we have

$$\tilde{\phi}_0(x) \leq c(x, y_0) - \tilde{\psi}_0(y_0) \leq c(x, y_0) + 1/2,$$
$$\tilde{\psi}_0(y) \leq c(x_0, y) - \tilde{\phi}_0(x_0) \leq c(x_0, y) + 1/2.$$

We improve the admissible pair by using *Rüschendorf's trick*, and define

$$\overline{\phi}_0(x) = \inf_{y \in \mathcal{Y}_0} \{c(x, y) - \tilde{\psi}_0(y)\}.$$

Essentially, this is the best possible choice of ϕ if we hold $\tilde{\psi}_0$ fixed. Since $\overline{\phi}_0 \geq \tilde{\phi}_0$, we can bound it above and below by the cost function:

$$\inf_{y \in \mathcal{Y}_0} \{c(x, y) - c(x_0, y)\} - 1/2 \leq \overline{\phi}_0(x) \leq c(x, y_0) + 1/2.$$

Then, if we define $\overline{\psi}_0(y) = \inf_{x \in \mathcal{X}} \{c(x, y) - \overline{\phi}_0(x)\}$, it's easy to bound $\overline{\psi}_0$ from below and above as before and to show that $(\overline{\phi}_0, \overline{\psi}_0) \in \Phi_c$. This shows that $\overline{\phi}_0(x) \geq -\|c\|_\infty - 1/2$ and $\overline{\psi}_0(y) \geq -\|c\|_\infty - 1/2$, and putting together our estimates gives the result; namely, it follows that $(\overline{\phi}_0, \overline{\psi}_0)$ is close to being a dual pair. We then finish the proof by approximating a general l.s.c. cost function from below by bounded Lipschitz functions (as in Proposition 3.2.2) and using the monotone convergence theorem. □

Definition 4.4.1 (c-concave functions)**.** A function $f : \mathcal{X} \to \mathbb{R}$ is called c-concave if it can be written as

$$f(x) = \inf_{y \in \mathcal{Y}} \{c(x, y) - g(y)\}$$

for some function $g : \mathcal{Y} \to \mathbb{R}$. If this is the case, we say that $f = g^c$ is the c-concave conjugate of g.

The proof of Theorem 4.4.1 shows that if c is bounded, the supremum can be taken over pairs of c-concave conjugates (ϕ^{cc}, ϕ^c) for bounded functions ϕ.

4.4.2 Metric cost functions

If the cost function $c(x, y) = d(x, y)$ is a metric, we have a stronger duality theorem.

Theorem 4.4.2 (Kantorovich-Rubinstein duality)**.** *Suppose $\mathcal{X} = \mathcal{Y}$ is a Polish space with an an l.s.c. metric cost d, and define $\mathrm{Lip}_1(\mathcal{X})$ to be the set of 1-Lipschitz functions on $\mathcal{X}$. Then, we have strong duality:*

$$\inf_{\pi \in \Pi(\mu,\nu)} \int d(x,y)\, d\pi(x,y) = \sup_{\phi \in \mathrm{Lip}_1(\mathcal{X})} \left\{ \int \phi(x)\, d(\mu - \nu)(x) \right\}.$$

Proof. We define $d_n = d/(1 + d/n)$, which is a distance bounded by 1 and whose 1-Lipschitz functions are a subset of $\mathrm{Lip}_1(\mathcal{X})$. By a similar argument to the proof of the Kantorovich duality theorem (Theorem 4.4.1), it suffices to show the result when $d = d_n$ (since $d_n \uparrow d$); hence, we assume d is bounded so all 1-Lipschitz functions are bounded. By the Kantorovich duality theorem, it suffices to study the dual problem

$$\sup_{(\phi,\psi) \in \Phi_d} \left\{ \int \phi(x)\, d\mu(x) + \int \psi(y)\, d\nu(y) \right\} = \sup_{\phi \in L^1(\mu)} \left\{ \int \phi^{dd}(x)\, d\mu(x) + \int \phi^d(y)\, d\nu(y) \right\}.$$

But since ϕ^d is defined as an infimum of 1-Lipschitz functions (bounded from below), it is also 1-Lipschitz. Furthermore, we see from the definition that $\phi^{dd} = -\phi^d$, and the result follows. □

For instance, the Kantorovich-Rubinstein duality theorem holds for $W_1(\mu, \nu)$.

4.5 Brenier's theorem

Assume that c is the quadratic cost function on $\mathbb{R}^d$ and let λ denote the Lebesgue measure on $\mathbb{R}^d$. In this section, we show the astonishing result (due to Brenier) that as long as $\mu \ll \lambda$, the optimal transport map is unique and is given by the graph of the gradient of a convex function. Furthermore, *any* valid transport coupling which is the gradient of a convex function is optimal.

Lemma 4.5.1. *Suppose π_* is an optimal coupling of μ and ν with respect to the quadratic cost function $c(x, y) = \|x - y\|_2^2$. Then, the support of π_* is cyclically monotone. In particular,* $\mathrm{supp}(\pi_*)$ *is a subset of the graph of the subgradient of a proper l.s.c. convex function.*

Proof. Suppose that $\mathrm{supp}(\pi_*)$ is not cyclically monotone, so that there exist points $(x_1, y_1), \ldots, (x_m, y_m) \in \mathrm{supp}(\pi_*)$ such that

$$0 < \sum_{i=1}^{m} \langle y_i,\, x_{i+1} - x_i \rangle.$$

Rearranging the terms, this is equivalent to

$$\sum_{i=1}^{m} \|x_i - y_i\|^2 > \sum_{i=1}^{m} \|x_{i+1} - y_i\|^2.$$

By continuity of the norm, we can find neighborhoods U_i of x_i and V_i of y_i such that the inequality holds for all $x_i' \in U_i$ and $y_i' \in V_i$. The point of this is that now we can contradict optimality of π_* by writing down a better coupling π. Define Borel measures π_i on $\mathbb{R}^d \times \mathbb{R}^d$ by $\pi_i(\cdot) = \pi_*(\cdot \,|\, U_i \times V_i)$, and let $\pi_i^{(1)}$ and $\pi_i^{(2)}$ denote the marginals of π_i. Now, for any $\epsilon > 0$ we define

$$\pi = \pi_* + \frac{\epsilon}{m} \sum_{i=1}^{m} (\pi_{i+1}^{(1)} \times \pi_i^{(2)} - \pi_i).$$

Note that this definition is motivated by the analogous structure of the inequality defining cyclical monotonicity. For all $A \in \mathcal{B}(\mathbb{R}^d \times \mathbb{R}^d)$, we have

$$\pi(A) \geq \pi_*(A) - \frac{\epsilon}{m} \sum_{i=1}^{m} \pi_i(A) \geq \pi_*(A) - \frac{\epsilon\, \pi_*(A)}{m} \sum_{i=1}^{m} \frac{1}{\pi_*(U_i \times V_i)}.$$

If ϵ is chosen small enough, then we can guarantee $\pi(A) \geq 0$. Now, it is easy to check that $\pi \in \Pi(\mu, \nu)$. But then this implies that

$$\begin{aligned} &\int \|x - y\|_2^2 \, d\pi(x, y) - \int \|x - y\|_2^2 \, d\pi_*(x, y) \\ &\quad = \frac{\epsilon}{m} \sum_{i=1}^{m} \left(\int_{U_{i+1} \times V_i} \|x - y\|_2^2 \, d\pi_{i+1}^{(1)}(x) \, d\pi_i^{(2)}(y) - \int_{U_i \times V_i} \|x - y\|_2^2 \, d\pi_i(x, y) \right). \end{aligned}$$

This expression is strictly negative (since each term in the sum is strictly negative) which contradicts the optimality of π_*. The second statement follows from Rockafellar's theorem (Theorem 3.6.1). □

Note. The converse of Lemma 4.5.1 is also true: if $\pi_* \in \Pi(\mu, \nu)$ has cyclically monotone support (for $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$) then it is optimal for the quadratic cost.

In the following, let λ denote the Lebesgue measure on $\mathbb{R}^d$.

Theorem 4.5.2 (Brenier)**.** *Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ be such that $\mu \ll \lambda$. Then, the following are equivalent.*

1. *$\pi_* \in \Pi(\mu, \nu)$ is optimal for the Kantorovich problem with $c(x, y) = \|x - y\|_2^2$.*
2. *There exists a proper l.s.c. convex function $\varphi : \mathbb{R}^d \to \mathbb{R}$ such that $\pi_* = (\mathrm{Id} \times \nabla\varphi)_\sharp \, \mu$, where $\nabla\varphi$ is defined μ-almost everywhere.*
3. *The supremum in the Kantorovich duality (Theorem 4.4.1)*

$$\int \|x - y\|_2^2 \, d\pi_*(x, y) = \sup_{(\phi, \psi) \in \Phi_c} \left\{ \int \phi(x) \, d\mu(x) + \int \psi(y) \, d\nu(y) \right\}$$

is attained for $\phi(x) = \|x\|_2^2 - 2\varphi(x)$ and $\psi(y) = \|y\|_2^2 - 2\varphi^*(y)$. The dual variables ϕ and ψ are called Kantorovich potentials for (μ, ν).

Notice that Brenier's theorem shows that the solution to the Kantorovich problem is essentially unique and coincides with the solution to the Monge problem.

Proof. First, we show (1) implies (2). We know by Lemma 4.5.1 that $\operatorname{supp}(\pi_*)$ is contained in the graph of $\partial\varphi$ for some proper l.s.c. convex function φ. Since proper convex functions are λ-a.e. differentiable on the interior of their domain, it follows that $\pi_* = (\mathrm{Id} \times \nabla\varphi)_\sharp \mu$, where $\nabla\varphi$ is defined μ-almost everywhere by absolute continuity.

Next, we show that (2) implies (3). By the Fenchel-Young inequality (Theorem 3.5.1), we have

$$\phi(x) + \psi(y) = \|x\|_2^2 + \|y\|^2 - 2(\varphi(x) + \varphi^*(y)) \leq \|x\|_2^2 + \|y\|_2^2 - 2\langle x, y\rangle = \|x - y\|_2^2.$$

In fact, by the equality condition of the Fenchel-Young inequality, we find that

$$\begin{aligned}\int \phi(x)\, d\mu(x) + \int \psi(y)\, d\nu(y) &= \int (\phi(x) + \psi(y))\, d\pi_*(x, y) \\ &= \int (\phi(x) + \psi(\nabla\varphi(x)))\, d\mu(x) \\ &= \int \|x - y\|_2^2\, d\pi_*(x, y).\end{aligned}$$

so ϕ and ψ are dual-optimal. All that's left is to show that $\phi \in L^1(\mu)$ and $\psi \in L^1(\nu)$, which will imply that $(\phi, \psi) \in \Phi_c$. Note that $\varphi = \varphi^{**}$ by the Fenchel-Moreau theorem (Theorem 3.5.2), so φ^* and $\varphi = \varphi^{**}$ are bounded below by affine functions (meaning that ϕ_+ and ψ_+ are integrable). Since

$$\int \phi(x)\, d\mu(x) + \int \psi(y)\, d\nu(y) = \int \|x - y\|_2^2\, d\pi_*(x, y) \geq 0,$$

it follows that ϕ and ψ are integrable with respect to μ and ν respectively.

Finally, (3) implies (1) is obvious from the Kantorovich duality (Theorem 4.4.1) and that for any $\pi \in \Pi(\mu, \nu)$, we have by the Fenchel-Young inequality (Theorem 3.5.1) that

$$\int \|x - y\|_2^2\, d\pi_*(x, y) = \int \phi(x)\, d\mu(x) + \int \psi(y)\, d\nu(y) \leq \int \|x - y\|_2^2\, d\pi(x, y). \qquad \square$$

This theorem allows us to easily construct optimal transport maps for the quadratic cost.

Corollary 4.5.2.1. *If $\mu, \nu \in \mathcal{P}_2(\mathbb{R})$ are such that $\mu \ll \lambda$, then the optimal transport map is $F_\nu^{-1} \circ F_\mu$, where F^{-1} denotes the quantile transformation.*

Proof. It's clear that $(F_\nu^{-1} \circ F_\mu)_\sharp \mu = \nu$ by the probability integral transform, and $F_\nu^{-1} \circ F_\mu$ is monotone so it is the derivative of a convex function; for instance, this follows from Rockafellar's theorem (Theorem 3.6.1). Now apply Brenier's theorem (Theorem 4.5.2) to see that it is optimal. $\square$

4.6 Bures-Wasserstein distances

In this section, we discuss the 2-Wasserstein distance between Gaussian measures on $\mathbb{R}^d$ and the resulting optimal coupling.

Theorem 4.6.1. *If $\mu = \mathcal{N}(m_1, \Sigma_1)$ and $\nu = \mathcal{N}(m_2, \Sigma_2)$ are two Gaussian measures on $\mathbb{R}^d$ with $\Sigma_1 \succ 0$ and $\Sigma_2 \succ 0$, then the optimal transport map for the quadratic cost is given by*

$$T(x) = \Sigma_1^{-1/2}(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\Sigma_1^{-1/2}(x - m_1) + m_2,$$

which induces the Wasserstein distance

$$W_2(\mu, \nu) = \sqrt{\|m_1 - m_2\|_2^2 + \operatorname{tr}\left(\Sigma_1 + \Sigma_2 - 2\,(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\right)}.$$

If $m_1 = m_2$, this is called the *Bures-Wasserstein distance* between positive semidefinite matrices Σ_1 and Σ_2. One option is to use Brenier's theorem to prove Theorem 4.6.1.

Proof (Knott and Smith (1984)). We make the ansatz that the transport map $T(x) = Ax + b$ is affine. Furthermore, we pick $A \succeq 0$ so that T is the gradient of a convex function, which is required by Brenier's theorem (Theorem 4.5.2). There are no constraints on b, so we may assume w.l.o.g. that $m_1 = m_2$. Then, since we need

$$\Sigma_2 = \int (Ax)(Ax)^\top \, d\mu(x) = A\Sigma_1 A^\top,$$

this forces

$$\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2} = \Sigma_1^{1/2}(A\Sigma_1 A^\top)\Sigma_1^{1/2} = (\Sigma_1^{1/2}A\Sigma_1^{1/2})^2.$$

Solving for A yields the optimal transport plan, and then some algebra gives the Wasserstein distance. □

Another option is to give a more direct proof of Theorem 4.6.1 by expanding the cost function, without relying on Brenier's theorem.

Proof (modified from Givens and Shortt (1984)). We begin by reducing to the case of centered measures; this step works generally for the 2-Wasserstein distance and does not require the assumption that μ and ν are Gaussian. Define $\tilde{\mu} = (\cdot - m_1)_\sharp \mu$ and $\tilde{\nu} = (\cdot - m_2)_\sharp \nu$. Then, for any coupling $\pi \in \Pi(\mu, \nu)$, define $\tilde{\pi} = ((\cdot - m_1), (\cdot - m_2))_\sharp \pi$ so that

$$\begin{aligned}\int \|x - y\|_2^2 \, d\pi(x, y) &= \int \|(x + m_1) - (y + m_2)\|_2^2 \, d\tilde{\pi}(x, y) \\ &= \int \|x - y\|_2^2 \, d\tilde{\pi}(x, y) + \int \langle x - y,\, m_1 - m_2 \rangle \, d\tilde{\pi}(x, y) + \|m_1 - m_2\|_2^2.\end{aligned}$$

The middle term vanishes since $\tilde{\pi}$ is centered, so it is clear that

$$W_2(\mu,\nu)^2 = \inf_{\pi\in\Pi(\mu,\nu)} \int \|x-y\|_2^2 \, d\pi(x,y) = \inf_{\pi\in\Pi(\tilde{\mu},\tilde{\nu})} \int \|x-y\|_2^2 \, d\pi(x,y) + \|m_1 - m_2\|_2^2$$

and it suffices to consider the case where $m_1 = m_2 = 0$. In this case, we obtain

$$W_2(\mu,\nu)^2 = \inf_{\pi\in\Pi(\mu,\nu)} \int \|x-y\|_2^2 \, d\pi(x,y) = \int \|x\|_2^2 \, d\mu(x) + \int \|y\|_2^2 \, d\nu(y) - 2 \sup_{\pi\in\Pi(\mu,\nu)} \int \langle x, y\rangle \, d\pi(x,y).$$

Letting $C := \mathbb{E}_{(X,Y)\sim\pi}[XY^\top]$, the 2-Wasserstein distance is the optimal value of the following semidefinite program:

$$\begin{aligned} \inf_{C\in\mathbb{R}^{d\times d}} \quad & \operatorname{tr}(\Sigma_1 + \Sigma_2 - 2C) \\ \text{s.t.} \quad & \begin{bmatrix} \Sigma_1 & C \\ C^\top & \Sigma_2 \end{bmatrix} \succeq 0. \end{aligned}$$

Since the only constraint is on the covariance of the coupling π, we may assume without loss of generality that π is a Gaussian coupling. Because $\Sigma_2 \succ 0$, the constraint is equivalent to the Schur complement constraint $\Sigma_1 - C\Sigma_2^{-1}C^\top \succeq 0$. The Lagrangian for this problem is

$$\mathcal{L}(C,\Lambda) = \operatorname{tr}(\Sigma_1 + \Sigma_2 - 2C) - \operatorname{tr}(\Lambda(\Sigma_1 - C\Sigma_2^{-1}C^\top))$$

for $\Lambda \succeq 0$. Finally, we use the KKT conditions to characterize the solution. By stationarity, we have

$$J_C \, \mathcal{L}(C^*,\Lambda^*) = -2I_d + 2\Lambda^* C^* \Sigma_2^{-1} = 0 \implies \Lambda^* C^* = \Sigma_2.$$

Complementary slackness (together with stationarity) gives

$$\Lambda^*(\Sigma_1 - C^*\Sigma_2^{-1}(C^*)^\top) = 0 \implies \Lambda^*\Sigma_1 = \Lambda^* C^* \Sigma_2^{-1}(C^*)^\top \implies \Lambda^*\Sigma_1 = (C^*)^\top \implies C^* = \Sigma_1\Lambda^*.$$

Plugging this back in to the stationarity condition, we find that $\Lambda^*\Sigma_1\Lambda^* = \Sigma_2$. Multiplying on the left and right by $\Sigma_1^{1/2}$ yields the equation

$$(\Sigma_1^{1/2}\Lambda^*\Sigma_1^{1/2})^2 = \Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2},$$

and solving for Λ^*, we find

$$\Lambda^* = \Sigma_1^{-1/2}(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\Sigma_1^{-1/2}.$$

By complementary slackness, we obtain the solution

$$C^* = \Sigma_1\Lambda^* = \Sigma_1^{1/2}(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\Sigma_1^{-1/2},$$

which fully characterizes the optimal coupling π^*. Using the cyclic property of trace yields the Wasserstein distance

$$\begin{aligned} W_2(\mu,\nu)^2 &= \operatorname{tr}(\Sigma_1 + \Sigma_2 - 2C^*) \\ &= \operatorname{tr}(\Sigma_1 + \Sigma_2 - 2\Sigma_1^{1/2}(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\Sigma_1^{-1/2}) \\ &= \operatorname{tr}(\Sigma_1 + \Sigma_2 - 2(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}), \end{aligned}$$

proving the theorem. □

4.7 Gromov-Wasserstein alignment

In this section, we motivate and formally define Gromov-Wasserstein distances (Mémoli (2011)), which provide a principled method to compare two metric measure spaces; in particular, we can use these to compare two probability distributions (possibly over different spaces) in a way that respects their underlying geometry.

We begin motivating Gromov-Wasserstein distances by asking the question: how can one compare two compact sets (shapes) in a metric space? One principled method is to use the Hausdorff distance (recall that $\operatorname{dist}(a, B) := \inf_{b\in B} d(a, b)$).

Definition 4.7.1 (Hausdorff distance)**.** The *Hausdorff distance* between compact subsets A and B of a metric space $(\mathcal{X}, d)$ is defined as

$$d_H(A,B) = \max\left\{\sup_{a\in A} \operatorname{dist}(a,B),\ \sup_{b\in B} \operatorname{dist}(b,A)\right\}.$$

However, this notion of distance is overly sensitive to minor intricacies in these shapes; for instance, two shapes may be mostly identical except for a thin spike in the second one, but the Hausdorff distance will detect this and label the two sets as far apart. One way to make the Hausdorff distance less stringent, as well as to make these distances faster to compute, is to consider an L^p relaxation. In particular, we can place probability measures μ_A and μ_B over A and B respectively (which intuitively measure relative importance of features in each shape) and compute the p-Wasserstein distance between these measures.

The Hausdorff distance is a metric on the set of compact subsets of a metric space $(\mathcal{X}, d)$. Then, we can define a notion of distance between any two compact metric spaces by minimizing over all isometric embeddings of the two spaces; this process is sometimes called *Gromovization*.

Definition 4.7.2 (Gromov-Hausdorff (GH) distance)**.** The *Gromov-Hausdorff* (GH) distance between two compact metric spaces $(\mathcal{X}, d_\mathcal{X})$ and $(\mathcal{Y}, d_\mathcal{Y})$ is

$$\operatorname{GH}(\mathcal{X},\mathcal{Y}) = \inf_{\iota_\mathcal{X},\iota_\mathcal{Y}} d_H(\iota_\mathcal{X}(\mathcal{X}), \iota_\mathcal{Y}(\mathcal{Y})),$$

where the infimum is taken over all isometric embeddings $\iota_{\mathcal{X}} : \mathcal{X} \to \mathcal{Z}$ and $\iota_{\mathcal{Y}} : \mathcal{Y} \to \mathcal{Z}$ for some shared metric space $(\mathcal{Z}, d_{\mathcal{Z}})$.

Intuitively, the infimum is doing the work of alignment, while the Hausdorff distance measures the distance between the resulting compact sets. However, it is not clear at all how to compute the GH distance; this motivates the Gromov-Wasserstein (GW) distance as an L^p relaxation of the GH distance. Of course, this means that the GW distance is a less stringent version of the GH distance. The relationships between these distances are roughly shown in Figure 4.1, taken from Mémoli (2011).

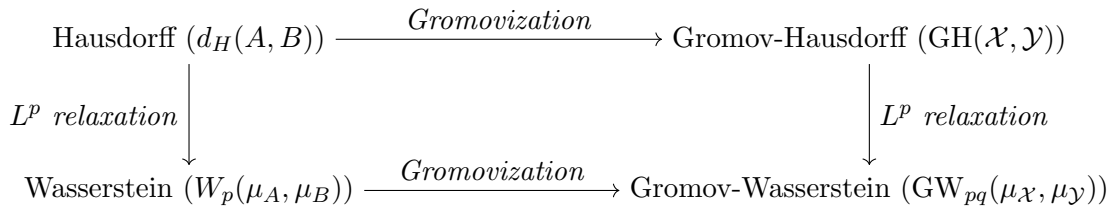

Figure 4.1: Relationships between various distances on spaces of objects and distributions.

We begin with a relevant definition. Recall that a *locally finite* measure space is one where every point has an open neighborhood of finite measure. If (X, d) is a metric space, we let $\mathcal{B}(X)$ denote the Borel σ-algebra on X induced by d.

Definition 4.7.3 (Metric measure space)**.** A *metric measure space* (m.m. space) is a tuple $(\mathcal{X},\, d,\, \mu)$ such that $(\mathcal{X}, d)$ is a Polish space and $(\mathcal{X},\, \mathcal{B}(\mathcal{X}),\, \mu)$ is a locally finite measure space.

Here are a few important examples of metric measure spaces.

Example 4.7.1 (Euclidean space)**.** The space $(\mathbb{R}^d,\, \|\cdot\|_p,\, \lambda)$ is an m.m. space, where $p \geq 1$ and λ denotes the Lebesgue measure on $\mathbb{R}^d$.

Example 4.7.2 (Random variables)**.** If X is an $\mathbb{R}^d$-valued random variable with distribution μ, then $(\mathbb{R}^d,\, \|\cdot\|_p,\, \mu)$ is an m.m. space for $p \geq 1$.

We are now ready to define the Gromov-Wasserstein distance between two (not necessarily compact) m.m. spaces. Recall that $\Pi(\mu_{\mathcal{X}},\, \mu_{\mathcal{Y}})$ denotes the set of couplings between $\mu_{\mathcal{X}}$ and $\mu_{\mathcal{Y}}$.

Definition 4.7.4 (Gromov-Wasserstein distance)**.** Let $(\mathcal{X},\, d_{\mathcal{X}},\, \mu_{\mathcal{X}})$ and $(\mathcal{Y},\, d_{\mathcal{Y}},\, \mu_{\mathcal{Y}})$ be two m.m. spaces. Given $p, q \in [1, \infty)$, the *pq-Gromov-Wasserstein (GW) distance* between $\mu_{\mathcal{X}}$ and $\mu_{\mathcal{Y}}$ is defined as

$$\begin{aligned}\mathrm{GW}_{pq}(\mu_{\mathcal{X}},\, \mu_{\mathcal{Y}}) &= \inf_{\pi \in \Pi(\mu_{\mathcal{X}},\, \mu_{\mathcal{Y}})} \left(\int |d_{\mathcal{X}}(x, x')^q - d_{\mathcal{Y}}(y, y')^q|^p \, d(\pi \times \pi)(x, y, x', y') \right)^{1/p} \\ &= \inf_{\pi \in \Pi(\mu_{\mathcal{X}},\, \mu_{\mathcal{Y}})} \|d_{\mathcal{X}}^q - d_{\mathcal{Y}}^q\|_{L^p(\pi \times \pi)},\end{aligned}$$

The GW distance looks for the measure coupling π between two m.m. spaces which is as close to an isometry as possible. The GW distance independently samples two pairs (X, Y) and (X', Y') according to π and minimizes the pth moment of the deviation of the qth powers of distances in $\mathcal{X}$ and $\mathcal{Y}$. Although the pq-GW distance looks similar in flavor to the p-Wasserstein distance, note that the objective function is no longer linear in π; this makes the GW distance more difficult to analyze and compute in practice.

However, there is a useful duality theory for quadratic GW distances under the Euclidean distance, which is discussed in Zhang et al. (2024b). Expanding the square, we find that

$$\begin{aligned}
&\mathrm{GW}_{2,2}(\mu,\nu)^2 \\
&\quad= \left(\int \|x-x'\|_2^4 \, d(\mu\times\mu)(x,x') + \int \|y-y'\|_2^4 \, d(\nu\times\nu)(y,y') - 4\int \|x\|_2^2 \, \|y\|_2^2 \, d(\mu\times\nu)(x,y)\right) \\
&\qquad+ \left(\inf_{\pi\in\Pi(\mu,\nu)} \left\{-4\int \|x\|_2^2 \, \|y\|_2^2 \, d\pi(x,y) - 8\sum_{i=1}^{d_\mu}\sum_{j=1}^{d_\nu}\left(\int x_i y_j \, d\pi(x,y)\right)^2\right\}\right).
\end{aligned}$$

The first term, which we call $S_1(\mu,\nu)$, has no dependence on the choice of coupling π. We call the second term (which does depend on the coupling) $S_2(\mu,\nu)$.

Theorem 4.7.1 (Gromov-Wasserstein duality)**.** *Given $A \in \mathbb{R}^{d_\mu \times d_\nu}$, define $\mathcal{T}_A(\mu,\nu)$ to be the optimal value of the Kantorovich problem between μ and ν under the cost*

$$c_A(x,y) = -4\,\|x\|_2^2\,\|y\|_2^2 - 32 x^\top A y.$$

Then we have the following strong duality, for S_2 defined as above:

$$S_2(\mu,\nu) = \inf_{A\in\mathbb{R}^{d_\mu\times d_\nu}} \{32\,\|A\|_F^2 + \mathcal{T}_A(\mu,\nu)\}.$$

Furthermore, the minimum is achieved at some

$$A^* \in \left[-\frac{1}{2}\sqrt{\mathbb{E}_{(X,Y)\sim\mu\times\nu}[X^2\,Y^2]},\ \frac{1}{2}\sqrt{\mathbb{E}_{(X,Y)\sim\mu\times\nu}[X^2\,Y^2]}\right]^{d_\mu\times d_\nu}.$$

Suppose $\mathcal{H}$ is a Hilbert space; we can then define the inner product Gromov-Wasserstein distance between Borel probability measures on $\mathcal{H}$ with finite second moment.

Definition 4.7.5 (Inner product Gromov-Wasserstein distances)**.** Let $\mu,\nu \in \mathcal{P}_2(\mathcal{H})$. Then, the *inner product Gromov-Wasserstein distance* (IGW) between μ and ν is

$$\begin{aligned}
\mathrm{IGW}(\mu,\nu) &= \inf_{\pi\in\Pi(\mu,\nu)} \left(\int (\langle x,x'\rangle - \langle y,y'\rangle)^2 \, d(\pi\times\pi)(x,y,x',y')\right)^{1/2} \\
&= \inf_{\pi\in\Pi(\mu,\nu)} \|\langle X,X'\rangle - \langle Y,Y'\rangle\|_{L^2(\pi\times\pi)}.
\end{aligned}$$

The existence of an optimal coupling for the IGW problem was shown in Vayer (2020). Then, IGW is a metric on the set of measures over $\mathcal{H}$ with finite second moment, modulo unitary transformations.

Theorem 4.7.2 (Zhang et al. (2024a), Proposition 3.1)**.** *The IGW distance is a metric on equivalence classes of $\mathcal{P}_2(\mathcal{H})$, where $\mu \equiv \nu$ if there exists a $(\mu \times \mu)$-almost sure unitary transformation $\iota : \mathcal{H} \to \mathcal{H}$ with $\iota_\sharp \mu = \nu$, in the sense that $\langle x, y \rangle = \langle \iota(x), \iota(y) \rangle$ for $(\mu \times \mu)$-almost every $x, y \in \mathcal{H}$.*

Chapter 5

Probability flows

In this chapter, we explore flows in the space of probability measures, which includes Wasserstein gradient flows and diffusion processes. Some material will come from Chewi et al. (2025), and the rest will be taken from various other sources.

5.1 The continuity equation

First, we describe the continuity equation, which roughly says that the change in density at a point is equal to the net flux of probability mass into that point. Suppose $X_0 \sim \mu_0$ and X_t evolves according to the dynamics

$$dX_t = v_t(X_t)\,dt,$$

where $v_t : \mathbb{R}^d \to \mathbb{R}^d$ is a time-dependent vector field. Implicitly, we assume that v_t is such that the ODE has a solution for all time $t \geq 0$; for instance, this will happen if $\|v_t\|_{\mathrm{Lip}}$ is uniformly bounded over time by the Picard-Lindelöf theorem. Let μ_t denote the distribution of X_t and let ρ_t denote the density of μ_t, which we assume exists for all $t \geq 0$.

We avoid technicalities regarding the regularity of ρ_t and v_t for now, so that we don't obfuscate the results. If $\varphi \in C_c^\infty(\mathbb{R}^d)$ is a test function, we get

$$\partial_t \mathbb{E}[\varphi(X_t)] = \mathbb{E}[\partial_t \varphi(X_t)] = \mathbb{E}[\nabla\varphi(X_t)^\top v_t(X_t)].$$

Using the definition of μ_t, we can rewrite this as

$$\partial_t \int \varphi\,d\mu_t = \int (\nabla\varphi)^\top v_t\,d\mu_t(x). \tag{5.1}$$

This is the *weak form* of the continuity equation, but we state the usual form below.

Theorem 5.1.1 (Continuity equation)**.** *The density ρ_t satisfies the* continuity equation

$$\partial_t \rho_t + \nabla \cdot (\rho_t v_t) = 0$$

if it exists and is sufficiently regular; otherwise, the weak form (5.1) holds.

Proof. For any test function $\varphi \in C_c^\infty(\mathbb{R}^d)$,

$$\int \varphi \, \partial_t \rho_t = \partial_t \int \varphi \rho_t = \int (\nabla \varphi)^\top \rho_t v_t$$

by the weak form of the continuity equation (5.1). Recall the product rule for divergence:

$$\nabla \cdot (\varphi \rho_t v_t) = (\nabla \varphi)^\top \rho_t v_t + \varphi \, \nabla \cdot (\rho_t v_t).$$

The integral of the first term vanishes by the divergence theorem since φ has compact support, so we get

$$\int \varphi \, \partial_t \rho_t = - \int \varphi \, \nabla \cdot (\rho_t v_t)$$

Since this equality is supposed to hold for all test functions φ, the result follows. □

The perspective $dX_t = v_t(X_t)\, dt$ is called the *Lagrangian* perspective, while the continuity equation $\partial_t \rho_t + \nabla \cdot (\rho_t v_t) = 0$ is called the *Eulerian* perspective. Going from the Eulerian to the Lagrangian perspective is more subtle. For example, if $\mu_t = \mathcal{N}(0, I_d)$ for all $t \geq 0$, then the Lagrangian dynamics could either be described by all particles staying still or rotating around the origin.

Somehow, we would like to pick a "movement-minimizing" flow, since this is the best explanation for the Eulerian dynamics that we are given. Formally, we want to pick a velocity field which minimizes the kinetic energy of the flow, defined as $\|v_t\|^2_{L^2(\mu_t)}$. We expect that such a velocity field must intuitively be curl-free, because any swirling is wasteful. First, we need a relevant definition.

Definition 5.1.1 (Metric derivative)**.** The *metric derivative* of a curve $(x_t)_{t \geq 0}$ in a metric space $(\mathcal{X}, d)$ is

$$|\dot{x}|_t \coloneqq \lim_{s \to t} \frac{d(x_s, x_t)}{|s - t|},$$

provided the limit exists (along paths where $s \neq t$).

Now, the following theorem connects the Eulerian and Lagrangian perspectives using optimal transport. Let $T_{\mu \to \nu}$ denote the Brenier map from μ to ν (Theorem 4.5.2).

Theorem 5.1.2. *If $(\mu_t)_{t \geq 0}$ is a curve of absolutely continuous measures (w.r.t. Lebesgue measure) in $\mathcal{P}_2(\mathbb{R}^d)$ such that the metric derivative $|\dot{\mu}|_t$ exists for all $t \geq 0$, then*

(i) For all velocity fields $(v_t)_{t\geq 0}$ satisfying the continuity equation (Theorem 5.1.1), we have

$$|\dot{\mu}|_t \leq \|v_t\|_{L^2(\mu_t)},$$

where the metric derivative is with respect to the 2-Wasserstein distance.

(ii) There exists a unique velocity field $(v_t)_{t\geq 0}$ satisfying the continuity equation with equality in (i). This field is given by

$$v_t = \lim_{h\downarrow 0} \frac{T_{\mu_t \to \mu_{t+h}} - \mathrm{Id}}{h},$$

where the limit is in $L^2(\mu_t)$.

It's worth noting that this theorem holds even if the measures μ_t are not absolutely continuous for the Lebesgue measure, in which case we consider the weak form of the continuity equation and the formula for the optimal v_t in (ii) will change. Furthermore, we only need the metric derivative to exist for *almost every* $t \geq 0$ for the theorem to hold.

Proof. The inequality in (i) follows immediately by considering the suboptimal coupling (X_t, X_{t+h}) where $X_t \sim \mu_t$ and $\partial_s X_s = v_s(X_s)$ follows the flow. In this case, we have

$$\frac{1}{h^2} W_2(\mu_t, \mu_{t+h})^2 \leq \frac{1}{h^2} \mathbb{E}[\|X_{t+h} - X_t\|_2^2]$$

and taking the limit as $h \downarrow 0$ gives the result. If we optimally couple μ_t and μ_{t+h}, the inequality becomes an equality and v_t will be of the form given in (ii). The last part of the proof is showing that v_t is well-defined and unique, which essentially follows from the fact that v_t is conservative and solves a strictly convex optimization problem. We omit the full proof here for clarity. □

This theorem shows that there is a unique way to go from the Eulerian to the Lagrangian perspective which minimizes movement, and the resulting velocity field is the one which locally moves particles along optimal transport maps.

5.2 The Benamou-Brenier formula

In this section, we present the Benamou-Brenier formula, which gives a dynamic formulation of the optimal transport problem for the quadratic cost.

Theorem 5.2.1 (Benamou-Brenier formula)**.** *Suppose $\mu_0, \mu_1 \in \mathcal{P}_2(\mathbb{R}^d)$ are absolutely continuous with respect to the Lebesgue measure. Then,*

$$W_2(\mu_0, \mu_1)^2 = \inf\left\{ \int_0^1 \|v_t\|^2_{L^2(\mu_t)}\, dt : \partial_t \rho_t + \nabla \cdot (\rho_t v_t) = 0 \right\},$$

where the infimum is over all curves $(\mu_t)_{t\in[0,1]}$ and velocity fields $(v_t)_{t\in[0,1]}$. The optimal curve is unique and defined by the displacement interpolation $\mu_t = ((1-t)\,\mathrm{Id} + t\,T_{\mu_0\to\mu_1})_\sharp\,\mu_0$.

The Benamou-Brenier theorem actually does not require μ_0 and μ_1 to be absolutely continuous, and holds for any two measures in $\mathcal{P}_2(\mathbb{R}^d)$ if we consider the weak form of the continuity equation (although the definition of the displacement interpolation must slightly change).

Proof. Let $X_0 \sim \mu_0$ with $dX_t = v_t(X_t)\,dt$, so that

$$W_2(\mu_0,\mu_1)^2 \le \mathbb{E}[\|X_1 - X_0\|_2^2] = \mathbb{E}\left[\left\|\int_0^1 v_t(X_t)\,dt\right\|_2^2\right] \le \mathbb{E}\left[\int_0^1 \|v_t(X_t)\|_2^2\,dt\right] = \int_0^1 \|v_t\|_{L^2(\mu_t)}^2\,dt.$$

For equality, we need (X_0, X_1) to be optimally coupled for the quadratic cost and $dX_t = v_t(X_t)\,dt = (X_1 - X_0)\,dt$ to be constant in time. Since (X_0, X_1) is optimally coupled, we have $X_1 = \nabla\varphi(X_0)$ for some convex function φ by Theorem 4.5.2. Defining

$$\varphi_t(x) \coloneqq (1-t)\,\frac{\|x\|_2^2}{2} + t\,\varphi(x)$$

gives $\nabla\varphi_t(x) = (1-t)\,x + t\,\nabla\varphi(x)$ and hence $X_t = \nabla\varphi_t(X_0)$. It's clear that φ_t is strictly convex so $\nabla\varphi_t$ is invertible. If we define

$$v_t \coloneqq (\varphi - \mathrm{Id}) \circ (\nabla\varphi_t)^{-1},$$

we get

$$v_t(X_t) = (\varphi - \mathrm{Id})(X_0) = X_1 - X_0,$$

and the result follows. □

The Benamou-Brenier formula shows that the squared 2-Wasserstein distance can be interpreted as the minimum total kinetic energy required to move one probability measure to another. The optimal flow is given by the displacement interpolation, which moves particles along straight lines in the direction of the optimal transport map.

5.3 Wasserstein gradient flows

In this section, we place a Riemannian structure on the Wasserstein space $\mathcal{P}_2(\mathbb{R}^d)$, following the work of Jordan, Kinderlehrer, and Otto in the late 1990s. Although $\mathcal{P}_2(\mathbb{R}^d)$ is not a bona fide Riemannian manifold, we can still define tangent spaces and a formal Riemannian metric. First, we define the tangent space at a measure $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ as

$$T_\mu\,\mathcal{P}_2(\mathbb{R}^d) \coloneqq \overline{\{\nabla\psi : \psi \in C_c^\infty(\mathbb{R}^d)\}},$$

where the closure is taken over $L^2(\mu)$. Intuitively, the tangent space consists of all conservative velocity fields, which makes sense because we know that optimal velocity fields must be curl-free. Then, we place the $L^2(\mu)$ inner product on the tangent space:

$$\langle \nabla\psi_1, \nabla\psi_2 \rangle_{T_\mu \mathcal{P}_2(\mathbb{R}^d)} \coloneqq \int (\nabla\psi_1)^\top (\nabla\psi_2)\, d\mu.$$

From the Benamou-Brenier formula (Theorem 5.2.1), we see that the constant-speed geodesic from μ_0 to μ_1 is given by the displacement interpolation

$$\mu_t = ((1-t)\,\mathrm{Id} + t\, T_{\mu_0 \to \mu_1})_\sharp\, \mu_0.$$

Therefore, the *logarithmic map* is $\log_\mu(\nu) = T_{\mu\to\nu} - \mathrm{Id}$, which gives the initial velocity field of the geodesic from μ to ν. The *exponential map* is then given by $\exp_\mu(\nabla\psi) = (\mathrm{Id} + \nabla\psi)_\sharp\, \mu$, which gives the endpoint at $t = 1$ of the geodesic starting at μ with initial velocity field $\nabla\psi$. Clearly, the exponential map is only defined for velocity fields $\nabla\psi$ such that $\mathrm{Id} + \nabla\psi$ is the gradient of a convex function, meaning that all of the eigenvalues of $\nabla^2\psi$ are at least -1. Next, we derive a formula for the Wasserstein gradient of a functional.

Definition 5.3.1 (First variation)**.** The *first variation* of a functional $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ at a measure μ is a function $\frac{\delta F}{\delta \mu} : \mathbb{R}^d \to \mathbb{R}$ such that, for any signed measure χ with $\mu + \epsilon\chi \in \mathcal{P}_2(\mathbb{R}^d)$ for sufficiently small $\epsilon > 0$,

$$\lim_{\epsilon \downarrow 0} \frac{F(\mu + \epsilon\chi) - F(\mu)}{\epsilon} = \int \frac{\delta F}{\delta \mu}\, d\chi.$$

Intuitively, the first variation $\frac{\delta F}{\delta \mu}(x)$ measures how much the functional F changes when we place a small amount of mass at the point x. The first variation is a Gâteaux derivative in the space of measures, such that the directional derivative of F at μ in the direction χ is given by the evaluation of the first variation (viewed as a linear functional on the space of signed measures) on χ. Using this, we can derive the Wasserstein gradient of F.

Proposition 5.3.1. *The Wasserstein gradient (the Riemannian gradient with respect to the geometry defined above) of a functional $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ at a measure μ is given by*

$$\nabla_{\mathcal{W}} F(\mu) = \nabla \left(\frac{\delta F}{\delta \mu} \right).$$

Proof. Fixing a curve $(\mu_t)_{t \geq 0}$ in $\mathcal{P}_2(\mathbb{R}^d)$, our goal is to show that

$$\partial_t F(\mu_t) = \langle \nabla_{\mathcal{W}} F(\mu_t), v_t \rangle_{T_{\mu_t} \mathcal{P}_2(\mathbb{R}^d)} = \left\langle \nabla \left(\frac{\delta F}{\delta \mu_t} \right), v_t \right\rangle_{T_{\mu_t} \mathcal{P}_2(\mathbb{R}^d)},$$

where v_t is the optimal velocity field satisfying the continuity equation for the curve $(\mu_t)_{t\geq 0}$. Let ρ_t denote the density of μ_t. Putting $\chi = \partial_t \rho_t$ in the definition of the first variation gives

$$\partial_t F(\mu_t) = \lim_{\epsilon \downarrow 0} \frac{F(\mu_t + \epsilon\, \partial_t \rho_t) - F(\mu_t)}{\epsilon} = \int \frac{\delta F}{\delta \mu_t}\, \partial_t \rho_t. \tag{5.2}$$

By the continuity equation, we get

$$\int \frac{\delta F}{\delta \mu_t}\, \partial_t \rho_t = -\int \frac{\delta F}{\delta \mu_t}\, (\nabla \cdot (\mu_t v_t)).$$

At last, the divergence theorem gives

$$-\int \frac{\delta F}{\delta \mu}\, (\nabla \cdot (\mu_t v_t)) = \int \left(\nabla \left(\frac{\delta F}{\delta \mu_t}\right)\right)^\top v_t \, d\mu_t = \left\langle \nabla \left(\frac{\delta F}{\delta \mu_t}\right),\, v_t \right\rangle_{T_{\mu_t} \mathcal{P}_2(\mathbb{R}^d)}.$$

The result follows because $\nabla \left(\frac{\delta F}{\delta \mu_t}\right)$ is a conservative vector field so it is in the tangent space $T_{\mu_t} \mathcal{P}_2(\mathbb{R}^d)$. □

Intuitively, the Wasserstein gradient of F at μ points in the direction of greatest increase of the first variation, which represents the direction in which pushing mass will increase F the most. Using this, we can define Wasserstein gradient flows, which locally move mass around to decrease a functional as quickly as possible.

Definition 5.3.2 (Wasserstein gradient flow)**.** A curve $(\mu_t)_{t\geq 0}$ in $\mathcal{P}_2(\mathbb{R}^d)$ is a *Wasserstein gradient flow* of a functional $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ if it satisfies the continuity equation with velocity field

$$v_t = -\nabla_{\mathcal{W}} F(\mu_t) = -\nabla \left(\frac{\delta F}{\delta \mu_t}\right).$$

Since the Wasserstein gradient flow is a gradient flow, we can easily show that it decreases the functional F over time:

$$\partial_t F(\mu_t) = \langle \nabla_{\mathcal{W}} F(\mu_t),\, \partial_t \rho_t \rangle_{T_{\mu_t} \mathcal{P}_2(\mathbb{R}^d)} = -\|\nabla_{\mathcal{W}} F(\mu_t)\|^2_{T_{\mu_t} \mathcal{P}_2(\mathbb{R}^d)} \leq 0.$$

Furthermore, the usual tools for analyzing gradient flows, such as the Polyak-Łojasiewicz inequality and Grönwall's inequality, can be applied to Wasserstein gradient flows. One can show that if F is strictly geodesically convex in the Wasserstein space, then the Wasserstein gradient flow converges to the unique minimizer of F at an exponential rate.

If the measure μ is singular, then the definition of the first variation of F is that it is any function satisfying

$$\lim_{\epsilon \downarrow 0} \frac{F((\mathrm{Id} + \epsilon\, \xi)_\sharp\, \mu) - F(\mu)}{\epsilon} = \int \nabla \left(\frac{\delta F}{\delta \mu}\right)^\top \xi \, d\mu$$

for all compactly supported smooth vector fields $\xi : \mathbb{R}^d \to \mathbb{R}^d$. In this case, the Wasserstein gradient is still given by $\nabla_{\mathcal{W}} F(\mu) = \nabla \left(\frac{\delta F}{\delta \mu}\right)\Big|_{\mathrm{supp}(\mu)}$. In addition, note that all of the Wasserstein gradient flow machinery works on Riemannian manifolds with minor modifications (Villani, 2009).

We can now do a couple examples of computing the Wasserstein gradient of a functional.

Example 5.3.1 (Potential energy)**.** Let $V \in C^1(\mathbb{R}^d)$ and define the *potential energy* functional $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ as

$$F(\mu) = \int V \, d\mu.$$

For a curve $(\mu_t)_{t\geq 0}$ in $\mathcal{P}_2(\mathbb{R}^d)$, we have

$$\partial_t F(\mu_t) = \int V \, \partial_t \rho_t.$$

Therefore, (5.2) gives $\frac{\delta F}{\delta \mu} = V$ and the Wasserstein gradient is $\nabla_{\mathcal{W}} F(\mu) = \nabla V$. Hence, the Wasserstein gradient flow of the potential energy functional satisfies

$$\partial_t \rho_t = \nabla \cdot (\rho_t \, \nabla V).$$

This is sometimes called the Fokker-Planck equation with zero diffusion.

Example 5.3.2 (Internal energy)**.** Let $U \in C^2((0,\infty))$ and define the *internal energy* functional $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ as

$$F(\mu) = \int U(\rho),$$

where ρ is the density of μ. For a curve $(\mu_t)_{t\geq 0}$ in $\mathcal{P}_2(\mathbb{R}^d)$ where ρ_t is the density of μ_t, we have

$$\partial_t F(\mu_t) = \int U'(\rho_t) \, \partial_t \rho_t.$$

Therefore, (5.2) gives $\frac{\delta F}{\delta \mu} = U'(\rho)$ and the Wasserstein gradient is $\nabla_{\mathcal{W}} F(\mu) = \nabla(U'(\rho))$.

Example 5.3.3 (Interaction energy)**.** Let $K \in C^1(\mathbb{R}^d \times \mathbb{R}^d)$ be a symmetric kernel. Define the *interaction energy* functional $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ as

$$F(\mu) = \frac{1}{2} \iint K(x,y) \, d(\mu \times \mu)(x,y).$$

For a curve $(\mu_t)_{t\geq 0}$ in $\mathcal{P}_2(\mathbb{R}^d)$ where ρ_t is the density of μ_t, we have

$$\partial_t F(\mu_t) = \frac{1}{2} \iint K(x-y) \, d\mu_t(y) \, \partial_t \rho_t(x).$$

Therefore, (5.2) gives $\frac{\delta F}{\delta \mu} = \int K(x,y) \, d\mu(y)$ and the Wasserstein gradient is $\nabla_{\mathcal{W}} F(\mu) = \int \nabla_x K(x,y) \, d\mu(y)$.

The *McKean-Vlasov equation* describes the Wasserstein gradient flow of the sum of all three types of functionals above.

Example 5.3.4 (Wasserstein gradient flow for entropy)**.** As an interesting example, consider the negative entropy functional $H : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ defined as

$$H(\mu) = \int \rho \log \rho,$$

where ρ is the density of μ. This is a special case of the internal energy functional with $U(\rho) = \rho \log \rho$, so the first variation is

$$\frac{\delta H}{\delta \mu} = U'(\rho) = \log(\rho) + 1.$$

Therefore, the Wasserstein gradient is

$$\nabla_{\mathcal{W}} H(\mu) = \nabla \log(\rho) = \frac{\nabla \rho}{\rho},$$

yielding the Wasserstein gradient flow

$$\partial_t \rho_t = \nabla \cdot \left(\rho_t \frac{\nabla \rho_t}{\rho_t} \right) = \Delta \rho_t.$$

We can now recognize the Wasserstein gradient flow as the heat equation! This makes sense informally; the fastest way to increase entropy (decrease negative entropy) of a measure is to follow the heat equation, which spreads out mass as quickly as possible. The first variation $\log(\rho) + 1$ can be interpreted because adding mass at a point with low density ρ increases entropy (decreases negative entropy) more than adding mass at a point with high density.

5.4 Diffusion processes

In this section, we discuss the basics of diffusion processes. We start by studying the behavior of the solution to the stochastic differential equation (SDE)

$$dX_t = v_t(X_t)\, dt + \sigma_t(X_t) \cdot dB_t,$$

where v is the drift and σ is the diffusion coefficient. As long as $X_0 \in L^2(\mathbb{P})$ and there exists $K > 0$ such that

$$\|v_t(x) - v_t(y)\|_2 + \|\sigma_t(x) - \sigma_t(y)\|_F \leq K \|x - y\|_2$$

and $\|v_t(0)\|_2 + \|\sigma_t(0)\|_F \leq K$ for all $t \geq 0$ and $x, y \in \mathbb{R}^d$, it is known that the SDE has a unique strong solution; this result was originally proven by Kiyosi Itô in the early 1940s by Picard-Lindelöf iteration and Grönwall's inequality. We start by deriving the Fokker-Planck equation, which describes the time evolution of the density of X_t.

Theorem 5.4.1 (Fokker-Planck equation)**.** *If ρ_t is the density of X_t, then it satisfies the Fokker-Planck equation*

$$\partial_t \rho_t = -\nabla \cdot (\rho_t v_t) + \frac{1}{2} \sum_{i=1}^{d} \sum_{j=1}^{d} \partial_{x_i} \partial_{x_j} (\rho_t (\sigma_t \sigma_t^\top)_{ij}).$$

Proof. For any test function $\varphi \in C_c^\infty(\mathbb{R}^d)$, Itô's formula gives

$$d\varphi(X_t) = \left(\nabla \varphi(X_t)^\top v_t(X_t) + \frac{1}{2} \sum_{i=1}^{d} \sum_{j=1}^{d} (\partial_{x_i} \partial_{x_j} \varphi(X_t)) \, (\sigma_t(X_t) \, \sigma_t(X_t)^\top)_{ij} \right) dt + \nabla \varphi(X_t)^\top \sigma_t(X_t) \, dB_t.$$

Taking expectations and using the definition of ρ_t gives

$$\partial_t \int \varphi \rho_t = \int (\nabla \varphi)^\top v_t \rho_t + \frac{1}{2} \sum_{i=1}^{d} \sum_{j=1}^{d} (\partial_{x_i} \partial_{x_j} \varphi) \, (\sigma_t \sigma_t^\top)_{ij} \, \rho_t.$$

Using the divergence theorem and the fact that φ has compact support, we get

$$\int (\nabla \varphi)^\top v_t \rho_t = - \int \varphi \, \nabla \cdot (\rho_t v_t).$$

Similarly, integrating by parts twice gives

$$\int (\partial_{x_i} \partial_{x_j} \varphi) \, (\sigma_t \sigma_t^\top)_{ij} \, \rho_t = \int \varphi \, \partial_{x_i} \partial_{x_j} (\rho_t (\sigma_t \sigma_t^\top)_{ij}).$$

Since φ was an arbitrary test function, the result follows. □

The Fokker-Planck equation can be interpreted as a continuity equation with an additional diffusion term. Next, we write down the *probability flow ODE*, which is an ordinary differential equation whose solution has the same time-marginal distributions as the SDE above. Consider a deterministic process X_t satisfying the dynamics $dX_t = \tilde{v}_t(X_t) \, dt$. We know from Theorem 5.1.1 that the density ρ_t of X_t satisfies the continuity equation $\partial_t \rho_t = -\nabla \cdot (\rho_t \tilde{v}_t)$. We compare this to the Fokker-Planck equation above

$$\partial_t \rho_t = -\nabla \cdot (\rho_t v_t) + \frac{1}{2} \nabla \cdot (\nabla \cdot (\rho_t \, \sigma_t \sigma_t^\top)).$$

and equate the things inside the divergences to get

$$\tilde{v}_t \rho_t = v_t \rho_t - \frac{1}{2} \nabla \cdot (\rho_t \, \sigma_t \sigma_t^\top) \implies \tilde{v}_t = v_t - \frac{\nabla \cdot (\rho_t \, \sigma_t \sigma_t^\top)}{2 \rho_t}.$$

Using the product rule for divergence, we obtain

$$\nabla \cdot (\rho_t \, \sigma_t \sigma_t^\top) = (\sigma_t \sigma_t^\top)(\nabla \rho_t) + \rho_t \, \nabla \cdot (\sigma_t \sigma_t^\top),$$

so that

$$\tilde{v}_t = v_t - \frac{\nabla \cdot (\rho_t \, \sigma_t \sigma_t^\top)}{2 \rho_t} = v_t - \frac{1}{2} \nabla \cdot (\sigma_t \sigma_t^\top) - \frac{1}{2} (\sigma_t \sigma_t^\top) \frac{\nabla \rho_t}{\rho_t} = v_t - \frac{1}{2} \nabla \cdot (\sigma_t \sigma_t^\top) - \frac{1}{2} (\sigma_t \sigma_t^\top) \nabla \log(\rho_t).$$

We summarize the above considerations in the following theorem.

Theorem 5.4.2 (Probability flow ODE)**.** *The probability flow ODE corresponding to the SDE*

$$dX_t = v_t(X_t)\,dt + \sigma_t(X_t)\cdot dB_t$$

is given by

$$dX_t = \left(v_t(X_t) - \frac{1}{2}\nabla\cdot(\sigma_t\sigma_t^\top)(X_t) - \frac{1}{2}(\sigma_t\sigma_t^\top)(X_t)\,\nabla\log\rho_t(X_t)\right)dt,$$

where ρ_t is the density of X_t. In particular, the time-marginal distributions of the solution to this ODE are the same as those of the SDE.

In the above formula, the term $\nabla\log\rho_t$ is called the *score function* of the density ρ_t. The probability flow ODE is particularly useful in generative modeling, since it allows us to generate samples from a target distribution by solving an ODE instead of an SDE, which is often easier and more efficient. However, it requires knowledge of the score function, which is often unknown and must be estimated from data. Finally, we derive the astonishing result that the time-reversal of a diffusion process is also a diffusion process, due to Anderson in 1982.

Theorem 5.4.3 (Anderson time-reversal)**.** *Let X_t satisfy the SDE*

$$dX_t = v_t(X_t)\,dt + \sigma_t(X_t)\cdot dB_t,$$

and let ρ_t be the density of X_t. Then, fixing $T>0$, the time-reversed process $\tilde{X}_\tau = X_{T-\tau}$ satisfies the SDE

$$d\tilde{X}_\tau = \left(v_{T-\tau}(\tilde{X}_\tau) - \nabla\cdot(\sigma_{T-\tau}\sigma_{T-\tau}^\top)(\tilde{X}_\tau) - (\sigma_{T-\tau}\sigma_{T-\tau}^\top)(\tilde{X}_\tau)\,\nabla\log\rho_{T-\tau}(\tilde{X}_\tau)\right)d\tau + \sigma_{T-\tau}(\tilde{X}_\tau)\cdot d\tilde{B}_\tau.$$

Proof. We introduce the reverse time variable $\tau = T - t$ and we are interested in the dynamics of the density $\tilde{\rho}_\tau = \rho_{T-\tau}$. Using the chain rule and the Fokker-Planck equation (Theorem 5.4.1), we get

$$\partial_\tau\tilde{\rho}_\tau = -\partial_t\rho_{T-\tau} = \nabla\cdot(\rho_{T-\tau}v_{T-\tau}) - \frac{1}{2}\sum_{i=1}^d\sum_{j=1}^d\partial_{x_i}\partial_{x_j}(\rho_{T-\tau}(\sigma_{T-\tau}\sigma_{T-\tau}^\top)_{ij}).$$

Now, we want to rearrange this equation to look like a Fokker-Planck equation; in particular, we want a positive semidefinite coefficient for the second term. To do this, we employ Anderson's trick of absorbing part of the diffusion term into the drift term:

$$\begin{aligned}\partial_\tau\tilde{\rho}_\tau &= \nabla\cdot(\rho_{T-\tau}v_{T-\tau}) - \frac{1}{2}\nabla\cdot(\nabla\cdot(\rho_{T-\tau}\sigma_{T-\tau}\sigma_{T-\tau}^\top))\\ &= \nabla\cdot\left(\rho_{T-\tau}v_{T-\tau} - \nabla\cdot(\rho_{T-\tau}\sigma_{T-\tau}\sigma_{T-\tau}^\top)\right) + \frac{1}{2}\nabla\cdot(\nabla\cdot(\rho_{T-\tau}\sigma_{T-\tau}\sigma_{T-\tau}^\top)).\end{aligned}$$

Finally, the product rule for divergence implies that

$$\nabla\cdot(\rho_{T-\tau}\sigma_{T-\tau}\sigma_{T-\tau}^\top) = (\sigma_{T-\tau}\sigma_{T-\tau}^\top)(\nabla\rho_{T-\tau}) + \rho_{T-\tau}\,\nabla\cdot(\sigma_{T-\tau}\sigma_{T-\tau}^\top),$$

so that

$$\begin{aligned}\partial_\tau \tilde{\rho}_\tau &= \nabla \cdot \left(\rho_{T-\tau} v_{T-\tau} - \rho_{T-\tau} \nabla \cdot (\sigma_{T-\tau}\sigma_{T-\tau}^\top) - (\sigma_{T-\tau}\sigma_{T-\tau}^\top) \nabla \rho_{T-\tau} \right) + \frac{1}{2} \nabla \cdot (\nabla \cdot (\rho_{T-\tau} \sigma_{T-\tau} \sigma_{T-\tau}^\top)) \\ &= \nabla \cdot \left(\rho_{T-\tau} \left(v_{T-\tau} - \nabla \cdot (\sigma_{T-\tau}\sigma_{T-\tau}^\top) - (\sigma_{T-\tau}\sigma_{T-\tau}^\top) \nabla \log \rho_{T-\tau} \right) \right) + \frac{1}{2} \nabla \cdot (\nabla \cdot (\rho_{T-\tau} \sigma_{T-\tau} \sigma_{T-\tau}^\top)).\end{aligned}$$

This is in the form of a Fokker-Planck equation, so the result follows. □

Again, the score function $\nabla \log \rho_{T-\tau}$ appears in the drift of the time-reversed SDE.

5.5 The Ohrnstein-Uhlenbeck process

In practice, modern diffusion models often consider the *Ornstein-Uhlenbeck process*, which is defined by

$$dX_t = -\beta_t X_t \, dt + \sqrt{2\beta_t} \, dB_t,$$

where $\beta_t > 0$ is a time-dependent scalar function which is such that $\int_0^\infty \beta_t \, dt = \infty$. First, we will show that the stationary distribution of this SDE is the standard Gaussian distribution.

Proposition 5.5.1. *The Ornstein-Uhlenbeck process has stationary distribution $\mathcal{N}(0, I_d)$.*

Proof. We use an integrating factor to cancel the drift term. The Itô product rule gives

$$\begin{aligned} d\left(X_t \exp\left(\int_0^t \beta_s \, ds \right) \right) &= \beta_t \exp\left(\int_0^t \beta_s \, ds \right) X_t \, dt + \exp\left(\int_0^t \beta_s \, ds \right) dX_t \\ &= \beta_t \exp\left(\int_0^t \beta_s \, ds \right) X_t \, dt - \beta_t \exp\left(\int_0^t \beta_s \, ds \right) X_t \, dt + \sqrt{2\beta_t} \exp\left(\int_0^t \beta_s \, ds \right) dB_t \\ &= \sqrt{2\beta_t} \exp\left(\int_0^t \beta_s \, ds \right) dB_t. \end{aligned}$$

In particular, this means that

$$\begin{aligned} X_t \exp\left(\int_0^t \beta_s \, ds \right) - X_0 &= \int_0^t \sqrt{2\beta_s} \exp\left(\int_0^s \beta_r \, dr \right) dB_s \\ \implies X_t &= X_0 \exp\left(-\int_0^t \beta_s \, ds \right) + \int_0^t \sqrt{2\beta_s} \exp\left(-\int_s^t \beta_r \, dr \right) dB_s. \end{aligned}$$

The first term goes to zero as $t \to \infty$ because $\int_0^\infty \beta_s \, ds = \infty$. The second term is a Gaussian random variable with mean zero. We compute its covariance using the Itô isometry:

$$\begin{aligned} &\mathbb{E}\left[\left(\int_0^t \sqrt{2\beta_s} \exp\left(-\int_s^t \beta_r \, dr \right) dB_s \right) \left(\int_0^t \sqrt{2\beta_u} \exp\left(-\int_u^t \beta_r \, dr \right) dB_u \right)^\top \right] \\ &= I_d \int_0^t 2\beta_s \exp\left(-2\int_s^t \beta_r \, dr \right) ds \\ &= I_d \exp\left(-2\int_0^t \beta_r \, dr \right) \int_0^t 2\beta_s \exp\left(2\int_0^s \beta_r \, dr \right) ds. \end{aligned}$$

Substituting $u = \exp\left(2\int_0^s \beta_r\, dr\right)$ gives $du = 2\beta_s \exp\left(2\int_0^s \beta_r\, dr\right)\, ds$, so that

$$\begin{aligned} I_d \exp\left(-2\int_0^t \beta_r\, dr\right) \int_0^t 2\beta_s \exp\left(2\int_0^s \beta_r\, dr\right)\, ds &= I_d \exp\left(-2\int_0^t \beta_r\, dr\right)\left(\exp\left(2\int_0^t \beta_r\, dr\right) - 1\right) \\ &= I_d\left(1 - \exp\left(-2\int_0^t \beta_r\, dr\right)\right). \end{aligned}$$

As $t \to \infty$, this converges to I_d. Therefore, the stationary distribution of the Ornstein-Uhlenbeck process is $\mathcal{N}(0, I_d)$ as claimed. □

Letting ρ_t denote the density of X_t, we usually approximate ρ_T for large $T > 0$ by the standard Gaussian density. The Anderson time-reversal (Theorem 5.4.3) then gives the time-reversed SDE

$$d\tilde{X}_\tau = \left(\beta_{T-\tau}\tilde{X}_\tau + 2\beta_{T-\tau}\nabla \log \rho_{T-\tau}(\tilde{X}_\tau)\right) d\tau + \sqrt{2\beta_{T-\tau}}\, d\tilde{B}_\tau.$$

The only unknown term in this SDE is the score function $\nabla \log \rho_{T-\tau}$. In practice, we estimate this score function using a neural network trained via score matching. Once we have a good estimate of the score function, we can simulate the time-reversed SDE to generate samples from the target distribution. Alternatively, we can simulate the corresponding probability flow ODE

$$d\tilde{X}_\tau = \left(\beta_{T-\tau}\tilde{X}_\tau + \beta_{T-\tau}\nabla \log \rho_{T-\tau}(\tilde{X}_\tau)\right) d\tau,$$

as in Theorem 5.4.2. This ODE is often easier to simulate than the SDE and it also produces high-quality samples. However, it has been observed in practice that simulating the SDE often yields better sample diversity compared to the ODE; this may be due to the inherent randomness in the SDE which allows it to self-correct during sampling despite errors in the score function estimate or discretization. Related, the *manifold hypothesis* suggests that real-world data often lies on a low-dimensional manifold in the ambient space, and the stochasticity in the SDE may help explore this manifold more effectively than the probability flow ODE.

In practice, the forward process is easy to sample in closed form due to the linear drift. However, we need to use a numerical solver to simulate the reverse process (turning Gaussian noise into samples from the original image distribution), such as the Euler-Maruyama method for SDEs or the Runge-Kutta method for ODEs.

5.6 Variational inference and Langevin dynamics

In variational inference, we have a measure ν which is difficult to sample from directly; in Bayesian inference, this is usually the posterior. For example, we may wish to sample from this distribution to estiamte its mean by a Monte Carlo algorithm (or any expectation over ν). Instead, we consider a family of measures

Φ (the variational family) which are easy to sample from, and we want to find the measure $\mu \in \Phi$ which is closest to ν in terms of KL divergence:

$$\arg\min_{\mu \in \Phi} \ \mathrm{KL}(\mu \,\|\, \nu).$$

In the following, we assume that ν has density $\rho_\nu \propto \exp(-V)$ for some potential function $V : \mathbb{R}^d \to \mathbb{R}$. If we let Φ denote any subset of $\mathcal{P}_2(\mathbb{R}^d)$ consisting only of absolutely continuous measures with respect to the Lebesgue measure, then suppose that μ has density ρ. Then, we can rewrite the KL divergence as

$$\begin{aligned}
\mathrm{KL}(\mu \,\|\, \nu) &= \int \rho(x) \log\left(\frac{\rho(x)}{\exp(-V(x))/Z}\right) dx \\
&= \int \rho(x) \log(\rho(x))\, dx + \int \rho(x) V(x)\, dx + \log(Z) \int \rho(x)\, dx \\
&= \int V \, d\mu + \int \rho \log(\rho) + \log(Z),
\end{aligned}$$

where $Z = \int \exp(-V(x))\, dx$ is the normalizing constant of ν. Therefore, minimizing the KL divergence is equivalent to minimizing the functional

$$F(\mu) = \int V \, d\mu + \int \rho \log(\rho).$$

Using our computations in Section 5.3 for the potential energy and negative entropy functionals, we find that the Wasserstein gradient of F is

$$\nabla_{\mathcal{W}} F(\mu) = \nabla V + \nabla \log(\rho).$$

Hence, the Wasserstein gradient flow of F satisfies the continuity equation

$$\partial_t \rho_t = \nabla \cdot (-\rho_t \nabla \log(\rho_\nu) + \rho_t \nabla \log(\rho_t)) = \nabla \cdot \left(\rho_t \nabla \log\left(\frac{\rho_t}{\rho_\nu}\right)\right).$$

If Φ is a geodesically convex subset of $\mathcal{P}_2(\mathbb{R}^d)$ and V is strongly convex, then it is easy to show that F is α-geodesically convex in Φ. Therefore, the Wasserstein gradient flow converges to the unique minimizer of F in Φ at an exponential rate. We can rewrite the Wasserstein gradient flow as

$$\partial_t \rho_t = \nabla \cdot (\rho_t \nabla V) + \Delta \rho_t,$$

which is the Fokker-Planck equation (Theorem 5.4.1) corresponding to the *Langevin dynamics* SDE

$$dX_t = -\nabla V(X_t)\, dt + \sqrt{2}\, dB_t.$$

Elegantly, this immediately implies that Langevin dynamics converges exponentially fast to the stationary distribution ν in KL divergence, as long as ν is strongly log-concave. This is a fundamental result in the theory of sampling via Langevin dynamics, and suggests the an unadjusted Langevin algorithm (ULA) which discretizes the Langevin dynamics SDE for sampling from log-concave distributions.

Chapter 6

Deep learning theory

In this chapter, we provide a brief overview of some interesting aspects of the theoretical foundations of deep learning.

6.1 Infinitely wide neural networks

In this section, we discuss the analysis of infinitely wide two-layer neural networks. One approach to studying such networks is via the neural tangent kernel (NTK) introduced by Jacot et al. (2018). Another approach is via the mean-field limit introduced independently by Chizat and Bach (2018) and Mei et al. (2018) within a few months of each other. The main difference between these two approaches is in the scaling of the parameters with respect to the width of the network. Still, note that both approaches require weights which have a very specific scaling with respect to the width of the network.

6.1.1 The neural tangent kernel and lazy training

We begin by considering the NTK approach taken by Jacot et al. (2018). Define a simple feedforward neural network with one hidden layer by

$$f(x; w, a) = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} a_i \, \sigma(w_i^\top x),$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ is a non-linear activation function and m is the width of the hidden layer. The particular normalization of $1/\sqrt{m}$ is important for the following analysis.

First, note that at initialization, if we take $w_i \sim \mathcal{N}(0, I)$ and $a_i \sim \mathcal{N}(0, 1)$ independently, then by the central limit theorem, for any fixed input x, as $m \to \infty$, we have that

$$f(x; w, a) \xrightarrow{d} \mathcal{N}\left(0, \, \mathbb{E}_{w \sim \mathcal{N}(0, I)}[\sigma(w^\top x)^2]\right)$$

as long as $\sigma(w^\top x) \in L^2(\mathbb{P})$. Furthermore, for any two inputs $x, x' \in \mathbb{R}^d$, we have

$$\mathrm{Cov}(f(x; w, a), f(x'; w, a)) = \mathbb{E}_{w \sim \mathcal{N}(0, I)}[f(x; w, a) f(x'; w, a)] = \mathbb{E}_{w \sim \mathcal{N}(0, I)}[\sigma(w^\top x)\, \sigma(w^\top x')].$$

Hence, at initialization, the network behaves like a mean-zero Gaussian process with covariance function given by the above expression. This observation was originally made by Neal (1996).

Next, consider training the network using gradient flow on the squared loss from data $\{(x_i, y_i)\}_{i=1}^n$. Denoting $\theta = (w, a)$, the gradient flow dynamics are

$$\frac{d\theta_t}{dt} = -\eta \nabla_\theta \left(\frac{1}{2n} \sum_{i=1}^n (f(x_i; \theta_t) - y_i)^2 \right),$$

where $\eta > 0$ is the learning rate. By the chain rule, the output for a fixed input x evolves according to

$$\frac{df(x; \theta_t)}{dt} = (\nabla_\theta f(x; \theta_t))^\top \frac{d\theta_t}{dt} = -\frac{\eta}{n} \sum_{i=1}^n (f(x_i; \theta_t) - y_i) \left((\nabla_\theta f(x; \theta_t))^\top (\nabla_\theta f(x_i; \theta_t)) \right).$$

So the output at a test point x evolves according to a linear combination of the loss gradients at each training point, each weighted by the term

$$(\nabla_\theta f(x; \theta_t))^\top (\nabla_\theta f(x_i; \theta_t))$$

This is why people refer to the kernel

$$\Theta_t^{(m)}(x, x') = (\nabla_\theta f(x; \theta_t))^\top (\nabla_\theta f(x'; \theta_t))$$

as the *neural tangent kernel*. Computing explicitly, we have

$$\partial_{a_i} f(x; w, a) = \frac{1}{\sqrt{m}} \sigma(w_i^\top x)$$

and

$$\partial_{w_i} f(x; w, a) = \frac{1}{\sqrt{m}} a_i\, \sigma'(w_i^\top x)\, x,$$

so that

$$\begin{aligned} \Theta_t^{(m)}(x, x') &= \sum_{i=1}^m \left(\partial_{a_i} f(x; \theta_t)\, \partial_{a_i} f(x'; \theta_t) + \partial_{w_i} f(x; \theta_t)^\top \partial_{w_i} f(x'; \theta_t) \right) \\ &= \frac{1}{m} \sum_{i=1}^m \left(\sigma((w_i)_t^\top x)\, \sigma((w_i)_t^\top x') + (a_i)_t^2\, \sigma'((w_i)_t^\top x)\, \sigma'((w_i)_t^\top x')\, x^\top x' \right). \end{aligned}$$

Note that the partial derivatives both scale as $1/\sqrt{m}$, so the parameters don't move very much as $m \to \infty$; this is called *lazy training*. As a consequence, the individual features $\sigma(w_i^\top x)$ do not evolve to learn data-dependent representations, but rather serve as a fixed basis. Therefore, if m is large enough, we expect

that $\Theta_t^{(m)}$ will be close to its value at initialization $\Theta_0^{(m)}$ for all time $t \geq 0$. By the strong law of large numbers, as $m \to \infty$, we have

$$\Theta_0^{(m)}(x, x') \xrightarrow{a.s.} \Theta(x, x') = \mathbb{E}_{w \sim \mathcal{N}(0, I)}[\sigma(w^\top x)\, \sigma(w^\top x')] + \mathbb{E}_{w \sim \mathcal{N}(0, I)}[\sigma'(w^\top x)\, \sigma'(w^\top x')]\, x^\top x'.$$

This argument can be formalized into a rigorous estimate if we place some assumptions on σ.

After taking the limit as $m \to \infty$, we obtain the following linear dynamics for the output at a test point x:

$$\frac{df(x; \theta_t)}{dt} = -\frac{\eta}{n} \sum_{i=1}^{n} (f(x_i; \theta_t) - y_i)\, \Theta(x, x_i).$$

If we write out the formula for $f(x_i; \theta_t) - y_i$ by integrating the dynamics and then integrate the resulting dynamics for $f(x; \theta_t)$, we obtain the closed-form expression

$$\lim_{t \to \infty} f(x; \theta_t) = f(x; \theta_0) + \Theta(x, X)\, \Theta(X, X)^{-1} (Y - f(X; \theta_0)),$$

which corresponds to kernel regression on the residuals with kernel Θ. Here, X and Y should be interpeted as the matrices of training inputs and outputs. Geometrically, this implies that the optimization landscape is well-approximated by the tangent plane at initialization; the nonlinear network $f(x; \theta)$ effectively behaves as its linear approximation $f(x; \theta_0) + \nabla f(x; \theta_0)^\top (\theta - \theta_0)$.

6.1.2 The mean-field regime

In the mean-field approach taken by Chizat and Bach (2018) and Mei et al. (2018), we consider a two-layer neural network of the form

$$f(x; w, a) = \frac{1}{m} \sum_{i=1}^{m} a_i\, \sigma(w_i^\top x),$$

where again $\sigma : \mathbb{R} \to \mathbb{R}$ is a non-linear activation function and m is the width of the hidden layer. Note that the normalization here is different from the NTK case; this difference in scaling leads to very different behavior in the infinite-width limit. In this case, the parameters move significantly during training, leading to feature learning. Let $\theta_i = (w_i, a_i)$ denote the parameters of the ith neuron and define the measure

$$\mu^{(m)} = \frac{1}{m} \sum_{i=1}^{m} \delta_{\theta_i}.$$

We can then rewrite the neural network as

$$f(x; w, a) = f(x; \mu^{(m)}) \coloneqq \int a\, \sigma(w^\top x)\, d\mu^{(m)}(w, a).$$

Now, we assume that the initialization of the parameters is such that $\mu^{(m)} \xrightarrow{d} \mu$ as $m \to \infty$ for some probability measure μ (which has density ρ). We can then consider training the network using gradient flow on the squared loss from data $\{(x_i, y_i)\}_{i=1}^n$. The risk functional is given by

$$R(\mu) = \frac{1}{2n} \sum_{i=1}^{n} (f(x_i; \mu) - y_i)^2.$$

It is equivalent to minimize the alternative risk functional

$$F(\mu) = \frac{1}{2n} \sum_{i=1}^{n} f(x_i; \mu)^2 - \frac{1}{n} \sum_{i=1}^{n} y_i \, f(x_i; \mu).$$

The first term can be written as

$$\frac{1}{2n} \sum_{i=1}^{n} f(x_i; \mu)^2 = \frac{1}{2} \int K((w, a), (w', a')) \, d(\mu \times \mu)((w, a), (w', a')),$$

where

$$K((w, a), (w', a')) = \frac{a a'}{n} \sum_{i=1}^{n} \sigma(w^\top x_i) \, \sigma((w')^\top x_i).$$

The second term can be written as

$$-\frac{1}{n} \sum_{i=1}^{n} y_i \, f(x_i; \mu) = \int V \, d\mu,$$

where

$$V(w, a) = -\frac{a}{n} \sum_{i=1}^{n} y_i \, \sigma(w^\top x_i).$$

Therefore, using the potential energy and interaction energy formulas derived in Section 5.3, the Wasserstein gradient of F is given by

$$\begin{aligned} \nabla_{\mathcal{W}} F(\mu) &= \nabla V + \int \nabla_{(w,a)} K((w, a), (w', a')) \, d\mu(w', a') \\ &= -\frac{1}{n} \sum_{i=1}^{n} y_i \begin{pmatrix} a \, \sigma'(w^\top x_i) \, x_i \\ \sigma(w^\top x_i) \end{pmatrix} + \int \frac{a'}{n} \sum_{i=1}^{n} \begin{pmatrix} a \, \sigma'(w^\top x_i) \, \sigma((w')^\top x_i) \, x_i \\ \sigma(w^\top x_i) \, \sigma((w')^\top x_i) \end{pmatrix} d\mu(w', a'), \end{aligned}$$

and the Wasserstein gradient flow follows the McKean-Vlasov equation

$$\begin{aligned} \partial_t \rho_t(w, a) = \nabla_{(w,a)} \cdot \Bigg(\rho_t(w, a) \Bigg[& \frac{1}{n} \sum_{i=1}^{n} y_i \left(a \, \sigma'(w^\top x_i) \, x_i \; \sigma(w^\top x_i) \right) \\ & - \int \frac{a'}{n} \sum_{i=1}^{n} \begin{pmatrix} a \, \sigma'(w^\top x_i) \, \sigma((w')^\top x_i) \, x_i \\ \sigma(w^\top x_i) \, \sigma((w')^\top x_i) \end{pmatrix} d\mu_t(w', a') \Bigg] \Bigg). \end{aligned}$$

Note that this is a generalization of the finite-width case, where each particle (w_i, a_i) moves according to the negative gradient of the risk; this is recovered by taking $\mu_t = \mu_t^{(m)}$. The mean-field limit is one

explanation for feature learning in infinite-width neural networks, as the individual features $\sigma(w^\top x)$ evolve during training. Furthermore, this approach explains why gradient descent is able to find global minimizers in the overparameterized regime; under suitable assumptions on the activation function and initialization, it can be shown that the Wasserstein gradient flow for the infinite-width network converges to a global minimizer. The process of converting guarantees for the infinite-width network to guarantees for large but finite-width networks is called *propagation of chaos*.

6.2 Further intuition for deep learning

In this section, we summarize some well-known intuition for deep learning theory. This spans a lot of papers within the last decade, so we won't bother with citations. Even so, this list is obviously incomplete and biased towards my own interests.

6.2.1 Generalization and benign overfitting

Deep neural networks often overfit the training data (even in the $p \gg n$ regime with many more parameters than training points) but still generalize well to unseen data. This phenomenon is sometimes referred to as *benign overfitting*. One heuristic explanation for this phenomenon is that SGD initialized at zero will converge to a minimum-norm solution for a linear regression task, and minimum-norm solutions often generalize well; this is called *implicit regularization*. Similar phenomena have been observed for more complex networks and tasks as well. In addition, there is a lot of empirical and theoretical work showing that *flat local minima* generalize better than sharp local minima.

The *double descent* phenomenon is that the test error initially decreases, then increases as the model complexity approaches the *interpolation threshold* $p = n$, and then descreases again as the model complexity increases further. The usual explanation for this phenomenon is that as the model complexity increases beyond the interpolation threshold, the model is able to use implicit regularization to find a good interpolating solution. Then, the expressivity of the model becomes a benefit rather than a hindrance.

6.2.2 Training dynamics of neural networks

The phenomenon of *grokking* (related to double descent) is that a model trained on a small training set may initially overfit to zero training error but poor test error, but after further training, the test error suddenly improves dramatically. One explanation for this phenomenon is that the model quickly memorizes the training data, but then slowly learns more generalizable features over time as the regularization effect of SGD comes into effect. This is one reason why weight decay is often helpful in practice, especially for training transformers, where grokking has been frequently observed.

The analysis of infinitely deep neural networks with residual connections usually models the network as a continuous-time neural ODE. In this case, the limiting behavior of the network as the depth is taken to infinity can be analyzed using tools from dynamical systems. For example, this is the approach taken when modeling transformers as an interacting particle system in continuous time.

The *f-principle* is the empirical observation that neural networks tend to learn low-frequency approximations to functions first, before learning higher-frequency components. For example, this can even be observed in one dimension by training a neural network and computing its Fourier transform at different points during training. The explanation for this phenomenon is usually called *spectral bias*, which links the f-principle to the eigenvalues of the neural tangent kernel. This is one explanation for why neural networks generalize well, as low-frequency functions often generalize better than high-frequency functions, and provides one justification for early stopping as a regularization technique.

If we are optimizing a function $f(x) = \frac{1}{2} x^\top H x$ by gradient descent, it is well-known that gradient descent is stable if and only if the learning rate η satisfies $\eta < 2/\lambda_{\max}(H)$. Otherwise, the iterates diverge along the eigenvector associated to the largest eigenvalue. Despite this, it was observed empirically that neural networks have two phases of training: an initial *progressive sharpening* phase where the sharpness (largest eigenvalue of the Hessian) increases until it reaches approximately $2/\eta$, followed by a *edge of stability* phase where the sharpness hovers around $2/\eta$. Here, the loss will oscillate in the short-term, but decrease in the long-term. Intuitively, when the optimizer steps into a region with sharp curvature, the gradient step is large enough to overshoot the minimum along that direction, *catapulting* the iterates out of the sharp region. This is loosely because the loss is not a perfect quadratic, and there is some effect due to the higher-order terms. The edge of stability phenomenon is one explanation for why large learning rates often lead to better generalization, as the optimizer is forced to avoid sharp minima (which have a maximum eigenvalue of the Hessian much larger than $2/\eta$).

There is a large body of work built around the idea that stochastic gradient descent will easily escape from strict saddle points, so many nonconvex problems still allow gradient descent to converge to a local minimum efficiently. There is also an interesting line of work showing that in certain settings, gradient descent can converge to a global minimum despite nonconvexity. For example, such phenomena have been observed in matrix factorization, phase retrieval, and the training of two-layer neural networks.

The *lottery ticket hypothesis* is the empirical observation that large neural networks often contain small subnetworks which, when trained in isolation, can achieve similar performance to the full network. This suggests that overparameterization is helpful for "getting lucky" and finding good subnetworks during training. There is some theoretical work related to this, showing that sufficiently overparameterized networks can approximate certain functions well after training only a small subset of the parameters.

Linear mode connectivity is the observation that two neural networks trained from different random

initializations often have a linear path in parameter space between them along which the loss remains low. This suggests that the loss landscape of neural networks has large connected regions of low loss, rather than isolated minima separated by high barriers. This occurs in two main regimes. First, if you train a network for a small number of steps and then clone it and continue training both copies with different SGD noise, the solutions remain linearly connected. This suggests that SGD noise doesn't really matter after the network settles into a convex basin in the loss landscape. Further, the transition point where linear mode connectivity coincides with the time that the network becomes stable to pruning (i.e., the "winning ticket" was found). If two networks are trained from completely different random initializations, they do not exhibit naive linear mode connectivity. However, if we take into account the symmetry of neural networks to permutations of neurons in a single layer, then we can often find a permutation of one network such that the two networks are linearly connected. There is very recent work using optimal transport to prove linear mode connectivity between two networks between two-layer neural networks.

6.2.3 Expressivity of deep neural networks

It is well-known that even a single-layer neural network with a sufficient number of neurons can uniformly approximate any continuous function on a compact domain. Therefore, even a single-layer network is called a *universal approximator*. However, there are results showing that there exist functions which can be efficiently approximated by a deep network but not by a shallow network unless the width is exponentially large. If we measure the expressivity of a network by the number of linear regions it can represent, then it can be shown that deep networks can represent exponentially more linear regions than shallow networks with the same number of parameters. Therefore, depth is an important factor in the expressivity of neural networks.

6.2.4 Theoretical aspects of transformers

A lot of theoretical work on transformers is about *mechanistic interpretability*; this is related to finding human-understandable explanations for their behavior. The most well-known circuit in a transformer is the *induction head*. If the current token is a, the induction head will attend to the previous occurence of a and then increase the probability of generating the token b that previously followed it. Some theoretical work also shows that this behavior emerges abruptly during training (a phase transition) corresponding almost exactly with the development of in-context learning ability.

The main mechanism by which transformers appear to know more concepts than dimensions in their embeddings is called *superposition*. This is the idea that one can store exponentially many "almost-orthogonal" vectors in $\mathbb{R}^d$ by the Johnson-Lindenstrauss lemma, which allows neurons to be *polysemantic*

(activate for multiple unrelated features).

In-context learning is the ability of large language models to perform new tasks at test time just from a few input-output examples, without any parameter updates. One explanation for this phenomenon is that transformers are implicitly implementing gradient descent (or another iterative algorithm) within their forward pass. On many classical tasks (like linear regression) it has been shown that there exist weights allowing a transformer to exactly implement in-context learning via gradient descent. Furthermore, it has been shown empirically that transformers trained on linear regression tasks learn to implement something very close to the theoretically predicted gradient descent formulas.

There is recent work modeling an unnormalized variant of the self-attention mechanism (with $Q = K = V = I_d$) as an interacting particle system. This system is shown to be the Wasserstein gradient flow for the interaction energy functional

$$E(\mu) = \frac{1}{2\beta} \iint \exp(\beta \langle x, x' \rangle) \, d(\mu \times \mu)(x, x')$$

on the unit sphere, and therefore the infinite-depth limit of this attention mechanism clusters all input tokens to a single point. In practice, it has been observed that deep transformers tend to cluster token representations in later layers, which is consistent with this theoretical result. Furthermore, this result explains why the last few layers of a transformer are often less important for performance; if all tokens are clustered together, then the representations are not very informative.

In terms of expressivity, it has been shown that transformers are strictly more expressive than recurrent neural networks (RNNs) on specific tasks. For example, transformers have better performance on "associative recall" tasks, where the model is required to have fast and precise random access to long-term memory. However, transformers are slower than RNNs at processing very long sequences due to their quadratic complexity in sequence length.

The training of a transformer has been shown to satisfy a particular *scaling law*: test loss decreases as a power law where the exponent depends on the intrinsic dimension of the data distribution and the computational budget. The fact that these scaling laws can be accurately predicted is one explanation for why large language models can be trained at such a large scale.

Interestingly, transformers trained on language modeling tasks have been shown to have overlapping training and testing loss curves. This suggests that transformers are able to generalize extremely well from training to testing data, which is one explanation for their success in practice. One possible explanation for this phenomenon relies on the immense size of the dataset used to train large language models. If the training data is sufficiently diverse and large, then the model may effectively generalize well to unseen data because the training data already covers most of the relevant distribution. Related, another theory is that there is leakage between the training and testing data in large language modeling datasets, but there are

significant efforts in practice to make sure that this is not the case.

Bibliography

Ambrosio, L., Gigli, N., & Savaré, G. (2005). *Gradient flows: In metric spaces and in the space of probability measures*. Springer.

Chewi, S., Niles-Weed, J., & Rigollet, P. (2025). *Statistical optimal transport* (Vol. 2364). Springer Nature.

Chizat, L., & Bach, F. (2018). On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in Neural Information Processing Systems*, *31*.

Givens, C. R., & Shortt, R. M. (1984). A class of Wasserstein metrics for probability distributions. *Michigan Mathematical Journal*, *31*(2), 231–240.

Jacot, A., Gabriel, F., & Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. *Advances in Neural Information Processing Systems*, *31*.

Knott, M., & Smith, C. S. (1984). On the optimal mapping of distributions. *Journal of Optimization Theory and Applications*, *43*, 39–49.

Kreyszig, E. (1991). *Introductory functional analysis with applications*. John Wiley & Sons.

Mei, S., Montanari, A., & Nguyen, P.-M. (2018). A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, *115*(33), E7665–E7671.

Mémoli, F. (2011). Gromov–Wasserstein distances and the metric approach to object matching. *Foundations of Computational Mathematics*, *11*, 417–487.

Neal, R. M. (1996). Priors for infinite networks. In *Bayesian learning for neural networks* (pp. 29–53). Springer.

Rockafellar, R. T. (1970). Convex analysis.

Santambrogio, F. (2015). Optimal transport for applied mathematicians.

Vayer, T. (2020). A contribution to optimal transport on incomparable spaces. *arXiv preprint arXiv:2011.04447*.

Villani, C. (2003). *Topics in optimal transportation* (Vol. 58). American Mathematical Society.

Villani, C. (2009). *Optimal transport: Old and new* (Vol. 338). Springer.

Zhang, Z., Goldfeld, Z., Greenewald, K., Mroueh, Y., & Sriperumbudur, B. K. (2024a). Gradient Flows and Riemannian Structure in the Gromov-Wasserstein Geometry. *arXiv preprint arXiv:2407.11800*.

Zhang, Z., Goldfeld, Z., Mroueh, Y., & Sriperumbudur, B. K. (2024b). Gromov–Wasserstein distances: entropic regularization, duality and sample complexity. *The Annals of Statistics*, *52*(4), 1616–1645.

Appendix A

Supplementary results

We use this appendix to discuss some results that are used in the main text (and which may be of interest in their own right).

A.1 Zorn's lemma

In this section, we discuss Zorn's lemma, which is a generalization of induction to infinite sets. It is equivalent to the axiom of choice and transfinite induction, so we omit the proof of equivalence here and assume it as an axiom.

Definition A.1.1 (Poset)**.** A *partially ordered set* (or *poset*) is a set P with a binary relation $\leq$ that is reflexive, antisymmetric, and transitive.

Note that there may be a, b in a poset where $\leq$ isn't defined between them.

Definition A.1.2 (Chain)**.** A *chain* (or *totally ordered set*) in a poset P is a subset $C \subseteq P$ such that $\leq$ is defined between all elements.

Definition A.1.3 (Maximal element)**.** An element $m \in P$ is *maximal* if $m \leq p$ implies $m = p$.

Axiom A.1.1 (Zorn's lemma)**.** *If all chains have an upper bound in a nonempty poset P, then there is a maximal element $m \in P$.*

For instance, Zorn's lemma implies that all vector spaces have Hamel bases and all Hilbert spaces have a total orthonormal set.

A.2 The Baire category theorem

In this section, we discuss the Baire category theorem, which prevents points in a complete metric space from being too sparse.

Definition A.2.1 (Nowhere dense)**.** A set $M \subseteq X$ is *nowhere dense* if $\overline{M}$ has empty interior.

Definition A.2.2 (Meager)**.** A set $M \subseteq X$ is *meager* if it is a countable union of nowhere dense sets and is *nonmeager* otherwise.

Theorem A.2.1 (Baire category theorem)**.** *If X is a complete metric space, then X is nonmeager. Equivalently, the intersection of countably many open dense sets in X is also dense.*

Proof. The proof proceeds by contradiction; we construct a Cauchy sequence using a nested sequence of balls. Let $X = \bigcup_{n=1}^{\infty} M_n$ where each M_n is nowhere dense. Then, $\overline{M_1}^c$ is nonempty and open, so we can find a ball $B_1 \subseteq \overline{M_1}^c$. We can then find a ball $B_2 \subseteq B_1 \cap \overline{M_2}^c$ with radius at most half of B_1's radius, and so on. This gives a nested sequence of balls $B_1 \supseteq B_2 \supseteq \cdots$. By the completeness of X, there is a unique point $x \in \bigcap_{n=1}^{\infty} B_n$. But now, by construction, $x \notin M_n$ for any n, which gives the desired contradiction. □

A.3 Tychonoff's theorem

In this section, we discuss Tychonoff's theorem, which states that the product of compact spaces is compact.

Definition A.3.1 (Finite intersection property)**.** A collection $\mathcal{F}$ of sets has the *finite intersection property* if every finite subcollection of $\mathcal{F}$ has nonempty intersection.

The following lemma is a useful characterization of compactness in topological spaces, and is easily equivalent to the usual definition.

Lemma A.3.1. *A topological space X is compact if and only if every collection of closed sets with the finite intersection property has nonempty intersection.*

Proof. The forward direction is elementary and follows from Cantor's intersection theorem. We'll now show the contrapositive of the reverse direction. Suppose that X is not compact and U is a collection of open sets with no finite subcover. Then, the complements of these sets form a collection of closed sets with the finite intersection property, but their intersection is empty. □

Theorem A.3.2 (Tychonoff)**.** *If X_i is compact for all $i \in I$, then $\prod_{i \in I} X_i$ is compact.*

Proof. Let $\mathcal{F}$ be a family of closed sets in $\prod_{i\in I} X_i$ with the finite intersection property; by Lemma A.3.1, it suffices to show that $\mathcal{F}$ has nonempty intersection. Now, let $\mathcal{P}$ be the set of pairs (B, x_B) such that $B \subseteq I$ is finite, $x_B \in \prod_{i\in B} X_i$, and for any finite subfamily $\mathcal{G} \subseteq \mathcal{F}$ there exists $y \in \bigcap_{G\in\mathcal{G}} G$ such that $\pi_B(y) = x_B$. We define a partial order on $\mathcal{P}$ by $(B, x_B) \leq (C, x_C)$ if $B \subseteq C$ and $\pi_B(x_C) = x_B$ (where π_B denotes the projection onto the coordinates in B). It's clear that any chain has an upper bound, so by Zorn's lemma (Axiom A.1.1), there exists a maximal element (M, x_M). If $M \neq I$, pick $i_0 \in I \setminus M$ and define (for each finite $\mathcal{G} \subseteq \mathcal{F}$):

$$S_{\mathcal{G}} = \left\{ x_0 \in X_{i_0} : \exists\, y \in \bigcap_{G\in\mathcal{G}} G \text{ such that } \pi_M(y) = x_M \text{ and } \pi_{i_0}(y) = x_0 \right\}.$$

Now, the collection of possible sets $S_{\mathcal{G}}$ satisfies the finite intersection property since $(M, x_M) \in \mathcal{P}$. Hence, the intersection of the possible sets $S_{\mathcal{G}}$ is nonempty by compactness of X_{i_0}. But this gives an extension of (M, x_M), which is a contradiction. □