EE660 Lecture Notes: Mathematical Foundations and Methods of Machine Learning

Stephen Tu

stephen.tu@usc.edu

Last Updated: January 8, 2024

Abstract

This set of lecture notes accompanies the Spring 2024 offering of EE660 at USC.

Note: This document is a continual work in progress and will be constantly updated throughout the semester. Be sure to frequently check for updates.

Contents

1 Supervised learning

We start our study by considering an unknown distribution $\mathcal{D}$ over an event space $\mathsf{Z} = \mathsf{X} \times \mathsf{Y}$. Here, the space X denotes the *covariates*, and we will typically assume for simplicity that X is a subset of Euclidean space. On the other hand, the space Y denotes the *labels*. We will let $p(x, y) = p(x)p(y \mid x)$ denote the joint density function over Z.[1]

We first consider *binary classification*, where $\mathsf{Y} = \{-1, +1\}$. The binary classification problem is: given i.i.d. *training examples* $S_n := \{(x_1, y_1), \ldots, (x_n, y_n)\}$ from $\mathcal{D}$, learn a predictor $\hat{f}_n : \mathsf{X} \mapsto \mathsf{Y}$, so that ideally on unseen *test examples* $(x, y) \sim \mathcal{D}$, we have that $\hat{f}_n(x) = y$.

1.1 What form should the predictor take?

In order to propose an algorithm to learn the predictor $\hat{f}_n$, we need to specify two things:

(a) What form should the predictor function $\hat{f}_n$ take?

(b) What algorithm should we use to learn its representation?

To shed some insight into question (a), let us set forth a framework to study what an *optimal* predictor looks like. Let $\ell : \mathsf{Y} \times \mathsf{Y} \mapsto \mathbb{R}$ be a loss function, where we think of $\ell(\hat{y}, y)$ as the cost of predicting $\hat{y}$ when the true label is y. For a concrete example, think of $\ell(\hat{y}, y) = \mathbf{1}\{\hat{y} \neq y\}$, which is the *zero-one loss* or the *mistake loss*.

With this setup, we can now pose an optimization problem to compute the best predictor:

$$\min_{f : \mathsf{X} \to \mathsf{Y}} \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(f(x), y)]. \tag{1.1}$$

(Note for what follows, we will usually drop the subscript under the expectation $\mathbb{E}$. However, please do not be afraid to ask if it is ever not clear what random variables we are taking expectations and/or probabilities with respect to.) Above, the optimization is taken over all (measurable) functions mapping covariates X to labels Y. It may seem at first that this is an impossible problem to solve, but it actually has a very simple solution.

Proposition 1.1. *Suppose the loss ℓ is* proper*, i.e.,*

$$\ell(1, -1) > \ell(-1, -1), \quad \textit{and} \quad \ell(-1, 1) > \ell(1, 1).$$

A solution $\hat{f} : \mathsf{X} \to \mathsf{Y}$ of Equation (1.2) takes on the form:[2]

$$\hat{f}(x) = \operatorname{sgn}\left\{p(y = +1 \mid x) - \frac{\ell(1, -1) - \ell(-1, -1)}{\ell(-1, 1) - \ell(1, 1)} p(y = -1 \mid x)\right\}. \tag{1.2}$$

Proof. By iterated expectations,

$$\mathbb{E}[\ell(f(x), y)] = \mathbb{E}[\mathbb{E}[\ell(f(x), y) \mid x]].$$

Let us now study the inner quantity, conditioned on x. The key is that $f(x)$, once conditioned on x, is no longer random (the jargon I will often use is that $f(x)$ is "x-measurable"). Hence, the inner conditional expectation decomposes very cleanly:

$$\mathbb{E}[\ell(f(x), y) \mid x] = \ell(f(x), 1)p(y = +1 \mid x) + \ell(f(x), -1)p(y = -1 \mid x).$$

[1] We will not be overly concerned with measure-theoretic concerns in this course.

[2] The sgn function is $\operatorname{sgn}(a) := \mathbf{1}\{a \geqslant 0\} - \mathbf{1}\{a < 0\}$. Note the tie at $a = 0$, such that $\operatorname{sgn}(0) = 1$, is broken arbitrarily.

Hence, an optimal $f(x)$, *for this value of* x that we are conditioning on, should pick either $+1$ or -1 to minimize the RHS of the above equation. That is, abbreviating $p_1 = p(y = 1 \mid x)$ and $p_{-1} = p(y = -1 \mid x)$:

$$f(x) = \begin{cases} +1 & \text{if } \ell(-1,1)p_1 + \ell(-1,-1)p_{-1} \geqslant \ell(1,1)p_1 + \ell(1,-1)p_{-1} \\ -1 & \text{otherwise.} \end{cases}$$

The result now follows by re-arranging the expression above, relying on the properness of the loss ℓ to divide by $\ell(-1,1) - \ell(1,1)$. □

Let us now consider a special case of Equation (1.2) where $\ell(\hat{y}, y) = \mathbf{1}\{\hat{y} \neq y\}$ is the zero-one loss. In this case, the predictor $\hat{f}$ reduces to the *maximum a-posterior* (MAP) predictor:

$$\hat{f}(x) = \arg\max_{y \in \mathsf{Y}} p(y \mid x). \tag{1.3}$$

Note that by Bayes' rule, we have that:

$$p(y \mid x) = \frac{p(y)p(x \mid y)}{p(x)}.$$

Hence, under a uniform prior on the labels, i.e., $p(y = 1) = p(y = -1) = 1/2$, we have that $\hat{f}$ is also equal to the *maximum-likelihood estimator* (MLE):

$$\hat{f}(x) = \arg\max_{y \in \mathsf{Y}} p(x \mid y).$$

Exercise 1.2. Suppose we are in a multi-class classification setting, so that $\mathsf{Y} = \{1, \ldots, K\}$. Show that the optimal predictor $\hat{f} : \mathsf{X} \mapsto \mathsf{Y}$ minimizing the zero-one loss is still the MAP predictor:

$$\hat{f}(x) = \arg\max_{y \in \mathsf{Y}} p(y \mid x).$$

1.2 Perceptron

We now turn to answer the second question (b), what algorithm do we use to learn the representation of a decision function $f : \mathsf{X} \mapsto \mathsf{Y}$. We will start from a very simple prototypical algorithm, called the Perceptron, and understand its various properties. It will turn out that while the Perceptron is very simple, the same underlying principles are used to train essentially every modern machine learning model.

The Perceptron starts by making a modeling assuming regarding the form of the predictor $\hat{f}$. Conceptually, every predictor $\hat{f}$ (whether optimal or not) induces a *partition* of X (which we now assume to be a subset of $\mathbb{R}^d$):

$$T_1 := \{x \in \mathsf{X} \mid \hat{f}(x) = 1\}, \quad \text{and} \quad T_{-1} := \{x \in \mathsf{X} \mid \hat{f}(x) = -1\}.$$

(In fact, there is a one-to-one correspondence between partitions of X and predictors $\hat{f}$.) In the Perceptron, it is posited that the optimal predictor induces a *linear* partition of X: that is, we cut X in half by a hyperplane (that we will assume crosses through the origin for convenience), and assign the label $+1$ to one side of the hyperplane, and -1 to the other. Mathematically:

$$f_w(x) = \operatorname{sgn}\{\langle w, x \rangle\}, \quad w \in \mathbb{R}^d.$$

Learning here then refers to learning the weights w which parameterize the separating hyperplane. The Perceptron prescribes the following procedure to learn these weights.

We will now study the behavior of the Perceptron on *linearly separable* data.

Algorithm 1 The Perceptron

Input: Dataset $S_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$.

Set $w = 0$.
while `true` **do**
  Let $\mathcal{I}$ be a (possibly random) permutation of $\{1, \ldots, n\}$.
  Set `mistake = false`.
  for $i \in \mathcal{I}$ **do**
    if $y_i \langle w, x_i \rangle < 1$ **then**
      Set $w \leftarrow w + y_i x_i$.
      Set `mistake = true`.
    end if
  end for
  if `not mistake` **then**
    return w.
  end if
end while

Definition 1.3. *A dataset $S_n = \{(x_1, y_1), \ldots, (x_n, y_n)\} \subset \mathsf{Z}$ is* linearly separable *if there exists a non-zero $w \in \mathbb{R}^d$ such that:*

$$\forall i \in \{1, \ldots, n\}, \quad y_i = \operatorname{sgn}\{\langle w, x_i \rangle\}.$$

Such a non-zero $w \in \mathbb{R}^d$ satisfying the above condition is called a linear separator.

Exercise 1.4. Which of the following $n = 2$ datasets with $\mathsf{X} = \mathbb{R}^2$ are linearly separable?

(a) $S_2 = \{((0, 1), +1), ((0, -1), +1)\}$.

(b) $S_2 = \{((0, 1), -1), ((0, -1), -1)\}$.

(c) $S_2 = \{((0, 0), +1), ((0, 1), -1)\}$.

Next, we will define the notion of a *margin*, which quantifies how linearly separable a dataset is. Before we do this, let S be a subset of Euclidean space. The distance function $\operatorname{dist}(x, S)$ is defined as:

$$\operatorname{dist}(x, S) = \inf\{\|x - z\| \mid z \in S\}.$$

Also, for a vector $w \in \mathbb{R}^d$, we define its hyperplane H_w as:

$$H_w := \{x \in \mathbb{R}^d \mid \langle x, w \rangle = 0\}.$$

We now have enough notation to define the notion of margin.

Definition 1.5. *Let S_n be a linearly separable dataset (cf. Definition 1.3), and let $W \subset \mathbb{R}^d$ denote the set of linear separators. The* margin $\gamma(S_n)$ *is defined as:*

$$\gamma(S_n) := \sup_{w \in W} \min_{i \in \{1, \ldots, n\}} \operatorname{dist}(x_i, H_w) = \sup_{w \in W} \min_{i \in \{1, \ldots, n\}} \frac{|\langle x_i, w \rangle|}{\|w\|} = \sup_{\|w\| = 1} \min_{i \in \{1, \ldots, n\}} y_i \langle x_i, w \rangle. \tag{1.4}$$

The idea behind margin is to assign a quantity to a linear separator which measures how "robust" it is. (Draw some pictures in class). The following exercises check your understanding of margin.

Exercise 1.6. Prove the last two equalities in Equation (1.4).

Exercise 1.7. Suppose S_n is linearly separable. Consider the alternative definition of margin, which does not take into account whether or not a hyperplane separates the dataset:

$$\gamma'(S_n) = \sup_{w \neq 0} \min_{i \in \{1,\dots,n\}} \operatorname{dist}(x_i, H_w).$$

Is it possible to construct a S_n such that $\gamma'(S_n) > \gamma(S_n)$? If so, provide one construction. If not, prove it is not possible.

1.2.1 Mistake bound

With this notion of margin, we are able to prove our first non-trivial result regarding Perceptron, which is its well-known *mistake bound*. While this is a classic result, we will follow the derivations based on the presentation in Hardt and Recht [2022, Ch. 3].

To start, we will define a sequences of weights $w_0, w_1, \dots$ recursively, which captures the main Perceptron update. Let $i_0, i_1, \dots$ denote an *arbitrary* sequence of indices in $\{1, \dots, n\}$ (meaning, these indices could be repeating, etc.). Our sequence $w_0, w_1, \dots$ is defined as:

$$w_{t+1} = w_t + y_{i_t} x_{i_t} \mathbf{1}\{y_{i_t} \langle w_t, x_{i_t} \rangle < 1\}, \quad w_0 = 0. \tag{1.5}$$

Note that our notation makes implicit the dependence of the iterates $\{w_t\}$ on the indices $\{i_t\}$. In the sequel, it will be clear from context how the indices $\{i_t\}$ are specified.

Lemma 1.8 (Perceptron mistake bound)**.** *Let S_n be linearly separable. Consider the Perceptron sequence $\{w_t\}_{t \geqslant 0}$ defined in Equation (1.5). Let $m_t := \mathbf{1}\{y_{i_t} \langle w_t, x_{i_t} \rangle < 1\}$ denote the indicator of whether a mistake occured at time t, and let $M_t := \sum_{s=0}^{t-1} m_t$ denote the culmulative number of mistakes made up to time t. We have that for all $t \in \mathbb{N}$:*

$$M_t \leqslant \frac{2 + \max_{i \in \{1,\dots n\}} \|x_i\|^2}{\gamma^2(S_n)}. \tag{1.6}$$

Before proceeding to the proof, it is worth reflecting on the nature of this result. It is quite remarkable in that it is *dimension independent*. Furthermore, since the RHS is independent of t, this result implies there exists a t such that for all $t' \geqslant t$, we have $m_{t'} = 0$ (why?). In other words, the Perceptron eventually stops making mistakes. Furthermore, since this result holds for *any* sequence $i_0, i_1, \dots$, by setting the sequence $\{i_t\}$ to be repeated copies of $\{1, \dots, n\}$, this ensures that eventually the Perceptron indeed learns a linear separator (why?).

Proof of Lemma 1.8. Put $B := \max_{i \in \{1,\dots n\}} \|x_i\|^2$. Fix a t and suppose that $m_t = 1$. Then, by expanding the square:

$$\begin{aligned} \|w_{t+1}\|^2 = \|w_t + y_{i_t} x_{i_t}\|^2 &= \|w_t\|^2 + 2 y_{i_t} \langle w_t, x_{i_t} \rangle + \|x_{i_t}\|^2 \\ &\leqslant \|w_t\|^2 + 2 y_{i_t} \langle w_t, x_{i_t} \rangle + B \leqslant \|w_t\|^2 + 2 + B. \end{aligned}$$

The key here is the last inequality, which is a consequence of $m_t = 1$, or equivalently $y_{i_t} \langle w_t, x_{t_i} \rangle < 1$. On the other hand, if $m_t = 0$, then $\|w_{t+1}\|^2 = \|w_t\|^2$ by definition. Hence, for every t:

$$\|w_{t+1}\|^2 \leqslant \|w_t\|^2 + (2 + B) m_t.$$

Unrolling this recursion to $t = 0$ and using the definition of M_t yields for every t:

$$\|w_t\|^2 \leqslant (2 + B) M_t.$$

Next, let $w_\star$ denote a unit norm max-margin linear separator, i.e., a vector such that (cf. Equation (1.4)):

$$\gamma(S_n) = \min_{1 \leqslant i \leqslant n} |\langle x_i, w_\star \rangle|.$$

Now, suppose that $m_t = 1$, and hence, since $w_\star$ linearly separates S_n,

$$\langle w_\star, w_{t+1} - w_t \rangle = y_{i_t} \langle w_\star, x_t \rangle = |\langle w_\star, x_t \rangle| \geqslant \gamma(S_n).$$

Next, we use a *telescoping sum*, which is a trick often used in optimization proofs. The telescoping sum simply states (recall that $w_0 = 0$):

$$w_t = \sum_{k=1}^{t} (w_k - w_{k-1}) = \sum_{k=1}^{t} (w_k - w_{k-1}) m_{k-1}.$$

While this may appear tautological, it is actually quite useful, since:

$$\|w_t\| \geqslant \langle w_\star, w_t \rangle = \sum_{k=1}^{t} \langle w_\star, w_k - w_{k-1} \rangle m_{k-1} \geqslant \gamma(S_n) \sum_{k=1}^{t} m_{k-1} = \gamma(S_n) M_t.$$

Above, the first inequality holds by Cauchy-Schwarz (recall that $w_\star$ is unit norm). We now have upper and lower bounds on $\|w_t\|$, from which we conclude:

$$\gamma^2(S_n) M_t^2 \leqslant \|w_t\|^2 \leqslant (2 + B) M_t \implies M_t \leqslant \frac{2 + B}{\gamma^2(S_n)}.$$

□

Exercise 1.9. Suppose Algorithm 1 is called with a linearly separable dataset S_n. Let B bound the covariate norm, i.e., $B = \max_{i \in \{1,\dots,n\}} \|x_i\|^2$. Show that Algorithm 1 terminates with a solution $w \in \mathbb{R}^d$, which linearly separates S_n, in at most $(2 + B)/\gamma^2(S_n)$ passes through S_n.

1.2.2 Generalization bound

The mistake bound from Section 1.2.1 gives us a way to prove a generalization bound about Perceptron, on *unseen* instances. We will utilize a clever technique known as leave-one-out analysis. We need one more piece of notation before we proceed. Given a linearly separable dataset S_n, we let $w(S_n)$ denote the limiting value of the Perceptron sequence $\{w_t\}$ defined in Equation (1.5), where we fix the indices $\{i_t\}$ to repeatedly cycle through $\{1, \dots, n\}$, i.e., $i_t = (t \mod n) + 1$. Note that $w(S_n)$ is well-defined by Lemma 1.8. For what follows, we will assume that the distribution $\mathcal{D}$ is linearly separable, meaning that almost surely, for any $n \in \mathbb{N}_+$, S_n sampled i.i.d. from $\mathcal{D}$ is linearly separable.

Lemma 1.10 (Perceptron generalization bound)**.** *Suppose that $\mathcal{D}$ is linearly separable. Also, suppose that $\|x\| \leqslant B$ almost surely when $x \sim p(x)$. Then, letting $(x, y) \in \mathsf{Z}$ be drawn independently from S_n,*

$$\mathbb{P}\{y \langle w(S_n), x \rangle < 1\} \leqslant \frac{2 + B}{n + 1} \mathbb{E}\left[\frac{1}{\gamma^2(S_{n+1})}\right]. \tag{1.7}$$

Proof. We will construct an equivalent probability space as follows. Let $S_{n+1} = \{z_1, \dots, z_{n+1}\}$ be $n+1$ i.i.d. draws from $\mathcal{D}$, and let $S_{n+1}^{-k} := \{z_1, \dots, z_{k-1}, z_{k+1}, \dots z_{n+1}\}$ be the dataset where the k-th point is left out. Note that by exchangability, for any k,

$$\mathbb{P}\{y \langle w(S_n), x \rangle < 1\} = \mathbb{P}\{y_k \langle w(S_{n+1}^{-k}), x_k \rangle < 1\}.$$

Hence we can average over all indices to conclude:

$$\mathbb{P}\{y \langle w(S_n), x \rangle < 1\} = \frac{1}{n + 1} \sum_{k=1}^{n+1} \mathbb{P}\{y_k \langle w(S_{n+1}^{-k}), x_k \rangle < 1\}.$$

(This type of relabeling argument is common to leave-one-out analyses, and we will see it come up again later when we study algorithmic stability. In fact, this argument can be interpreted as a form of algorithmic stability analysis for Perceptron.)

Now, let $I_m \subset \{1, \dots, n+1\}$ denote the indices for which the Perceptron on S_{n+1} made a mistake at any point during execution, i.e.,

$$I_m := \{(t \mod (n+1)) + 1 \mid m_t = 1,\, t \geqslant 0\}.$$

Let $k \in \{1, \dots, n+1\}$, and suppose $k \notin I_m$. Then, we have the key equality:

$$w(S_{n+1}^{-k}) = w(S_{n+1}).$$

It is worth taking a moment to think why this equality is true. If $k \notin I_m$, then this means that each time Perceptron on S_{n+1} cycles through the index k, it makes no update. Hence, it is equivalent to running Perceptron on S_{n+1}^{-k}, where the k-th datapoint is altogether removed.

Now here is the final trick. Again let $k \notin I_m$. Then we have:

$$1 \leqslant y_k \langle w(S_{n+1}), x_k \rangle = y_k \langle w(S_{n+1}^{-k}), x_k \rangle.$$

(Check your understanding: why is the first inequality is true?) Therefore, chaining these arguments together and applying the mistake bound from Lemma 1.8,

$$\begin{aligned}\sum_{k=1}^{n+1} \mathbf{1}\{y_k \langle w(S_{n+1}^{-k}), x_k \rangle < 1\} &= \sum_{k \in I_m} \mathbf{1}\{y_k \langle w(S_{n+1}^{-k}), x_k \rangle < 1\} + \sum_{k \notin I_m} \mathbf{1}\{y_k \langle w(S_{n+1}^{-k}), x_k \rangle < 1\} \\ &= \sum_{k \in I_m} \mathbf{1}\{y_k \langle w(S_{n+1}^{-k}), x_k \rangle < 1\} \leqslant |I_m| \leqslant \frac{2+B}{\gamma^2(S_{n+1})}.\end{aligned}$$

The result now follows by taking expectations on both sides of the above inequality. □

Remark 1.11. Recall in the definition of $w(S_n)$, we specifically fixed the indices $\{i_t\}$ to repeatedly cycle through $\{1, \dots, n\}$. Where in the proof of Lemma 1.10 did we use this assumption? What modifications to how the sequence $\{i_t\}$ is generated can you think of, which allow the proof to still go through?

Remark 1.12. Lemma 1.10 states that the margin mistake error $\mathbb{P}\{y\langle w(S_n), w\rangle < 1\}$ decreases at a rate of $O(1/n)$. In the literature, this is referred to as a *fast rate*, since it is faster than the usual $O(1/\sqrt{n})$ rate which arises out of typical concentration of measure arguments (which we will see in a few lectures). This fast-rate is not to be taken for granted: obtaining a fast-rate in general settings typically involves a much more non-trivial argument, which is out of scope for this course.

Interpreting the margin mistake bound. How does the margin mistake bound from Lemma 1.10 relate to the zero-one loss $\mathbb{P}\{y \neq \operatorname{sgn}(\langle w(S_n), x\rangle)\}$? Let us abbreviate $w = w(S_n)$. We can upper bound the zero-one loss by the margin mistake bound via the following argument:

$$\begin{aligned}\{y \neq \operatorname{sgn}(\langle w, x\rangle)\} &= \{y = 1, \langle w, x\rangle < 0\} \cup \{y = -1, \langle w, x\rangle \geqslant 0\} \\ &= \{y = 1, y\langle w, x\rangle < 0\} \cup \{y = -1, y\langle w, x\rangle \leqslant 0\} \\ &\subset \{y\langle w, x\rangle \leqslant 0\} \subset \{y\langle w, x\rangle < 1\}.\end{aligned}$$

(Note the asymmetry in the cases is due to the tie-breaking choice made in our definition of the sgn function.) Hence by monotonicity of probability,

$$\mathbb{P}\{y \neq \operatorname{sgn}(\langle w(S_n), x\rangle)\} \leqslant \mathbb{P}\{y\langle w(S_n), x\rangle < 1\}.$$

Therefore, Lemma 1.10 implies a bound on the classification error of the Perceptron.

1.3 Empirical Risk Minimization (ERM)

The Perceptron algorithm we studied in Section 1.2, while having nice theoretical properties, is quite limiting. Indeed, its theoretical analysis requires that the underlying distribution $\mathcal{D}$ is linearly separable, a fairly restrictive assumption for many real world datasets. Therefore, we desire a more general framework which does not require such limiting assumptions.

1.3.1 Framework setup

The main framework we will study in this course, which covers essentially most of modern machine learning, is the empirical risk minimization (ERM) framework. We will spend quite a bit of time studying ERM, as it is the de-facto standard workhorse for modern machine learning.

At its core, ERM is a very simple concept. There are only three core ingredients:

1. A loss function $\ell : \Lambda \times \mathsf{Y} \mapsto \mathbb{R}$, where Λ is an auxiliary space.
2. A known, deterministic selection function $\psi : \Lambda \mapsto \mathsf{Y}$.
3. A function (hypothesis) class $\mathscr{F} = \{f : \mathsf{X} \mapsto \Lambda\}$ over which to restrict our search for predictors over. Note, this class induces a set of predictors by composition with the selection function ψ.
4. A training dataset $S_n = \{z_1, \dots, z_n\}$, sampled i.i.d. from an underlying distribution $\mathcal{D}$ over $\mathsf{Z} = \mathsf{X} \times \mathsf{Y}$.

Remark 1.13. While the ERM framework above is stated somewhat abstractly, for binary classification, we will always consider models with auxiliary space $\Lambda = \mathbb{R}$ and selection rule $\psi = \text{sgn}$. This corresponds to models $f(x)$ which classify examples by thresholding $\text{sgn}(f(x))$. This structure will be implicitly assumed moving forward. Note that the extra generality allows for e.g., multi-class classification, regression, etc. to fit under the same framework.

The goal of a learner is to produce a hypothesis $f \in \mathscr{F}$ to minimize the *population risk*:

$$L[f] := \mathbb{E}[\ell(f(x), y)].$$

Towards this goal, the ERM prescribes to us the following predictor:

$$\hat{f}_{\text{erm}} \in \arg\min_{f \in \mathscr{F}} \hat{L}_n[f] := \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i) \right\}. \tag{1.8}$$

The function $\hat{L}_n[f]$ is referred to as the *empirical risk*. The main object of study in this section of the course is the behavior of the population risk of the ERM, i.e., $L[\hat{f}_{\text{erm}}]$. (Note: our notation for both the population risk $L[f]$ and the empirical risk $L[f]$ makes the dependency on the loss ℓ implicit. In the sequel, it should be clear from context what ℓ is, but please do not hesitate to ask if it is not.)

1.3.2 Generalization error

Given a *fixed* predictor $f(x)$ and a dataset S_n, the empirical loss is an unbiased estimator of the true population level error, i.e.,

$$\mathbb{E}[\hat{L}_n[f]] = L[f].$$

However, when the predictor f is no longer independent of the dataset S_n (e.g., f is constructed as a function of S_n, as in the ERM (1.8)), then the empirical risk is no longer an unbiased estimator (convince yourself this is true by constructing an example). Thus, what we will need to do is understand how much the empirical risk can differ from the population risk. We start with a tautological decomposition:[3]

$$L[f] = \underbrace{\hat{L}_n[f]}_{\text{empirical risk}} + \underbrace{L[f] - \hat{L}_n[f]}_{\text{generalization error}} =: \hat{L}_n[f] + \text{gen}[f]. \tag{1.9}$$

[3] My graduate school advisor Ben would often call this decomposition the "fundamental theorem of machine learning".

In ERM, we purposefully make the first quantity, the empirical risk $\hat{L}_n[f]$, as small as possible. However, the question that remains is how does this affect $\mathrm{gen}[f]$? Again, the difficulty is that $\hat{f}_{\mathrm{erm}}$ is a function of the dataset S_n, and therefore $\mathbb{E}[L_n[\hat{f}_{\mathrm{erm}}]] \neq L[\hat{f}_{\mathrm{erm}}]$.

One of the key ideas behind generalization theory in machine learning is that by *restricting the function class $\mathscr{F}$ which we optimize over*, we can indeed control the generalization error by quantities which scale according to the complexity of the function class. The way this works is by doing something admittedly crude. Since characterizing the random variable $\hat{f}_{\mathrm{erm}}$ is non-trivial in general, we instead consider the *worst-case* generalization error over the class $\mathscr{F}$:

$$\mathrm{gen}[\hat{f}_{\mathrm{erm}}] \leqslant \sup_{f \in \mathscr{F}} \mathrm{gen}[f].$$

Analyzing the RHS of the above expression will be the focus of our next few lectures.

1.3.3 Finite hypothesis classes

As a warmup which contains the key ideas, let us suppose that $\mathscr{F}$ is finite, i.e., $|\mathscr{F}| < \infty$, and that ℓ is the zero-one loss, i.e., $\ell(\lambda, y) = \mathbf{1}\{\mathrm{sgn}(\lambda) \neq y\}$. We will use Hoeffding's inequality combined with a union bound to control the generalization error of every single predictor in $\mathscr{F}$.

Proposition 1.14. *Let ℓ be the zero-one loss and $\mathscr{F}$ be a finite function class. Fix any $\delta \in (0,1)$. With probability at least $1-\delta$,*

$$\max_{f \in \mathscr{F}} \mathrm{gen}[f] \leqslant \sqrt{\frac{2\log(|\mathscr{F}|/\delta)}{n}}.$$

Proof. As stated above, the main tool is Hoeffding's inequality (cf. Proposition B.12). Fix any $f \in \mathscr{F}$. We apply to the sum $M_n = \sum_{i=1}^n X_i$, with $X_i = \mathbf{1}\{y_i \neq \mathrm{sgn}(f(x_i))\}$. Note that by construction $X_i \in [-1,1]$ and hence for any $t > 0$, by Hoeffding's inequality,

$$\mathbb{P}\{\mathrm{gen}[f] \geqslant t\} = \mathbb{P}\{n^{-1}(\mathbb{E}[M_n] - M_n) \geqslant t\} \leqslant \exp(-nt^2/2).$$

Hence by a union bound,

$$\mathbb{P}\left\{\max_{f \in \mathscr{F}} \mathrm{gen}[f] \geqslant t\right\} = \mathbb{P}\left\{\bigcup_{f \in \mathscr{F}} \{\mathrm{gen}[f] \geqslant t\}\right\} \leqslant \sum_{f \in \mathscr{F}} \mathbb{P}\{\mathrm{gen}[f] \geqslant t\} \leqslant |\mathscr{F}| \exp(-nt^2/2).$$

Now, set the RHS above equal to the given δ and solve for t to conclude. □

The above bound in Proposition 1.14 is a prototypical statistical learning bound. The $1/\sqrt{n}$ rate in the denominator is generally unimprovable, as a consequence of the Central Limit Theorem. The $\sqrt{\log \mathscr{F}}$ scaling in the numerator states that we are allowed to test exponentially many in n hypothesis before our generalization gap dominates the error.

The generalization bound in Proposition 1.14 is quite powerful already. It tells us, via (1.9), that:

$$\forall f \in \mathscr{F},\ L[f] \leqslant \hat{L}_n[f] + \sqrt{\frac{2\log(|\mathscr{F}|/\delta)}{n}}.$$

Again, it is important to emphasize that this is quite remarkable. It implies that we can test exponentially many hypothesis in the number of data points n, until the training error $\hat{L}_n[f]$ fails to be an accurate proxy of the population error $L[f]$. Let us conclude this section with an example. Suppose that $\mathscr{F} = \{f_\theta : \mathsf{X} \mapsto \mathbb{R} \mid \theta \in \mathbb{R}^d\}$. That is, we optimize over a function class parameterized by d parameters. Let $N_{\texttt{f32}}$ denote the number of valid `float32` numbers (so $N_{\texttt{f32}} \leqslant 2^{32}$). Then, on a computer, $\mathscr{F}$ is a finite hypothesis class of cardinality $|\mathscr{F}| = N_{\texttt{f32}}^d$. So, if we perform optimization in our favorite ML framework,[4] we immediate have:

$$\forall f \in \mathscr{F},\ L[f] \leqslant \hat{L}_n[f] + \sqrt{\frac{2(d\log N_{\texttt{f32}} + \log(1/\delta))}{n}}.$$

[4] You are using `jax` (`https://github.com/google/jax`) right?

From this inequality, we have already uncovered a core tenant of conventional machine learning wisdom: *generalization happens when one has more datapoints n than parameters d (i.e., $n \gg d$).* Now, this wisdom has come into question in the past decade or so, with the success of deep overparmaeterized neural networks (which seem to be generalizing in ways that defy this logic). We will revisit this later on, but for now $n \gg d$ implies generalization is a good mental framework to have.

Exercise 1.15. Suppose that ℓ is the zero-one loss and $\mathscr{F}$ is an countably infinite function class of the form:

$$\mathscr{F} = \{f_\kappa \mid \kappa \in \mathbb{N}_+\}.$$

Show that with probability at least $1-\delta$, we have:

$$\forall f_\kappa \in \mathscr{F},\ \mathrm{gen}[f_\kappa] \leqslant c\sqrt{\frac{\log(\kappa/\delta)}{n}},$$

where $c > 0$ is a universal constant.

Hint: you may use without proof the well-known solution to the Basel problem: $\sum_{i=1}^{\infty} i^{-2} = \pi^2/6$.

1.3.4 Fast rates with realizable finite hypothesis classes

In addition to the assumptions of the last section, let us now impose an additional assumption, that there exists an $f \in \mathscr{F}$ such that $L[f] = 0$, that is, no mistakes are made at the population level. This is obviously a strong assumption, but we will see that it provides fast rates (as we saw with the Perceptron when data was linearly separable).

Proposition 1.16. *Suppose that ℓ is the zero-one loss, $\mathscr{F}$ is finite, and there exists $f \in \mathscr{F}$ satisfying $L[f] = 0$. Then, the empirical risk minimizer $\hat{f}_n$ satisfies, with probability at least $1-\delta$,*

$$L[\hat{f}_n] \leqslant \frac{\log(|\mathscr{F}|/\delta)}{n}.$$

Proof. For any $t > 0$, we will study the event $\{L(\hat{f}_n) > t\}$. Define the set of sub-optimal hypothesis as $B(t) = \{f \in \mathscr{F} \mid L(f) > t\}$. Note that $\{L(\hat{f}_n) > t\} = \{\hat{f}_n \in B(t)\}$. Now observe that since there exists an $f \in \mathscr{F}$ with $L[f] = 0$, then the ERM $\hat{f}_n$ will always achieve zero training error, i.e., $\hat{L}_n[\hat{f}_n] = 0$.[5] Hence if $\hat{f}_n \in B(t)$, that means there exists some $f \in B(t)$ with $\hat{L}_n[f] = 0$. So now we compute, for a fixed $f \in B(t)$,

$$\begin{aligned}\mathbb{P}\{\hat{L}_n[f] = 0\} &= \mathbb{P}\left\{\bigcap_{i=1}^n \{\mathrm{sgn}(f(x_i)) = y_i\}\right\} = \prod_{i=1}^n (1 - \mathbb{P}\{\mathrm{sgn}(f(x_i)) \neq y_i\}) \\ &= (1 - L[f])^n \leqslant (1-t)^n \leqslant \exp(-tn),\end{aligned}$$

where we used the elementary inequality $1 - x \leqslant \exp(x)$ for any $x \in \mathbb{R}$. Therefore by a union bound,

$$\begin{aligned}\mathbb{P}\{\hat{f}_n \in B(t)\} &\leqslant \mathbb{P}\{\exists f \in B(t),\ \hat{L}_n(f) = 0\} \leqslant \sum_{f \in B(t)} \mathbb{P}\{\hat{L}_n(f) = 0\} \\ &\leqslant |B(t)| \exp(-tn) \leqslant |\mathscr{F}| \exp(-tn).\end{aligned}$$

Now set the RHS equal to δ and solve for t. □

1.3.5 Moving towards non-finite hypothesis classes: hinge loss and logistic regression

The analyses of Section 1.3.3 and Section 1.3.4 both suffer from one key issue: they only apply to finite function classes, i.e., when $|\mathscr{F}| < \infty$. This is an unnecessary restriction. Indeed we saw that the Perceptron algorithm, which operates over a continuum of linear hypothesis, also enjoys a strong generalization bound (cf. Lemma 1.10). Towards removing this restriction, we first observe that Perception is an instance of ERM.

[5] Technically, there could be a measure-zero event where this is not true, but we ignore this technicality.

Hinge loss and the Perceptron. We claim that the Perceptron algorithm is an instance of ERM over hypothesis class of linear function $\mathscr{F} = \{x \mapsto \langle w, x\rangle \mid w \in \mathbb{R}^d\}$. What is the loss function? Here, we slightly overload notation and consider ℓ as a univariate function $\ell(\lambda, y) = \ell(\lambda \cdot y)$. With this univariate notation, we define the *hinge loss*

$$\ell_{\text{hinge}}(s) := \max\{1 - s, 0\}.$$

Now observe that for a datapoint $(x, y) \in \mathsf{Z}$ and parameters w, we have that:

$$\nabla_w \ell_{\text{hinge}}(y\langle w, x\rangle) = -yx \cdot \mathbf{1}\{y\langle w, x\rangle < 1\}.$$

This update rule is exactly that of the Perceptron algorithm (cf. Equation (1.5)). Hence, we can see that Perceptron is precisely running stochastic gradient descent (SGD) on the following empirical risk:

$$L_n[w] = \frac{1}{n}\sum_{i=1}^n \ell_{\text{hinge}}(y_i\langle w, x_i\rangle) = \frac{1}{n}\sum_{i=1}^n \max\{1 - y_i\langle w, x_i\rangle, 0\}.$$

We will revisit SGD in more depth later on in the course. Next, we will see that even for linear models, the hinge loss is not the only well-motivated loss function.

Logistic regression. Motivated by the MAP estimator (1.3), we now consider a different type of loss function based on directly learning the posterior probability $p(y \mid x)$. Sticking with a hypothesis class of linear functions $\mathscr{F} = \{x \mapsto \langle w, x\rangle \mid w \in \mathbb{R}^d\}$, consider the following model:

$$p_w(y = 1 \mid x) \propto \exp(\langle w, x\rangle), \quad p_w(y = -1 \mid x) \propto 1.$$

Normalizing across $\mathsf{Y} = \{\pm 1\}$,

$$p_w(y = 1 \mid x) = \frac{\exp(\langle w, x\rangle)}{1 + \exp(\langle w, x\rangle)}, \quad p_w(y = -1 \mid x) = \frac{1}{1 + \exp(\langle w, x\rangle)}.$$

To construct a loss function, let us use the KL-divergence to match $p_w(y \mid x)$ with $p(y \mid x)$. Specifically, conditioned on $x \in \mathsf{X}$,

$$\begin{aligned}
&\mathrm{KL}(p(y \mid x) \,\|\, p_w(y \mid x)) - \mathbb{E}[\log p(y \mid x) \mid x] \\
&= -\mathbb{E}[\log p_w(y \mid x) \mid x] \\
&= -p(y = 1 \mid x) \log p_w(y = 1 \mid x) - p(y = -1 \mid x) \log p_w(y = -1 \mid x) \\
&= p(y = 1 \mid x) \log(1 + \exp(-\langle w, x\rangle)) + p(y = -1 \mid x) \log(1 + \exp(\langle w, x\rangle)) \\
&= \mathbb{E}[\log(1 + \exp(-y\langle w, x\rangle)) \mid x].
\end{aligned}$$

Hence, by the tower property,

$$\mathbb{E}[\mathrm{KL}(p(y \mid x) \,\|\, p_w(y \mid x))] = \mathbb{E}[\log p(y \mid x)] + \mathbb{E}[\log(1 + \exp(-y\langle w, x\rangle))].$$

Observe that the first (negative) entropy term does not depend on the parameters w. Hence, this motivates the population level estimator:

$$w_\star \in \arg\min_{w \in \mathbb{R}^d} \mathbb{E}[\ell_{\log}(y\langle w, x\rangle)], \quad \ell_{\log}(s) := \log(1 + \exp(-s)),$$

and corresponding empirical risk:

$$\hat{L}_n[w] = \frac{1}{n}\sum_{i=1}^n \ell_{\log}(y_i\langle w, x_i\rangle) = \frac{1}{n}\sum_{i=1}^n \log(1 + \exp(-y_i\langle w, x_i\rangle)).$$

(Indeed, we have $\mathbb{E}[\hat{L}_n[w]] = L[w]$ for every fixed $w \in \mathbb{R}^d$.) The loss $\ell_{\log}(s)$ is the *logistic loss*, and this setup is referred to as *logistic regression*.

Exercise 1.17. Suppose that we are in the multi-class classification setting, with $\mathsf{Y} = \{1, \dots, K\}$. Suppose that we posit the following probabilistic model:

$$p_{w_{1:K}}(y = k \mid x) \propto \exp(\langle w_k, x \rangle),\ k \in \{1, \dots, K\}.$$

Show that the corresponding empirical risk for the population risk

$$L[w] = \mathbb{E}[\mathrm{KL}(p(y \mid x) \parallel p_{w_{1:K}}(y \mid x))]$$

is given by the following *cross-entropy loss*:

$$\hat{L}_n[w] = \frac{1}{n} \sum_{i=1}^{n} \left[\langle w_{y_i}, x_i \rangle - \log \left[\sum_{j=1}^{K} \exp(\langle w_j, x_i \rangle) \right] \right].$$

That is, show that $\mathbb{E}[\hat{L}_n[w]] = L[w]$ for every fixed $w = \{w_k\}_{k=1}^K$.

1.3.6 Margin theory

We have seen two examples of loss functions: the hinge loss and the logistic loss. However, at this point two key questions remain to be answered:

(a) Both our derivations of the hinge loss and logistic loss were based on linear predictors. Can we apply these specific losses beyond linear models?

(b) Ultimately, our goal is to access model quality via the zero-one loss. What is the relationship between the hinge/logistic loss and the zero-one loss? Does low population risk under the former imply low classification error?

Resolving (a) is simple. Observe that both the hinge and logistic loss involve the corresponding linear model $x \mapsto \langle w, x \rangle$ only through quantities of the form $y\langle w, x \rangle$. Hence, for a possibly non-linear hypothesis $f : \mathsf{X} \mapsto \mathbb{R}$, we construct the following population and empirical loss:

$$L[f] = \mathbb{E}[\ell(yf(x))], \quad \hat{L}_n[f] = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i f(x_i)), \quad \ell \in \{\ell_{\text{hinge}}, \ell_{\log}\}.$$

We now turn to resolving (b). For both hinge and logistic loss, there is actually a simple answer. Indeed, a key property of both losses is that they both dominate the zero-one loss in the following way. Specifically, consider a model $f : \mathsf{X} \mapsto \mathbb{R}$, and following similar arguments as we did to conclude Section 1.3.3:

$$\mathbf{1}\{y \neq \mathrm{sgn}(f(x))\} \leqslant \mathbf{1}\{yf(x) \leqslant 0\}.$$

This characterization of the zero-one loss can then be used to show that both losses dominate the zero-one loss, as formalized by the following exercise.

Exercise 1.18. Show that there exists a universal constant $c > 0$ such that:

$$\forall s \in \mathbb{R},\ \mathbf{1}\{s \leqslant 0\} \leqslant \ell_{\text{hinge}}(s) \text{ and } \mathbf{1}\{s \leqslant 0\} \leqslant c \cdot \ell_{\log}(s).$$

What is the sharpest constant c?

Hence, having low population risk under both the hinge or logistic loss implies low classification error, i.e.,

$$\mathbb{P}\{\mathrm{sgn}(f(x)) \neq y\} \leqslant \mathbb{E}[\ell_{\text{hinge}}(yf(x))] \quad \text{and} \quad \mathbb{P}\{\mathrm{sgn}(f(x)) \neq y\} \leqslant c \cdot \mathbb{E}[\ell_{\log}(yf(x))].$$

We can go beyond these two losses by considering a *family* of *ramp losses* parameterized by $\gamma > 0$:

$$\ell_\gamma(s) = \begin{cases} 1 & \text{if } s < 0 \\ 1 - s/\gamma & \text{if } s \in [0, \gamma], \\ 0 & \text{if } s > \gamma \end{cases}.$$

By construction, this loss is sandwiched by the following inequalities:

$$\mathbf{1}\{s \leqslant 0\} \leqslant \ell_\gamma(s) \leqslant \mathbf{1}\{s < \gamma\}. \tag{1.10}$$

Consequently, for every $f \in \mathscr{F}$,

$$\begin{aligned} \mathbb{P}\{y \neq \operatorname{sgn}(f(x))\} &\leqslant \mathbb{E}[\ell_\gamma(yf(x))] && \text{using (1.10)} \\ &\leqslant \frac{1}{n}\sum_{i=1}^n \ell_\gamma(y_i f(x_i)) + \sup_{f \in \mathscr{F}} \operatorname{gen}[f; \ell_\gamma] && \text{recall } \operatorname{gen}[f] = L[f] - \hat{L}_n[f] \\ &\leqslant \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{y_i f(x_i) < \gamma\} + \sup_{f \in \mathscr{F}} \operatorname{gen}[f; \ell_\gamma] && \text{using (1.10).} \end{aligned} \tag{1.11}$$

(Here, we use the notation $\operatorname{gen}[f; \ell_\gamma]$ to emphasize that the ramp loss is involved in the generalization bound). This derivation states that we can control the zero-one loss in general by:

(a) Controlling the generalization error $\sup_{f \in \mathscr{F}} \operatorname{gen}[f; \ell_\gamma]$ under the ramp loss ℓ_γ, and

(b) Counting the number of mistakes $y_i f(x_i) < \gamma$ on the training dataset S_n. Recall that if f is linear, then we defined (Definition 1.5) the margin of a correctly classified point (x, y) as $y\langle w, x\rangle / \|w\|$. In the general case, the quantity $yf(x)$ serves as the *un-normalized* margin.

Note that, $\hat{f}_{\text{erm}}$ can actually optimize any loss function not necessarily related to ℓ_γ and still apply the bound from Equation (1.11). We shall see shortly studying generalization of ℓ_γ in the non-finite hypothesis class case, is much simpler than directly trying to study the generalization over the zero-one loss, thanks to the Lipschitz continuity of ℓ_γ (observe that ℓ_γ is γ^{-1}-Lipschitz).

1.3.7 Rademacher complexity

Towards a theory for general function classes, we first observe that simply counting the number of parameters of a function class is insufficient. Instead we must consider the geometric structure. For instance, consider the following linear hypothesis classes:

$$\mathscr{F}_p = \{x \mapsto \langle w, x\rangle \mid w \in \mathbb{R}^d,\ \|w\|_p \leqslant 1\},\ p \in [1, \infty].$$

Note that for all p, $\mathscr{F}_p$ is parameterized by d parameters. However, in high dimension, the parameter spaces of these sets are drastically different. By a volume calculation,

$$\operatorname{Vol}(\{\|w\|_1 \leqslant 1\}) = 2^d/d!, \quad \operatorname{Vol}(\{\|w\|_2 \leqslant 1\}) = \pi^{d/2}/\Gamma(1 + d/2), \quad \operatorname{Vol}(\{\|w\|_\infty \leqslant 1\}) = 2^d.$$

Note that as $d \to \infty$, the first two volumes tend to 0, whereas the last tends to ∞. Hence, these parameter sets are geometrically very different, despite all sets being described by d parameters.

To address this issue, we now introduce a powerful tool for analyzing generalization error, which captures the geometry of the underlying hypothesis class.

Definition 1.19 (Rademacher complexity)**.** *Let $\mathscr{F}$ be a set of functions mapping $\mathsf{X} \mapsto \mathbb{R}$. The Rademacher complexity of $\mathscr{F}$ is defined as:*

$$\mathcal{R}_n(\mathscr{F}) := \mathbb{E} \sup_{f \in \mathscr{F}} n^{-1} \sum_{i=1}^n \varepsilon_i f(x_i),$$

where $\{\varepsilon_i\}_{i=1}^n$ are independent Rademacher random variables (and also independent of the $\{x_i\}_{i=1}^n$). Note that the expectation is taken jointly over both the data $\{x_i\}$ and the Rademacher random variables $\{\varepsilon_i\}$.

The power of the Rademacher complexity comes from the following classical symmetrization lemma.

Proposition 1.20 (Symmetrization)**.** *Let $\{x_i\}_{i=1}^n$ be independent random variables in X, and let $\mathscr{F}$ be a set of functions mapping $\mathsf{X} \mapsto \mathbb{R}$. We have that:*

$$\mathbb{E} \sup_{f \in \mathscr{F}} n^{-1} \sum_{i=1}^n (\mathbb{E}[f(x_i)] - f(x_i)) \leqslant 2\mathcal{R}_n(\mathscr{F}).$$

Proof. Let $\{x_i'\}$ be an independent copy of $\{x_i\}$, and let $\{\varepsilon_i\}$ be independent Rademacher random variables. We have:

$$\begin{aligned}
&\mathbb{E} \sup_{f \in \mathscr{F}} n^{-1} \sum_{i=1}^n (\mathbb{E}[f(x_i)] - f(x_i)) \\
&= \mathbb{E} \sup_{f \in \mathscr{F}} n^{-1} \sum_{i=1}^n (\mathbb{E}[f(x_i')] - f(x_i)) && \text{since } x_i \overset{d}{=} x_i' \\
&= \mathbb{E}_{\{x_i\}} \sup_{f \in \mathscr{F}} \mathbb{E}_{\{x_i'\}} \left[n^{-1} \sum_{i=1}^n (f(x_i') - f(x_i)) \right] && \text{linearity of expectation} \\
&\leqslant \mathbb{E} \sup_{f \in \mathscr{F}} n^{-1} \sum_{i=1}^n (f(x_i') - f(x_i)) && \text{since } \sup_f \mathbb{E}[Z_f] \leqslant \mathbb{E} \sup_f Z_f \\
&= \mathbb{E} \sup_{f \in \mathscr{F}} n^{-1} \sum_{i=1}^n \varepsilon_i (f(x_i') - f(x_i)) && \text{since } f(x_i') - f(x_i) \overset{d}{=} \varepsilon_i (f(x_i') - f(x_i)) \\
&\leqslant \mathbb{E} \sup_{f \in \mathscr{F}} n^{-1} \sum_{i=1}^n \varepsilon_i f(x_i') + \mathbb{E} \sup_{f \in \mathscr{F}} n^{-1} \sum_{i=1}^n \varepsilon_i f(x_i) && \text{since } \sup_x [f(x) + g(x)] \leqslant \sup_x f(x) + \sup_x g(x) \\
&= 2\mathcal{R}_n(\mathscr{F}) && \text{since } x_i \overset{d}{=} x_i'.
\end{aligned}$$

□

Next, we prove a high probability deviation bound on the quantity $\sup_{f \in \mathscr{F}} n^{-1} \sum_{i=1}^n (\mathbb{E}[f(x_i)] - f(x_i))$ using the bounded-differences inequality (Proposition B.16).

Proposition 1.21. *Let $\{x_i\}_{i=1}^n$ be independent random variables, and let $\mathscr{F}$ be a set of uniformly bounded functions mapping $\mathsf{X} \mapsto [-B, B]$. Then, with probability at least $1 - \delta$:*

$$\sup_{f \in \mathscr{F}} n^{-1} \sum_{i=1}^n (\mathbb{E}[f(x_i)] - f(x_i)) \leqslant 2\mathcal{R}_n(\mathscr{F}) + 2B\sqrt{\frac{2\log(1/\delta)}{n}}.$$

Proof. Let $h(x_{1:n}) := \sup_{f \in \mathscr{F}} n^{-1} \sum_{i=1}^n (\mathbb{E}[f(x_i)] - f(x_i))$. Pick any index $i \in \{1, \dots n\}$, and let $x_1, \dots, x_n, x_i'$ be arbitrary. Observe that, since $\mathscr{F}$ is uniformly bounded,

$$\begin{aligned}
&|h(x_{1:i-1}, x_i, x_{i+1:n}) - h(x_{1:i-1}, x_i', x_{i+1:n})| \\
&= \left| \sup_{f \in \mathscr{F}} n^{-1} \sum_{i=1}^n (\mathbb{E}[f(x_i)] - f(x_i)) - \sup_{f \in \mathscr{F}} n^{-1} \left[(\mathbb{E}[f(x_i)] - f(x_i')) + \sum_{i \neq i}^n (\mathbb{E}[f(x_i)] - f(x_i)) \right] \right| \\
&\leqslant \frac{1}{n} \sup_{f \in \mathscr{F}} |f(x_i) - f(x_i')| \leqslant \frac{2B}{n}.
\end{aligned}$$

Therefore the function h satisfies bounded-differences, and we can apply the bounded-differences inequality (Proposition B.16) to conclude:

$$\mathbb{P}\{h(x_{1:n}) - \mathbb{E}[h(x_{1:n})] \geqslant t\} \leqslant \exp\left(-nt^2/(8B^2)\right).$$

The claim now follows by using Proposition 1.20 to bound $\mathbb{E}[h(x_{1:n})] \leqslant 2\mathcal{R}_n(\mathscr{F})$. □

Finally, we state a key property of Rademacher complexities which we will need in our analysis.

Proposition 1.22. *Let $\ell : \mathbb{R} \mapsto \mathbb{R}$ be L-Lipschitz, i.e.,:*

$$\forall x, y \in \mathbb{R},\ |\ell(x) - \ell(y)| \leqslant L|x - y|.$$

Let $\mathscr{F}$ be a function class mapping $\mathsf{X} \mapsto \mathbb{R}$. We have that:

$$\mathcal{R}_n(\{\ell \circ f \mid f \in \mathscr{F}\}) \leqslant L \cdot \mathcal{R}_n(\mathscr{F}).$$

The proof of Proposition 1.22 is almost a direct consequence of the following special case of Ledoux and Talagrand's contraction lemma.

Lemma 1.23 (Ledoux and Talagrand Contraction)**.** *Let $\phi : \mathbb{R} \mapsto \mathbb{R}$ be a contraction, i.e.,*

$$\forall x, y \in \mathbb{R},\ |\phi(x) - \phi(y)| \leqslant |x - y|.$$

Let $T \subset \mathbb{R}^n$ be any set. We have that:

$$\mathbb{E} \sup_{t \in T} \sum_{i=1}^{n} \varepsilon_i \phi(t_i) \leqslant \mathbb{E} \sup_{t \in T} \sum_{i=1}^{n} \varepsilon_i t_i.$$

Proof. We will prove that, for every index $i \in \{1, \ldots, n\}$ and any arbitrary $A : T \mapsto \mathbb{R}$,

$$\mathbb{E} \sup_{t \in T}[A(t) + \varepsilon\phi(t_i)] \leqslant \mathbb{E} \sup_{t \in T}[A(t) + \varepsilon t_i]. \tag{1.12}$$

We start by expanding the LHS of (1.12):

$$\mathbb{E} \sup_{t \in T}[A(t) + \varepsilon\phi(t_i)] = \frac{1}{2} \sup_{t \in T}[A(t) + \phi(t_i)] + \frac{1}{2} \sup_{t \in T}[A(t) - \phi(t_i)].$$

Suppose that $u \in T$ achieves the supremum for the first term on the RHS, and $v \in T$ achieves the supremum for the second term on the RHS (if both supremums are not achieved, consider sequences $\{u^{(i)}\}, \{v^{(i)}\}$ in T which approach the supremums instead). First, let us suppose that $u_i \leqslant v_i$. Then, the contraction of ϕ yields $\phi(u_i) - \phi(v_i) \leqslant v_i - u_i$, and therefore

$$\begin{aligned}
\frac{1}{2}(A(u) + \phi(u_i)) + \frac{1}{2}(A(v) - \phi(v_i)) &\leqslant \frac{1}{2}(A(u) - u_i) + \frac{1}{2}(A(v) + v_i) \\
&\leqslant \frac{1}{2} \sup_{t \in T}(A(t) - t_i) + \frac{1}{2} \sup_{t \in T}(A(t) + t_i) \qquad \text{since } u, v \in T \\
&= \mathbb{E} \sup_{t \in T}[A(t) + \varepsilon t_i].
\end{aligned}$$

On the other hand, if $v_i \leqslant u_i$, then $\phi(u_i) - \phi(v_i) \leqslant u_i - v_i$, and a nearly identical argument yields:

$$\frac{1}{2}(A(u) + \phi(u_i)) + \frac{1}{2}(A(v) - \phi(v_i)) \leqslant \frac{1}{2}(A(u) + u_i) + \frac{1}{2}(A(v) - v_i) \leqslant \mathbb{E} \sup_{t \in T}[A(t) + \varepsilon t_i].$$

Combining these two cases, we have shown that (1.12) holds. To conclude the proof, we use the independence of the Rademacher random variables $\{\varepsilon_i\}$ along with repeated applications of (1.12) to "peel" off the ϕ in a step by step manner:

$$\mathbb{E} \sup_{t \in T} \sum_{i=1}^{n} \varepsilon_i \phi(t_i) = \mathbb{E}_{\varepsilon_{\neg n}} \mathbb{E}_{\varepsilon_n} \sup_{t \in T} \left[\sum_{i=1}^{n-1} \varepsilon_i \phi(t_i) + \varepsilon_n \phi(t_n) \right]$$

$$
\begin{aligned}
&\leqslant \mathbb{E}_{\varepsilon_{\neg n}} \mathbb{E}_{\varepsilon_n} \sup_{t \in T} \left[\sum_{i=1}^{n-1} \varepsilon_i \phi(t_i) + \varepsilon_n t_n \right] && \text{using (1.12)} \\
&= \mathbb{E}_{\varepsilon_{\neg n-1}} \mathbb{E}_{\varepsilon_{n-1}} \sup_{t \in T} \left[\left\{ \sum_{i=1}^{n-2} \varepsilon_i \phi(t_i) + \varepsilon_n t_n \right\} + \varepsilon_{n-1} \phi(t_{n-1}) \right] \\
&\leqslant \mathbb{E}_{\varepsilon_{\neg n-1}} \mathbb{E}_{\varepsilon_{n-1}} \sup_{t \in T} \left[\left\{ \sum_{i=1}^{n-2} \varepsilon_i \phi(t_i) + \varepsilon_n t_n \right\} + \varepsilon_{n-1} t_{n-1} \right] && \text{using (1.12) again} \\
&\vdots \\
&\leqslant \mathbb{E} \sup_{t \in T} \sum_{i=1}^{n} \varepsilon_i t_i.
\end{aligned}
$$

□

Exercise 1.24. Prove Proposition 1.22 using Lemma 1.23.

1.3.8 Generalization bounds via Rademacher complexity

We now have the tools in place to state our first generalization bound via Rademacher complexities.

Theorem 1.25 (Generalization bound via Rademacher complexity)**.** *Let $\ell : \mathbb{R} \mapsto \mathbb{R}$ be any L-Lipschitz loss function. Let $\mathscr{F}$ be a function class mapping $\mathsf{X} \mapsto \mathbb{R}$. Suppose that:*

$$\sup_{f \in \mathscr{F}} \sup_{(x,y) \in \mathsf{Z}} |\ell(yf(x))| \leqslant B_\ell.$$

With probability at least $1 - \delta$,

$$\forall f \in \mathscr{F},\ L[f] \leqslant \frac{1}{n} \sum_{i=1}^{n} \ell(y_i f(x_i)) + 2L \cdot \mathcal{R}_n(\mathscr{F}) + 2B_\ell \sqrt{\frac{2 \log(1/\delta)}{n}}.$$

Proof. Much as we proceeded in (1.9), we have for any $f \in \mathscr{F}$,

$$
\begin{aligned}
L[f] &= \frac{1}{n} \sum_{i=1}^{n} \ell(y_i f(x_i)) + \mathbb{E}[\ell(yf(x))] - \frac{1}{n} \sum_{i=1}^{n} \ell(y_i f(x_i)) \\
&\leqslant \frac{1}{n} \sum_{i=1}^{n} \ell(y_i f(x_i)) + \sup_{f \in \mathscr{F}} n^{-1} \sum_{i=1}^{n} (\mathbb{E}[\ell(y_i f(x_i))] - \ell(y_i f(x_i))).
\end{aligned}
$$

We use Proposition 1.21 to analyze the second term, which tells us with probability at least $1 - \delta$,

$$\sup_{f \in \mathscr{F}} n^{-1} \sum_{i=1}^{n} (\mathbb{E}[\ell(y_i f(x_i))] - \ell(y_i f(x_i))) \leqslant 2\mathcal{R}_n(\{(x,y) \mapsto \ell(yf(x)) \mid f \in \mathscr{F}\}) + 2B_\ell \sqrt{\frac{2 \log(1/\delta)}{n}}. \quad (1.13)$$

The Rademacher complexity term $\mathcal{R}_n(\{(x,y) \mapsto \ell(yf(x)) \mid f \in \mathscr{F}\})$ in (1.7) is challenging to analyze directly, since it involves the composition of two functions. However, we can use Rademacher contraction to "peel" off the loss ℓ directly exposing the Rademacher complexity of $\mathscr{F}$. Specifically, by Proposition 1.22, we have:

$$\mathcal{R}_n(\{(x,y) \mapsto \ell(yf(x)) \mid f \in \mathscr{F}\}) \leqslant L \cdot \mathbb{E} \sup_{f \in \mathscr{F}} n^{-1} \sum_{i=1}^{n} \varepsilon_i y_i f(x_i)$$

$$= L \cdot \mathbb{E} \sup_{f \in \mathscr{F}} n^{-1} \sum_{i=1}^{n} \varepsilon_i f(x_i) \qquad \varepsilon_i y_i \overset{d}{=} \varepsilon_i \text{ by symmetry}$$
$$= L \cdot \mathcal{R}_n(\mathscr{F}).$$

Recalling our discussion on margin theory (Section 1.3.6), an immediate corollary of Theorem 1.25 is the following margin bound.

Corollary 1.26 (Margin generalization bound)**.** *Fix any $\gamma > 0$. Let $\mathscr{F}$ be a function class mapping $\mathsf{X} \mapsto \mathbb{R}$. With probability at least $1 - \delta$,*

$$\forall f \in \mathscr{F},\ \mathbb{P}\{\operatorname{sgn}(f(x)) \neq y\} \leqslant \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{y_i f(x_i) < \gamma\} + \frac{2}{\gamma} \cdot \mathcal{R}_n(\mathscr{F}) + 2\sqrt{\frac{2\log(1/\delta)}{n}}.$$

Proof. Apply Theorem 1.25 to the ramp loss ℓ_γ (which satisfies $|\ell_\gamma| \leqslant 1$ and ℓ_γ is γ^{-1}-Lipschitz) from Section 1.3.6, and then apply the ramp loss inequality $\ell_\gamma(s) \leqslant \mathbf{1}\{s < \gamma\}$ from (1.10) to conclude. □

Exercise 1.27. One drawback of Corollary 1.26 is that it only applies for a *fixed* margin γ. Show that one can extend the bound to apply for all margins as follows. Suppose that $\mathscr{F}$ is uniformly bounded, so that $|f| \leqslant B$ for $f \in \mathscr{F}$. With probability at least $1 - \delta$, for all $f \in \mathscr{F}$,

$$\mathbb{P}\{\operatorname{sgn}(f(x)) \neq y\} \leqslant \inf_{\gamma \in [0,B]} \left[\frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{y_i f(x_i) < \gamma\} + \frac{4}{\gamma} \cdot \mathcal{R}_n(\mathscr{F}) + c\sqrt{\frac{\log\log(B/\gamma) + \log(1/\delta)}{n}} \right],$$

where $c > 0$ is a universal constant.

At this point, a natural question to ask is why Theorem 1.25 is applied to the ramp loss ℓ_γ and not directly to the zero-one loss. While the easy answer is that the zero-one loss is not Lipschitz thanks to the discontinuity at the origin, inspecting the proof of Theorem 1.25 shows that the Lipschitzness of the loss is only used to control the final Rademacher complexity term. Indeed, the proof up to and including (1.13) can be replayed with the zero-one loss substituted in. The Rademacher complexity term to be controlled then becomes:

$$\begin{aligned}
\mathbb{E} \sup_{f \in \mathscr{F}} n^{-1} \sum_{i=1}^{n} \varepsilon_i \mathbf{1}\{y_i \neq \operatorname{sgn}(f(x_i))\} &= \mathbb{E} \sup_{f \in \mathscr{F}} n^{-1} \sum_{i=1}^{n} \varepsilon_i \left(\frac{1 - \operatorname{sgn}(f(x_i)) y_i}{2} \right) && \text{using Exercise 1.28} \\
&= \frac{1}{2} \mathbb{E} \sup_{f \in \mathscr{F}} n^{-1} \sum_{i=1}^{n} \varepsilon_i (-y_i) \operatorname{sgn}(f(x_i)) \\
&= \frac{1}{2} \mathbb{E} \sup_{f \in \mathscr{F}} n^{-1} \sum_{i=1}^{n} \varepsilon_i \operatorname{sgn}(f(x_i)) && \text{since } \varepsilon_i \overset{d}{=} -\varepsilon_i y_i. \qquad (1.14)
\end{aligned}$$

We will see later that the quantity above can be controlled by the VC-dimension of the predictor class $\mathscr{H} = \{x \mapsto \operatorname{sgn}(f(x)) \mid f \in \mathscr{F}\}$. The VC-dimension, an classical object in the machine learning literature, is a combinatorial quantity which indicates the cardinality of the largest dataset which can be *shattered* by a predictor. However, this approach of going through the VC-dimension has its drawbacks, one of which is the geometric structure of the underlying hypothesis class $\mathscr{F}$ is abstracted into a combinatorial quantity, rather than directly exposed via the Rademacher complexity. The latter is much easier to work with as we will see shortly.

Exercise 1.28. Show that for all $\lambda \in \mathbb{R}$ and $y \in \{\pm 1\}$,

$$\mathbf{1}\{y \neq \operatorname{sgn}(\lambda)\} = \frac{1 - \operatorname{sgn}(\lambda) y}{2}.$$

1.3.9 Explicit computations of Rademacher complexity

We will first consider Rademacher complexities of linear function classes, $\mathscr{F}_p = \{x \mapsto \langle x, w\rangle \mid \|w\|_p \leqslant 1\}$. We will also assume that $x_i \sim N(0, I)$ to simplify expressions as much as possible. Our first computation is to simplify $\mathcal{R}_n(\mathscr{F}_p)$. Let q be dual to p, i.e., $1/p + 1/q = 1$. We have,

$$\mathcal{R}_n(\mathscr{F}_p) = \mathbb{E} \sup_{\|w\|_p \leqslant 1} n^{-1} \sum_{i=1}^n \varepsilon_i \langle x_i, w\rangle = n^{-1}\mathbb{E} \sup_{\|w\|_p \leqslant 1} \left\langle w, \sum_{i=1}^n \varepsilon_i x_i \right\rangle = n^{-1}\mathbb{E} \left\| \sum_{i=1}^n \varepsilon_i x_i \right\|_q . \tag{1.15}$$

That is, the Rademacher complexity of $\mathscr{F}_p$ involves computing the dual ℓ_q-norm of a random vector sum. For $p = q = 2$, Jensen's inequality yields a simple upper bound:

$$\mathcal{R}_n(\mathscr{F}_2) = n^{-1}\mathbb{E} \left\| \sum_{i=1}^n \varepsilon_i x_i \right\| \leqslant n^{-1} \sqrt{\mathbb{E} \left\| \sum_{i=1}^n \varepsilon_i x_i \right\|^2} = n^{-1} \sqrt{\sum_{i=1}^n \mathbb{E}\|x_i\|^2} = \sqrt{\frac{d}{n}}. \tag{1.16}$$

This argument is actually fairly sharp, as shown in the following exercise.

Exercise 1.29. Show that $\mathcal{R}_n(\mathscr{F}_2) \geqslant c\sqrt{d/n}$ for a universal constant c.

Hint: You may use without proof the following fact, known as the *Gaussian Poincaré inequality*. Let $g \in \mathbb{R}^k$ be an isotropic Gaussian random vector. For any $f : \mathbb{R}^k \mapsto \mathbb{R}$, we have that

$$\mathrm{Var}(f(g)) \leqslant \mathbb{E}\|\nabla f(g)\|^2.$$

Note that this fact is dimension free.

We now turn to study several other linear hypothesis spaces.

Exercise 1.30. Show that $\mathcal{R}_n(\mathscr{F}_\infty) = 2d/(\pi\sqrt{n})$.

Proposition 1.31. *We have that:*

$$\mathcal{R}_n(\mathscr{F}_1) \leqslant \sqrt{\frac{2\log(2d)}{n}}.$$

Proof. Fix an index $j \in \{1, \ldots, d\}$. It is not hard to see that $Z_j := \sum_{i=1}^n \varepsilon_i \langle x_i, e_j\rangle \overset{d}{=} N(0, n)$ (why?). Furthermore $Z_j \perp Z_k$ when $j \neq k$. Hence by Exercise B.18,

$$\mathbb{E} \left\| \sum_{i=1}^n \varepsilon_i x_i \right\|_\infty = \mathbb{E} \max_{j=1,\ldots,d} |Z_j| \leqslant \sqrt{2n\log(2d)}.$$

The claim now follows from the dual ℓ_∞-norm characterization of $\mathcal{R}_n(\mathscr{F}_1)$ (cf. Equation (1.15)). □

Exercise 1.32. Let $\mathscr{F}_{2,s}$ denote the following sparse hypothesis class:

$$\mathscr{F}_{2,s} := \{x \mapsto \langle w, x\rangle \mid w \in \mathbb{R}^d,\ \|w\|_0 \leqslant s,\ \|w\| \leqslant 1\}.$$

Show that there exists a universal positive constant c_0 such that,

$$\mathcal{R}_n(\mathscr{F}_{2,s}) \leqslant c_0\sqrt{\frac{s\log(d/s)}{n}}.$$

Hint: Use the following fact, which you may take for granted without proof. Let $g \sim N(0, I)$ be an isotropic random Gaussian vector. Then, the random variable $\|g\|$ is c_1-sub-Gaussian, for a universal positive constant. Note that this fact is also dimension free.

Exercise 1.33. Construct a hypothesis class $\mathscr{F} \subseteq [-1,1]^{\mathsf{X}}$ such that $\mathcal{R}_n(\mathscr{F}) \geqslant 1$ for all $n \geqslant 1$.

Next, we will study the hypothesis space of a single hidden-layer neural network with ReLU activations. The following example comes from Ma [2022, Section 5.3].

Proposition 1.34. *Let $\phi(x) = \max\{x, 0\}$ be the ReLU activation function. Consider the hypothesis space of single hidden-layer neural networks with width $m \in \mathbb{N}_+$:*

$$\mathscr{F} = \left\{ x \mapsto \sum_{h=1}^{m} w_h \phi(\langle u_h, x \rangle) \,\middle|\, \{w_h\}_{h=1}^m \subset \mathbb{R},\, \{u_h\}_{h=1}^m \subset \mathbb{R}^d,\, \sum_{h=1}^{m} |w_i| \|u_h\| \leqslant 1 \right\}.$$

We have that:

$$\mathcal{R}_n(\mathscr{F}) \leqslant 2\sqrt{\frac{d}{n}}.$$

Proof. Let $\theta = (\{w_h\}, \{u_h\})$ denote the parameters of $\mathscr{F}$, and let Θ denote the parameter set. We have:

$$\begin{aligned}
n \cdot \mathcal{R}_n(\mathscr{F}) &= \mathbb{E} \sup_{\theta \in \Theta} \sum_{i=1}^{n} \varepsilon_i \sum_{h=1}^{m} w_h \phi(\langle u_h, x_i \rangle) \\
&= \mathbb{E} \sup_{\theta \in \Theta} \sum_{h=1}^{m} w_h \sum_{i=1}^{n} \varepsilon_i \phi(\langle u_h, x_i \rangle) && \text{by re-arranging sums} \\
&= \mathbb{E} \sup_{\theta \in \Theta} \sum_{h=1}^{m} w_h \|u_h\| \sum_{i=1}^{n} \varepsilon_i \phi(\langle u_h, x_i \rangle / \|u_h\|) && \text{since ReLU is positive homogeneous} \\
&\leqslant \mathbb{E} \sup_{\theta \in \Theta} \max_{h=1,\ldots,m} \left| \sum_{i=1}^{n} \varepsilon_i \phi(\langle u_h, x_i \rangle / \|u_h\|) \right| && \text{since } \sum_{h=1}^{m} |w_i| \|u_h\| \leqslant 1 \\
&\leqslant \mathbb{E} \sup_{\|u\| \leqslant 1} \left| \sum_{i=1}^{n} \varepsilon_i \phi(\langle u, x_i \rangle) \right| && \text{since } u_h / \|u_h\| \in B_2^d.
\end{aligned}$$

We now prove an quick fact that will allow us to remove the absolute value in the last inequality above. Let $T \subset \mathbb{R}^n$ be an arbitrary set. Suppose we are guaranteed that for every $\varepsilon \in \{\pm 1\}^n$, $\exists\, t \in T$ such that $\langle \varepsilon, t \rangle \geqslant 0$. Then, we have:

$$\forall \varepsilon \in \{\pm 1\}^n,\ \sup_{t \in T} \langle \varepsilon, t \rangle = \sup_{t \in T} \left[\langle \varepsilon, t \rangle \mathbf{1}\{\langle \varepsilon, t \rangle \geqslant 0\} \right]. \tag{1.17}$$

(Check this!). Hence for any $\varepsilon \in \{\pm 1\}^n$:

$$\begin{aligned}
\sup_{t \in T} |\langle \varepsilon, t \rangle| &= \sup_{t \in T} \left[\langle \varepsilon, t \rangle \mathbf{1}\{\langle \varepsilon, t \rangle \geqslant 0\} - \langle \varepsilon, t \rangle \mathbf{1}\{\langle \varepsilon, t \rangle \leqslant 0\} \right] \\
&\leqslant \sup_{t \in T} \left[\langle \varepsilon, t \rangle \mathbf{1}\{\langle \varepsilon, t \rangle \geqslant 0\} \right] + \sup_{t \in T} \left[\langle -\varepsilon, t \rangle \mathbf{1}\{\langle -\varepsilon, t \rangle \geqslant 0\} \right] \\
&= \sup_{t \in T} \langle \varepsilon, t \rangle + \sup_{t \in T} \langle -\varepsilon, t \rangle && \text{using (1.17).}
\end{aligned}$$

Taking expectation w.r.t. $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ where the ε_i's are i.i.d. Rademacher random variables yields:

$$\mathbb{E} \sup_{t \in T} |\langle \varepsilon, t \rangle| \leqslant 2\mathbb{E} \sup_{t \in T} \langle \varepsilon, t \rangle. \tag{1.18}$$

We now define the projection set:

$$T = \{ z = (\phi(\langle u, x_i \rangle))_{i=1}^n \mid \|u\| \leqslant 1 \}.$$

Notice that $0 \in T$ always (by setting $u = 0$), so we can apply (1.18) to conclude that:

$$n \cdot \mathcal{R}_n(\mathscr{F}) \leqslant \mathbb{E} \sup_{\|u\| \leqslant 1} \left| \sum_{i=1}^n \varepsilon_i \phi(\langle u, x_i \rangle) \right| \leqslant 2\mathbb{E} \sup_{\|u\| \leqslant 1} \sum_{i=1}^n \varepsilon_i \langle x_i, u \rangle \leqslant 2\sqrt{nd}.$$

Above, the last inequality holds by our calculation in (1.16). The claim now follows. □

Exercise 1.35. Let $\phi : \mathsf{X} \mapsto \ell_2(\mathbb{N})$ be an infinite-dimensional feature map satisfying $\|\phi(x)\|_\infty \leqslant 1$ for all $x \in \mathsf{X}$. Consider the following hypothesis class:

$$\mathscr{F} = \left\{ x \mapsto \sum_{i=1}^\infty w_i \phi(x)_i \;\middle|\; \sum_{i=1}^\infty \frac{w_i^2}{\mu_i} \leqslant 1 \right\},$$

where $(\mu_i)_{i \geqslant 1}$ is a positive element in $\ell_1(\mathbb{N})$ (i.e., $\sum_{i=1}^\infty \mu_i < \infty$). Show that:

$$\mathcal{R}_n(\mathscr{F}) \leqslant \sqrt{\frac{\sum_{i=1}^n \mu_i}{n}}.$$

Proposition 1.36 (Massart's lemma)**.** *Let $Q \subset \mathbb{R}^n$ be a finite set of points, and put $M = \max_{q \in Q} n^{-1/2}\|q\|$. Let $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ denote a Rademacher random vector. Then,*

$$\mathbb{E}_\varepsilon \max_{q \in Q} n^{-1} \langle \varepsilon, q \rangle \leqslant \sqrt{\frac{2M^2 \log |Q|}{n}}.$$

Proof. For a fixed $q \in Q$, we have that $n^{-1/2}\langle \varepsilon, q \rangle$ is a $n^{-1/2}\|q\|$-sub-Gaussian random vector. Hence by Proposition B.17,

$$\mathbb{E}_\varepsilon \max_{q \in Q} n^{-1} \langle \varepsilon, q \rangle = n^{-1/2} \mathbb{E}_\varepsilon \max_{q \in Q} n^{-1/2} \langle \varepsilon, q \rangle \leqslant \sqrt{\frac{2M^2 \log |Q|}{n}}.$$

□

Corollary 1.37. *Let $\mathscr{F}$ be a finite function class, and suppose that there exists a finite M satisfying*

$$\sup_{x \in \mathsf{X}, f \in \mathscr{F}} n^{-1} \sum_{i=1}^n f(x_i)^2 \leqslant M^2.$$

Then, we have that:

$$\mathcal{R}_n(\mathscr{F}) \leqslant \sqrt{\frac{2M^2 \log |\mathscr{F}|}{n}}.$$

Proof. Let $Q(x_{1:n}) = \{(f(x_1), \ldots, f(x_n)) \in \mathbb{R}^n \mid f \in \mathscr{F}\}$ denote the empirical projection of $\mathscr{F}$ onto $\mathbb{R}^n$. By assumption, we have that for any $q \in Q(x_{1:n})$, $n^{-1/2}\|q\| \leqslant M$. Hence,

$$\begin{aligned}
\mathcal{R}_n(\mathscr{F}) &= \mathbb{E} \sup_{f \in \mathscr{F}} n^{-1} \sum_{i=1}^n \varepsilon_i f(x_i) \\
&= \mathbb{E}_{x_{1:n}} \mathbb{E}_\varepsilon \sup_{q \in Q(x_{1:n})} n^{-1} \langle \varepsilon, q \rangle \\
&\leqslant \sqrt{\frac{2M^2 \log |Q|}{n}} && \text{using Proposition 1.36} \\
&\leqslant \sqrt{\frac{2M^2 \log |\mathscr{F}|}{n}} && \text{since } |Q| \leqslant |\mathscr{F}|.
\end{aligned}$$

□

1.3.10 Vapnik-Chervonenkis dimension

With Massart's lemma in place, we can now complete the derivation started with (1.14). Recall the goal is to bound the Rademacher complexity:

$$\mathbb{E}\sup_{f\in\mathscr{F}} n^{-1}\sum_{i=1}^{n}\varepsilon_i \mathbf{1}\{\operatorname{sgn}(f(x_i))\neq y_i\} = \frac{1}{2}\mathbb{E}\sup_{f\in\mathscr{F}} n^{-1}\sum_{i=1}^{n}\varepsilon_i \operatorname{sgn}(f(x_i)).$$

As before, let $\mathscr{H}$ denote the predictor class $\mathscr{H} = \{x\mapsto \operatorname{sgn}(f(x)) \mid f\in\mathscr{F}\}$. Also, for a fixed set of datapoints $x_{1:n}$, let $\mathscr{H}(x_{1:n})\subseteq\{\pm 1\}^n$ denote the projection set:

$$\mathscr{H}(x_{1:n}) = \{(h(x_1),\dots,h(x_n)) \mid h\in\mathscr{H}\}.$$

The *growth function* $\tau_{\mathscr{H}}(n)$ is the following quantity:

$$\tau_{\mathscr{H}}(n) := \sup_{x_{1:n}} |\mathscr{H}(x_{1:n})|.$$

Finally, the *Vapnik-Chervonenkis dimension* $\operatorname{VCdim}(\mathscr{H})$ is the following quantity:

$$\operatorname{VCdim}(\mathscr{H}) := \sup\{n\in\mathbb{N}_+ \mid \exists\, x_{1:n} \text{ s.t. } |\mathscr{H}(x_{1:n})| = 2^n\}.$$

The well-known Sauer-Shelah lemma states that the growth function is upper bounded by the VC-dimension $d = \operatorname{VCdim}(\mathscr{H})$ in the following way:

$$\tau_{\mathscr{H}}(n) \leqslant \sum_{i=0}^{d}\binom{n}{i} \quad\text{and}\quad \sum_{i=0}^{d}\binom{n}{i} \leqslant \left(\frac{en}{d}\right)^d \text{when } n > d. \tag{1.19}$$

(We will not provide a proof of this classic result here). With this result in place, we have all the tools needed for the following exercise.

Exercise 1.38. Suppose that $\mathscr{H} = \{x\mapsto \operatorname{sgn}(f(x)) \mid f\in\mathscr{F}\}$ has finite $\operatorname{VCdim}(\mathscr{H}) = d$. Show that

$$\mathbb{E}\sup_{f\in\mathscr{F}} n^{-1}\sum_{i=1}^{n}\varepsilon_i \mathbf{1}\{\operatorname{sgn}(f(x_i))\neq y_i\} \leqslant c\sqrt{\frac{d\log(n/d)}{n}},$$

where $c>0$ is a universal positive constant.

1.3.11 Chaining and Dudley's inequality

Definition 1.39 (Covering number)**.** *Let (T,ρ) be a metric space. A collection of points $S=\{s_i\}\subset T$ is an* ε-covering *if for every $t\in T$ there exists an $s_i\in S$ such that $\rho(s_i,t)\leqslant\varepsilon$. The* covering number *of T at resolution ε, denoted $N(\varepsilon;T,\rho)$, is the minimum number of points needed to form an ε-cover of T.*

Definition 1.40 (Packing number)**.** *Let (T,ρ) be a metric space. A collection of points $S=\{s_i\}\subset T$ is an* ε-packing *if for any $s_i,s_j\in S$ with $s_i\neq s_j$, we have $\rho(s_i,s_j)>\epsilon$. The* packing number *of T at resolution ε, denoted $M(\varepsilon;T,\rho)$, is the maximum achievable size amongst all ε-packings of T.*

Covering and packing numbers are closely related.

Proposition 1.41. *Let (T,ρ) be a metric space. We have:*

$$M(2\varepsilon;T,\rho)\leqslant N(\varepsilon;T,\rho)\leqslant M(\varepsilon;T,\rho).$$

Proof. We will abbreviate the covering and packing numbers as $N(\varepsilon)$ and $M(\varepsilon)$, respectively. We first prove that $N(\varepsilon) \leqslant M(\varepsilon)$. Let $\{t_1, \dots, t_M\}$ be a maximal ε-packing. We claim that $\{t_1, \dots, t_M\}$ is also an ε-covering. To see this, suppose by contradiction there exists a $t \in T$ satisfying $\min_{i=1,\dots,M} \rho(t, t_i) > \varepsilon$. Then, $\{t_1, \dots, t_M, t\}$ is an ε-packing by definition with cardinality $M+1$, a contradiction.

Now we show that $M(2\varepsilon) \leqslant N(\varepsilon)$. Let $\{t_1, \dots, t_M\}$ be a maximal 2ε-packing and let $\{u_1, \dots, u_N\}$ be a minimal ε-covering. Suppose that $M \geqslant N+1$. Since $\{u_1, \dots, u_N\}$ covers T and $M > N$, there must exists a $t_i \neq t_j$ and u_k such that $t_i, t_j \in B_\rho(u_k, \varepsilon)$ (here, $B_\rho(u, \varepsilon) = \{v \in T \mid \rho(u, v) \leqslant \varepsilon\}$.) By triangle inequality, $\rho(t_i, t_j) \leqslant \rho(t_i, u_k) + \rho(u_k, t_j) \leqslant 2\varepsilon$. But, since $\{t_i\}$ is a 2ε-packing, by definition we must have $\rho(t_i, t_j) > 2\varepsilon$, a contradiction, and hence $M < N+1$ or equivalently $M \leqslant N$. □

Remark 1.42. The quantities $\log N(\varepsilon; T, \rho)$ and $\log M(\varepsilon; T, \rho)$ are often referred to as the *metric entropy* of T (at resolution ε).

Proposition 1.43 (Volume comparison inequalities)**.** *Let $T \subset \mathbb{R}^d$, and let $\|\cdot\|$ be a norm on $\mathbb{R}^d$ (not necessarily Euclidean). Let $B(x, r)$ denote the closed ball of radius r around x (measured using the $\|\cdot\|$ norm). Finally, let $\mathrm{Vol}(\cdot)$ denote the Lesbesgue measure on $\mathbb{R}^d$. For any $\varepsilon > 0$,*

$$\left(\frac{1}{\varepsilon}\right)^d \frac{\mathrm{Vol}(T)}{\mathrm{Vol}(B(0,1))} \leqslant N(\varepsilon; T, \|\cdot\|) \leqslant \frac{\mathrm{Vol}(T + B(0, \varepsilon/2))}{\mathrm{Vol}(B(0, \varepsilon/2))}.$$

Proof. We first prove the LHS inequality. Let $\{t_1, \dots, t_N\}$ be any ε-covering of T. By the definition of an ε covering, $T \subseteq \cup_{i=1}^N B(t_i, \varepsilon)$, and hence,

$$\mathrm{Vol}(T) \overset{(a)}{\leqslant} \mathrm{Vol}(\cup_{i=1}^N B(t_i, \varepsilon)) \overset{(b)}{\leqslant} \sum_{i=1}^N \mathrm{Vol}(B(t_i, \varepsilon)) \overset{(c)}{=} \sum_{i=1}^N \varepsilon^d \mathrm{Vol}(B(0,1)) = \varepsilon^d N \cdot \mathrm{Vol}(B(0,1)).$$

Above, (a) follows from monotonicity of measure, (b) follows from sub-additivity of measure, and (c) follows from the translation invariance and scaling properties of the Lebesgue measure. Rearranging the above inequality yields $N \geqslant \varepsilon^{-d} \mathrm{Vol}(T)/\mathrm{Vol}(B(0,1))$, and taking the infimum over ε-covers yields the LHS inequality.

We now proceed to the RHS inequality. Let $\{t_1, \dots, t_M\}$ be a maximal ε-packing of T. By construction,

$$\mathrm{Vol}(\cup_{i=1}^M B(t_i, \varepsilon/2)) = \sum_{i=1}^M \mathrm{Vol}(B(t_i, \varepsilon/2)) = M \cdot \mathrm{Vol}(B(0, \varepsilon/2)),$$

since the balls $B(t_i, \varepsilon/2)$ are disjoint. On the other hand, we clearly have $\cup_{i=1}^M B(t_i, \varepsilon/2) \subseteq T + B(0, \varepsilon/2)$, and hence by monotonicity of measure,

$$\mathrm{Vol}(\cup_{i=1}^M B(t_i, \varepsilon/2)) \leqslant \mathrm{Vol}(T + B(0, \varepsilon/2)).$$

Combining these two inequalities yields $M \leqslant \mathrm{Vol}(T + B(0, \varepsilon/2))/\mathrm{Vol}(B(0, \varepsilon/2))$. The claim follows by using the inequality $N(\varepsilon; T, \|\cdot\|) \leqslant M(\varepsilon; T, \|\cdot\|)$ from Proposition 1.41. □

Corollary 1.44. *Let $\|\cdot\|$ be any norm on $\mathbb{R}^d$. For any $\varepsilon > 0$,*

$$\left(\frac{1}{\varepsilon}\right)^d \leqslant N(\varepsilon; B(0,1), \|\cdot\|) \leqslant \left(1 + \frac{2}{\varepsilon}\right)^d.$$

Proof. The LHS inequality follows immediately from Proposition 1.43. For the RHS inequality, we simply observe that $B(0,1) + B(0, \varepsilon/2) \subseteq B(0, 1 + \varepsilon/2)$, in which case:

$$\mathrm{Vol}(B(0,1) + B(0, \varepsilon/2)) \leqslant (1 + \varepsilon/2)^d \mathrm{Vol}(B(0,1)).$$

The RHS inequality now follows again from Proposition 1.43. □

For a function class $\mathscr{F}$ mapping $\mathsf{X} \mapsto \mathbb{R}$ and a realization of data points $\{x_i\}_{i=1}^n$, we define the empirical $L_2(\mathsf{P}_n)$ (semi-)norm as:

$$\|f\|_{L_2(\mathsf{P}_n)}^2 := \frac{1}{n}\sum_{i=1}^n f(x_i)^2.$$

Let us make a first attempt to bound the Rademacher complexity by covering numbers. Recall Massart's lemma (Proposition 1.36), which gives a simple bound for finite function classes. Thus, we can use the idea of a covering to discretize a continuous function class, and pay a little extra for the discretization error (which is controlled by the scale of the covering).

Proposition 1.45 (One step discretization)**.** *The following holds for every realization $x_1, \dots, x_n \in \mathsf{X}$:*

$$\mathbb{E}_\varepsilon \sup_{f\in\mathscr{F}} n^{-1}\sum_{i=1}^n \varepsilon_i f(x_i) \leqslant \inf_{\alpha>0}\left\{\alpha + B_n\sqrt{\frac{2\log N(\alpha;\mathscr{F},L_2(\mathsf{P}_n))}{n}}\right\}, \quad B_n = \sup_{f\in\mathscr{F}}\|f\|_{L_2(\mathsf{P}_n)}.$$

Proof. We start with the projection set $Q = \{(f(x_1),\dots,f(x_n)) \in \mathbb{R}^n \mid f \in \mathscr{F}\}$. Fix any $\alpha > 0$ and let C denote a minimal α-covering of Q in the weighted $x \mapsto n^{-1/2}\|x\|$ norm. For any $q \in Q$ let $\varphi(q)$ denote the element in C of minimal distance, such that $n^{-1/2}\|q - \varphi(q)\| \leqslant \alpha$. We can now proceed by:

$$\begin{aligned}\mathbb{E}\sup_{q\in Q} n^{-1}\langle\varepsilon, q\rangle &= \mathbb{E}\sup_{q\in Q} n^{-1}\langle\varepsilon, \varphi(q) + q - \varphi(q)\rangle \\ &\leqslant \mathbb{E}\sup_{q\in Q} n^{-1}\langle\varepsilon,\varphi(q)\rangle + \mathbb{E}\sup_{q\in Q} n^{-1}\langle\varepsilon, q-\varphi(q)\rangle.\end{aligned}$$

Now, we first observe that:

$$n^{-1}\langle\varepsilon, q-\varphi(q)\rangle \leqslant n^{-1}\|\varepsilon\|\|q-\varphi(q)\| = n^{-1/2}\|q-\varphi(q)\| \leqslant \alpha,$$

and hence the second term is bounded by α. As for the first term, we apply Massart's lemma (Proposition 1.36), noting that $n^{-1/2}\|q\| \leqslant B_n$, and therefore,

$$\mathbb{E}\sup_{q\in Q} n^{-1}\langle\varepsilon,\varphi(q)\rangle \leqslant B_n\sqrt{\frac{2\log|C|}{n}} \leqslant B_n\sqrt{\frac{2\log N(\alpha;\mathscr{F},L_2(\mathsf{P}_n))}{n}}.$$

Since $\alpha > 0$ was arbitrary, we can optimize the bound over α, from which the claim follows. □

While the one step discretization argument is certainly a step in the right direction in terms of using covering numbers to bound Rademacher complexity, as we will see later it can be un-necessarily loose. The issue is that the argument is quite conservative by only choosing to discretize at one resolution. By repeatedly discretizing at finer and finer resolutions, we get a sharper result known as Dudley's inequality.

Lemma 1.46 (Dudley's entropy inequality)**.** *The following holds for every realization $x_1, \dots, x_n \in \mathsf{X}$:*

$$\mathbb{E}_\varepsilon \sup_{f\in\mathscr{F}} n^{-1}\sum_{i=1}^n \varepsilon_i f(x_i) \leqslant 12\int_0^{B_n}\sqrt{\frac{\log N(\varepsilon;\mathscr{F},L_2(\mathsf{P}_n))}{n}}\,\mathrm{d}\varepsilon, \quad B_n = \sup_{f\in\mathscr{F}}\|f\|_{L_2(\mathsf{P}_n)}.$$

Proof. Put $\varepsilon_0 := B_n$ and for $i = 1, 2, \dots$, define a sequence of decreasing resolutions $\varepsilon_i = \varepsilon_0 2^{-i}$. As before, let the projection set Q be

$$Q = \{(f(x_1),\dots,f(x_n)) \in \mathbb{R}^n \mid f \in \mathscr{F}\}.$$

With the definition of ε_0, observe that for any $q \in Q$,

$$n^{-1/2}\|q\| = \sqrt{\frac{1}{n}\sum_{i=1}^n f(x_i)^2} \leqslant \varepsilon_0.$$

For any $i = 1, 2, \ldots$, let C_i denote a minimal ε_i-cover of Q in the weighted ℓ_2 norm $n^{-1/2}\|\cdot\|$. Furthermore, for any $q \in Q$, let $\varphi_i(q)$ denote an element in C_i satisfying $n^{-1/2}\|q - \varphi_i(q)\| \leqslant \varepsilon_i$. Fix any $q \in Q$, and note that for any finite N,

$$\varphi_1(q) + \sum_{i=1}^{N} (\varphi_{i+1}(q) - \varphi_i(q)) = \varphi_{N+1}(q).$$

Hence, since $\lim_{i\to\infty} \varphi_i(q) = q$,

$$q = \varphi_1(q) + \sum_{i=1}^{\infty} (\varphi_{i+1}(q) - \varphi_i(q)).$$

Therefore,

$$\begin{aligned}
\mathbb{E} \sup_{q\in Q} n^{-1}\langle \varepsilon, q\rangle &= \mathbb{E} \sup_{q\in Q} n^{-1} \left\langle \varepsilon, \varphi_1(q) + \sum_{i=1}^{\infty} (\varphi_{i+1}(q) - \varphi_i(q)) \right\rangle \\
&\leqslant \mathbb{E} \sup_{q\in Q} n^{-1}\langle \varepsilon, \varphi_1(q)\rangle + \sum_{i=1}^{\infty} \mathbb{E} \sup_{q\in Q} n^{-1}\langle \varepsilon, \varphi_{i+1}(q) - \varphi_i(q)\rangle.
\end{aligned}$$

Let us control the second term first. First, observe that for any $q \in Q$,

$$n^{-1/2}\|\varphi_{i+1}(q) - \varphi_i(q)\| \leqslant n^{-1/2}\|q - \varphi_{i+1}(q)\| + n^{-1/2}\|q - \varphi_i(q)\| \leqslant \varepsilon_{i+1} + \varepsilon_i = 3\varepsilon_{i+1}.$$

On the other hand, the number of unique vectors $\{\varphi_{i+1}(q) - \varphi_i(q) \mid q \in Q\}$ is clearly upper bounded by $|C_{i+1}| \times |C_i|$. Hence by Massart's lemma (Proposition 1.36),

$$\mathbb{E} \sup_{q\in Q} n^{-1}\langle \varepsilon, \varphi_{i+1}(q) - \varphi_i(q)\rangle \leqslant 3\varepsilon_{i+1}\sqrt{\frac{2\log|C_{i+1}||C_i|}{n}} \overset{(a)}{\leqslant} 6\varepsilon_{i+1}\sqrt{\frac{\log|C_{i+1}|}{n}},$$

where (a) follows since $|C_i| \leqslant |C_{i+1}|$.

Now turning to the first term, we observe that $n^{-1/2}\|\varphi_1(q)\| \leqslant \varepsilon_0$, and hence again by Massart's lemma,

$$\mathbb{E} \sup_{q\in Q} n^{-1}\langle \varepsilon, \varphi_1(q)\rangle \leqslant \varepsilon_0\sqrt{\frac{2\log|C_1|}{n}} = 2\varepsilon_1\sqrt{\frac{2\log|C_1|}{n}}.$$

Hence combining both inequalities,

$$\mathbb{E} \sup_{q\in Q} n^{-1}\langle \varepsilon, q\rangle \leqslant 6\sum_{i=1}^{\infty} \varepsilon_i\sqrt{\frac{\log|C_i|}{n}} = 6\sum_{i=1}^{\infty} \varepsilon_i\sqrt{\frac{\log N(\varepsilon_i; \mathscr{F}, L_2(\mathsf{P}_n))}{n}}.$$

To finish the proof, we observe that since $\varepsilon \mapsto N(\varepsilon; \mathscr{F}, L_2(\mathsf{P}_n))$ is a monotonically decreasing function of ε, we have the following inequality:

$$(\varepsilon_i - \varepsilon_{i+1})\sqrt{\log N(\varepsilon_i; \mathscr{F}, L_2(\mathsf{P}_n))} \leqslant \int_{\varepsilon_{i+1}}^{\varepsilon_i} \sqrt{\log N(\varepsilon; \mathscr{F}, L_2(\mathsf{P}_n))}\,\mathrm{d}\varepsilon.$$

Since $\varepsilon_i = 2(\varepsilon_i - \varepsilon_{i+1})$, we conclude that

$$\begin{aligned}
6\sum_{i=1}^{\infty} \varepsilon_i\sqrt{\log N(\varepsilon_i; \mathscr{F}, L_2(\mathsf{P}_n))} &\leqslant 12\sum_{i=1}^{\infty} \int_{\varepsilon_{i+1}}^{\varepsilon_i} \sqrt{\log N(\varepsilon; \mathscr{F}, L_2(\mathsf{P}_n))}\,\mathrm{d}\varepsilon \\
&= 12\int_0^{\varepsilon_1} \sqrt{\log N(\varepsilon; \mathscr{F}, L_2(\mathsf{P}_n))}\,\mathrm{d}\varepsilon \\
&\leqslant 12\int_0^{\varepsilon_0} \sqrt{\log N(\varepsilon; \mathscr{F}, L_2(\mathsf{P}_n))}\,\mathrm{d}\varepsilon.
\end{aligned}$$

□

Remark 1.47. Dudley's inequality (Lemma 1.46) is stated conditionally on the data $\{x_i\}_{i=1}^n$. Taking another expectation over the data, we obtain:

$$\mathcal{R}_n(\mathscr{F}) \leqslant 12\mathbb{E}\int_0^{B_n} \sqrt{\frac{\log N(\varepsilon; \mathscr{F}, L_2(\mathsf{P}_n))}{n}}\,\mathrm{d}\varepsilon, \quad B_n = \sup_{f\in\mathscr{F}} \|f\|_{L_2(\mathsf{P}_n)}.$$

While Dudley's inequality is an improvement over the one step discretization argument in that it considers covering numbers at multiple resolutions, the version stated in Lemma 1.46 suffers from one key drawback, which is that in situations where $\log N(\varepsilon; \mathscr{F}, L_2(\mathsf{P}_n))$ is not integrable near the origin, the bound is vacuous. On the other hand the one step argument Proposition 1.45 does not suffer from this issue. We can, however, combine both bounds and obtain the best of both—multiscale covering numbers and truncating the lower limit of the integral before $\varepsilon \to 0$.

Exercise 1.48. Modify the proof of Lemma 1.46 to prove the following refined Dudley's inequality:

$$\mathbb{E}_\varepsilon \sup_{f\in\mathscr{F}} n^{-1}\sum_{i=1}^n \varepsilon_i f(x_i) \leqslant \inf_{\alpha>0}\left\{4\alpha + 12\int_\alpha^{B_n} \sqrt{\frac{\log N(\varepsilon; \mathscr{F}, L_2(\mathsf{P}_n))}{n}}\,\mathrm{d}\varepsilon\right\}.$$

Let us now apply Dudley's inequality to study linear hypothesis classes, as we did in Section 1.3.9. Specifically, let us consider $\mathscr{F} = \{x \mapsto \langle x, w\rangle \mid w \in \mathbb{R}^d, \|w\| \leqslant 1\}$. We need to estimate the $L_2(\mathsf{P}_n)$ covering number of $\mathscr{F}$. Let $X \in \mathbb{R}^{n\times d}$ denote the data matrix with row i containing example x_i. Now observe that for any $f_u, f_v \in \mathscr{F}$,

$$\|f_u - f_v\|_{L_2(\mathsf{P}_n)}^2 = \frac{1}{n}\sum_{i=1}^n \langle x_i, u-v\rangle^2 = \frac{1}{n}(u-v)^\mathsf{T} X^\mathsf{T} X(u-v) \leqslant \frac{\lambda_{\max}(X^\mathsf{T}X)}{n}\|u-v\|^2.$$

Thus, we immediately see that, with $B_n := \|X/\sqrt{n}\|_{\mathrm{op}}$,

$$N(\varepsilon; \mathscr{F}, L_2(\mathsf{P}_n)) \leqslant N(\varepsilon/B_n, B(0,1), \|\cdot\|) \leqslant \left(1 + \frac{2B_n}{\varepsilon}\right)^d,$$

where the last estimate follows from Corollary 1.44. Now, letting B be an almost sure upper bound on B_n, we compute:

$$\begin{aligned}
\int_0^B \sqrt{\log N(\varepsilon; \mathscr{F}, L_2(\mathsf{P}_n))}\,\mathrm{d}\varepsilon &\leqslant \sqrt{d}\int_0^B \sqrt{\log(1+2B/\varepsilon)}\,\mathrm{d}\varepsilon \\
&= B\sqrt{d}\int_0^1 \sqrt{\log(1+2/\varepsilon)}\,\mathrm{d}\varepsilon && \text{change of variables } \varepsilon \leftarrow \varepsilon/B \\
&\leqslant 2B\sqrt{d} && \text{using Mathematica.}
\end{aligned}$$

Hence, from Dudley's inequality,

$$\mathcal{R}_n(\mathscr{F}) \leqslant 24B\sqrt{\frac{d}{n}}.$$

Now, as a point of comparison, observe that the one-step discretization bound Proposition 1.45 yields the looser bound $\mathcal{R}_n(\mathscr{F}) \lesssim B\sqrt{\frac{d\log(n/d)}{n}}$ by choosing $\alpha \asymp \sqrt{d/n}$. Hence, the Dudley argument allows us to shave off a log factor in the bound for this example.

Let us now compare the bound from Dudley's inequality to the direct calculations made in Section 1.3.9. Recall that we computed that if x_i are i.i.d. from $N(0, I)$, then $\mathcal{R}_n(\mathscr{F}) \leqslant \sqrt{d/n}$. Immediately we see that the Dudley bound above does not apply to case, since there is no finite almost sure bound on B. However, let us suppose that instead the x_i's are i.i.d. where each entry is an independent Rademacher random variable. Then, the first and second moments $\mathbb{E}\|x\|$ and $\mathbb{E}\|x\|^2$ match the Gaussian distribution up to constants, but we have that $B = \sqrt{d}$ (check!). Thus, we actually get a looser result $\mathcal{R}_n(\mathscr{F}) \lesssim d/\sqrt{n}$, which is not sharp. However, this can be fixed, as the contents of the following exercise:

Exercise 1.49. Suppose that the random vector x_i's are i.i.d. where each entry is an independent Rademacher random variable. Let $\mathscr{F} = \{x \mapsto \langle x, w\rangle \mid w \in \mathbb{R}^d, \|w\| \leqslant 1\}$. Show using Remark 1.47 that when $n \geqslant d$, there exists a universal constant c such that:

$$\mathcal{R}_n(\mathscr{F}) \leqslant c\sqrt{\frac{d}{n}}.$$

Hint: use the following random matrix deviation bound (which you may use without proof, although at this point you actually have all the tools to prove this inequality!). Let $X \in \mathbb{R}^{n\times d}$ be the data matrix with row i corresponding to x_i. There exists universal constants c_0, c_1 such that:

$$\forall t > 0,\ \mathbb{P}\left\{\|X\|_{\mathrm{op}} \geqslant \sqrt{n} + c_0\sqrt{d} + t\right\} \leqslant \exp(-c_1 t^2).$$

The real advantage of Dudley's inequality, however, is to utilize it beyond linear hypothesis classes.

Exercise 1.50. Let $\mathscr{F} = \{x \mapsto f_\theta(x) \mid \theta \in \Theta \subset \mathbb{R}^d, \|\theta\| \leqslant B\}$. Suppose furthermore that the following conditions hold:

$$\forall x \in \mathsf{X},\ \theta, \theta_1, \theta_2 \in \Theta,\ |f_\theta(x)| \leqslant 1,\ |f_{\theta_1}(x) - f_{\theta_2}(x)| \leqslant L\|\theta_1 - \theta_2\|.$$

Show that there exists a universal constant c such that:

$$\mathcal{R}_n(\mathscr{F}) \leqslant c\sqrt{\frac{d\log(BL)}{n}}.$$

Exercise 1.51. Let $\mathscr{F}$ denote the following 1-Lipschitz function class:

$$\mathscr{F} = \{f : [-1,1]^d \mapsto [-1,1] \mid \forall x, y \in [-1,1]^d,\ |f(x) - f(y)| \leqslant \|x - y\|_\infty\}.$$

Suppose that $d > 2$. There exists an constant c such that:

$$\mathcal{R}_n(\mathscr{F}) \leqslant \frac{c}{n^{1/d}}.$$

Hint: Use the fact that $\log N(\varepsilon; \mathscr{F}, \|\cdot\|_\infty) \leqslant (c_0/\varepsilon)^d$ for some universal constant c_0. Here, $\|f\|_\infty = \sup_{x\in[-1,1]^d} |f(x)|$ is the supremum norm over $[-1,1]^d$.

1.4 Algorithmic stability

We now study alternative techniques for ensuring generalization. One particular method is based on algorithmic stability. The main idea is that algorithms which are more robust to small changes in the data (i.e., small perturbations to the data do not drastically alter the result) generalize better. We will slightly redefine notation in this section to aide the analysis. As before, we fix a hypothesis class $\mathscr{F}$ and a loss function ℓ. However, in this second, the loss function maps $\ell : \mathscr{F} \times \mathsf{Z} \to \mathbb{R}$. That is, the loss function takes the pair of hypothesis and labelled data point, and produces a score (think $\ell(f, (x, y)) = \ell(f(x), y)$). Finally, we will introduce the notion of an algorithm, which is a mapping $A : \mathsf{Z}^n \times \Xi \to \mathscr{F}$. The set Ξ encapsulates any randomness used by the algorithm.

Definition 1.52 (Uniform stability)**.** *An algorithm $A : \mathsf{Z}^n \times \Xi \to \mathscr{F}$ is* ε-uniformly-stable *if for every pair of datasets $S, S' \in \mathsf{Z}^n$ which differ in at most one example,*

$$\sup_{z\in\mathsf{Z}} \mathbb{E}_\xi[\ell(A(S,\xi), z) - \ell(A(S',\xi), z)] \leqslant \varepsilon.$$

For what follows we need some notation. Let $\{z_i\}_{i=1}^n$ and $\{z_i'\}_{i=1}^n$ be i.i.d. draws. Put $S = (z_1, \dots, z_n)$, $S' = (z_1', \dots, z_n')$, and $S^{(i)} = (z_1, \dots, z_{i-1}, z_i', z_{i+1}, \dots, z_n)$ for $i = 1, \dots, n$. We slightly overload the $\hat{L}_n[f]$ notation from earlier by $\hat{L}_S[f] = n^{-1} \sum_{i=1}^n \ell(f, z_i)$.

1.4.1 Uniform stability implies generalization

Now, we show that uniform stability directly implies generalization.

Lemma 1.53 (Uniform stability implies generalization in expectation)**.** *Let $A : \mathsf{Z}^n \times \Xi \to \mathscr{F}$ be an ε-uniformly-stable algorithm (cf. Definition 1.52). Then,*

$$|\mathbb{E}_{S,\xi}[L[A(S,\xi)] - \hat{L}_S[A(S,\xi)]]| \leqslant \varepsilon.$$

Proof. This argument is essentially an exercise in exchangability. For what follows, we will abbreviate $A(S,\xi) = A(S)$ to reduce the notional overhead. The first trick is to observe that for any i, because (S, z) and $(S^{(i)}, z_i)$ are exchangeable pairs (since $(S, z) \overset{d}{=} (S^{(i)}, z_i)$),

$$\mathbb{E}_{S,\xi}[L[A(S)]] = \mathbb{E}_{S,z,\xi}[\ell(A(S), z)] = \mathbb{E}_{S',z_i,\xi}[\ell(A(S^{(i)}, z_i))] = \mathbb{E}_{S,S',\xi}[\ell(A(S^{(i)}), z_i)].$$

Since this identity holds for any $i = 1, \dots, n$, we can average the RHS expression as follows:

$$\mathbb{E}_{S,\xi}[L[A(S)]] = \mathbb{E}_{S,S',\xi}\left[\frac{1}{n}\sum_{i=1}^n \ell(A(S^{(i)}, z_i))\right].$$

Therefore by ε-uniform-stability,

$$\mathbb{E}_{S,\xi}[L[A(S)] - \hat{L}_S[A(S)]] = \mathbb{E}_{S,S',\xi}\left[\frac{1}{n}\sum_{i=1}^n (\ell(A(S^{(i)}, z_i)) - \ell(A(S), z_i))\right] \leqslant \varepsilon.$$

Bounding $\mathbb{E}_{S,\xi}[\hat{L}_S[A(S)] - L[A(S)]]$ proceeds in a nearly identical way, from which the claim follows. □

With this result in hand, we can utilize the bounded differences inequality (Proposition B.16) to prove a high probability bound on the generalization error.

Lemma 1.54. *Suppose that $|\ell| \leqslant M$, and that $A : \mathsf{Z}^n \times \Xi \to \mathscr{F}$ is an ε-uniformly-stable algorithm. With probability at least $1 - \delta$,*

$$\mathbb{E}_\xi[L(A(S),\xi)] \leqslant \mathbb{E}_\xi[\hat{L}_S[A(S,\xi)]] + \varepsilon + 2(\varepsilon n + B)\sqrt{\frac{2\log(1/\delta)}{n}}.$$

Proof. Again as before, we will abbreviate $A(S,\xi) = A(S)$ to reduce notational burden. Define the random variable $f(S) = \mathbb{E}_\xi[L[A(S)] - \hat{L}_S[A(S)]]$. We will show that $f(S)$ satisfies the bounded differences property. Let $S^{(i)}$ be a dataset which differs in S in the i-th location. We first observe that,

$$\mathbb{E}_\xi[L[A(S)]] - \mathbb{E}_\xi[L(A(S^{(i)}))] = \mathbb{E}_z[\mathbb{E}_\xi[\ell(A(S), z) - \ell(A(S^{(i)}), z)]].$$

Hence by the ε-uniform-stability assumption, $\mathbb{E}_\xi[L[A(S)]] - \mathbb{E}_\xi[L(A(S^{(i)}))] \leqslant \varepsilon$. Furthermore, bounding $\mathbb{E}_\xi[L(A(S^{(i)}))] - \mathbb{E}_\xi[L[A(S)]] \leqslant \varepsilon$ proceeds identically.

Next, we observe that,

$$\begin{aligned}
&\mathbb{E}_\xi[\hat{L}_S[A(S)] - \hat{L}_{S^{(i)}}[A(S^{(i)})]] \\
&= \mathbb{E}_\xi\left[\frac{1}{n}\sum_{j\neq i}[\ell(A(S), z_j) - \ell(A(S^{(i)}), z_j)] + \frac{1}{n}[\ell(A(S), z_i) - \ell(A(S^{(i)}), z_i')]\right].
\end{aligned}$$

Hence by the ε-uniform-stability assumption and the boundedness of ℓ,

$$\mathbb{E}_\xi[\hat{L}_S[A(S)] - \hat{L}_{S^{(i)}}[A(S^{(i)})]] \leqslant \varepsilon + \frac{2B}{n}.$$

Again, bounding $\mathbb{E}_\xi[\hat{L}_{S^{(i)}}[A(S^{(i)})] - \hat{L}_S[A(S)]] \leqslant \varepsilon + 2B/n$ proceeds identically. Putting these inequalities together, we see that

$$|f(S) - f(S^{(i)})| = |\mathbb{E}_\xi[L[A(S)]] - \mathbb{E}_\xi[L(A(S^{(i)}))] + \mathbb{E}_\xi[\hat{L}_S[A(S)] - \hat{L}_{S^{(i)}}[A(S^{(i)})]]| \leqslant 2\varepsilon + \frac{2B}{n}.$$

This shows that f satisfies the bounded differences inequality. Hence, by Proposition B.16, with probability at least $1 - \delta$,

$$f(S) \leqslant \mathbb{E}[f(S)] + \sqrt{8n\left(\varepsilon + \frac{B}{n}\right)^2 \log(1/\delta)} = \mathbb{E}[f(S)] + 2(\varepsilon n + B)\sqrt{\frac{2\log(1/\delta)}{n}}.$$

Finally, by Lemma 1.53, we have that $\mathbb{E}[f(S)] \leqslant \varepsilon$. The claim now follows. □

An immediate consequence of Lemma 1.54 is that we require that the uniform stability constant ε to scale as $\varepsilon = O(n^{-1})$ in order for the generalization bound from Lemma 1.54 to scale as the expected $O(n^{-1/2})$.

Remark 1.55. Note that Lemma 1.54 only prescribes a high probability bound over the randomness of the training data, and *not* over the algorithm's internal randomness. This is because our definition of uniform stability (Definition 1.52) only considers the average stability over the algorithm randomness (for a fixed (S, S', z)).

Exercise 1.56. Modify the definition of Definition 1.52 and the corresponding results in Lemma 1.53 and Lemma 1.54, so that the high probability bound holds jointly over drawing both the training data and the specific algorithm randomness. Note that there is quite a bit of flexibility in altering the definitions (as in there is no one right answer), so be sure to also discuss the pros/cons of your approach.

1.4.2 Stability of stochastic gradient descent

With the previous section in place, we now turn to the studying the stability of specific algorithms. We will first consider the classical stochastic gradient descent (SGD) algorithm. In order to simplify the analysis, we will study a stylized version with no batching and where the data is drawn uniformly at random (instead of random shuffling of the data). Specifically, suppose we have a loss function $\ell(w, z)$, where now we let $w \in \mathbb{R}^d$. The SGD algorithm uses the following recipe for $t = 0, 1, \ldots,$

1. Let $i_t \sim \mathrm{Unif}(\{1, \ldots, n\})$.
2. Set $w_{t+1} = w_t - \eta_t \nabla_w \ell(w_t, z_{i_t})$.

To study this update rule, let us define the update operator $G_{\ell,\eta}(w, z)$ as:

$$G_{\ell,\eta}(w, z) := w - \eta \nabla_w \ell(w, z),$$

so that the SGD update rule is $w_{t+1} = G_{\ell,\eta_t}(w_t, z_{i_t})$.

Proposition 1.57 (Update operator is non-expansive)**.** *Suppose for every $z \in \mathsf{Z}$, we have $w \mapsto \ell(w, z)$ is convex and β-smooth. Then for any $u, v \in \mathbb{R}^d$, as long as $\eta \leqslant 2/\beta$, we have that the update operator $G_{\ell,\eta}$ is non-expansive, i.e.,*

$$\forall z \in \mathsf{Z},\ \|G_{\ell,\eta}(u, z) - G_{\ell,\eta}(v, z)\| \leqslant \|u - v\|.$$

Proof. From Proposition C.5, we have that the gradients of ℓ are co-coercive:

$$\langle \nabla_w \ell(u,z) - \nabla_w \ell(v,z), u - v \rangle \geqslant \frac{1}{\beta} \|\nabla_w \ell(u,z) - \nabla_w \ell(v,z)\|^2.$$

Therefore,

$$\begin{aligned}
&\|G_{\ell,\eta}(u,z) - G_{\ell,\eta}(v,z)\|^2 \\
&= \|u - v - \eta(\nabla_w \ell(u,z) - \nabla_w \ell(v,z))\|^2 \\
&= \|u-v\|^2 - 2\eta\langle u - v, \nabla_w \ell(u,z) - \nabla_w \ell(v,z)\rangle + \eta^2 \|\nabla_w \ell(u,z) - \nabla_w \ell(v,z)\|^2 \\
&\leqslant \|u-v\|^2 - (2\eta/\beta - \eta^2)\|\nabla_w \ell(u,z) - \nabla_w \ell(v,z)\|^2 \\
&\leqslant \|u-v\|^2 && \text{since } \eta \leqslant 2\beta.
\end{aligned}$$

□

Lemma 1.58 (SGD is stable on convex functions)**.** *Suppose that for every $z \in \mathsf{Z}$, we have $w \mapsto \ell(w,z)$ is convex, L-Lipschitz, and β-smooth. Then, as long as $\eta_t \leqslant 2/\beta$ for all t, running the SGD algorithm for T iterations is ε-uniformly-stable for*

$$\varepsilon \leqslant \frac{2L^2}{n} \sum_{t=0}^{T-1} \eta_t.$$

Proof. Let us consider running SGD on two datasets $S = \{z_i\}_{i=1}^n$ and $S' = \{z_i'\}_{i=1}^n$, where S, S' differ in exactly one data point. Let $\kappa \in \{1, \dots, n\}$ denote the index which the datasets differ. We are going to use what is known as a *coupling argument*. We let the iterates $\{u_t\}$ denote SGD running on S, and $\{v_t\}$ denote SGD running on S'. The coupling argument works by using the *same* randomness ξ for both executions (in this case, the same indices $\{i_t\}$ are drawn). Thus, since SGD initializes its weighs independently of the data (but possibly using the internal randomness ξ), we start with $u_0 = v_0$. Let $\delta_t := \|u_t - v_t\|$. We will write a recursion for δ_t by decomposing as follows:

$$\delta_{t+1} = \delta_{t+1} \mathbf{1}\{i_t = \kappa\} + \delta_{t+1} \mathbf{1}\{i_t \neq \kappa\}.$$

Let us first consider $\delta_{t+1}\mathbf{1}\{i_t \neq \kappa\}$, for which we have:

$$\begin{aligned}
\delta_{t+1}\mathbf{1}\{i_t \neq \kappa\} &= \|G_{\ell,\eta_t}(u_t, z_{i_t}) - G_{\ell,\eta_t}(v_t, z'_{i_t})\| \mathbf{1}\{i_t \neq \kappa\} \\
&= \|G_{\ell,\eta_t}(u_t, z_{i_t}) - G_{\ell,\eta_t}(v_t, z_{i_t})\| \mathbf{1}\{i_t \neq \kappa\} && \text{since } z_{i_t} = z'_{i_t} \text{ when } i_t \neq \kappa \\
&\leqslant \|u_t - v_t\| \mathbf{1}\{i_t \neq \kappa\} && \text{using Proposition 1.57} \\
&= \delta_t \mathbf{1}\{i_t \neq \kappa\}.
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\delta_{t+1}\mathbf{1}\{i_t = \kappa\} &= \|u_t - v_t - \eta_t(\nabla_w \ell(u_t, z_{i_t}) - \nabla_w \ell(v_t, z'_{i_t}))\| \mathbf{1}\{i_t = \kappa\} \\
&\leqslant (\|u_t - v_t\| + 2\eta_t L)\mathbf{1}\{i_t = \kappa\} && \text{since } \ell \text{ is } L\text{-Lipschitz} \\
&= (\delta_t + 2\eta_t L)\mathbf{1}\{i_t = \kappa\}.
\end{aligned}$$

Combining these two inequalities,

$$\delta_{t+1} \leqslant \delta_t + 2\eta_t L \mathbf{1}\{i_t = \kappa\} \implies \delta_T \leqslant 2L \sum_{t=0}^{T-1} \eta_t \mathbf{1}\{i_t = \kappa\}.$$

Since this inequality holds for every ξ, by linearity of expectation,

$$\mathbb{E}_\xi[\delta_T] \leqslant \frac{2L}{n} \sum_{t=0}^{T-1} \eta_t.$$

The claim now follows since ℓ is L-Lipschitz:

$$\mathbb{E}_\xi|\ell(u_T,z)-\ell(v_T,z)| \leqslant L\mathbb{E}_\xi|u_T-v_T| = L\mathbb{E}_\xi[\delta_T] \leqslant \frac{2L^2}{n}\sum_{t=0}^{T-1}\eta_t.$$

□

Remark 1.59. Note that the proof of Lemma 1.58 illustrates why the definition of uniform stability in Definition 1.52 requires that stability only hold in expectation over the internal randomness of the algorithm. If, for instance, the definition required that the algorithm was stable for almost every realization of the randomness, then SGD would not be stable under this definition, since the realization ξ which repeatedly picks the index corresponding to the differing point κ is not stable (despite being very unlikely) .

Remark 1.60. Note that Lemma 1.58 does *not* require at most one pass through the data; repeated passes over the same datapoint are allowed. This is not a typical property of SGD generalization analysis.

Sub-optimal rates for typical SGD step sizes. Note that one drawback of Lemma 1.58 is that for the typical $\eta_t \asymp t^{-1/2}$ step sizes usually prescribed for running SGD on convex functions, the ε-uniform-stability constant scales as $\varepsilon = O(\sqrt{T}/n)$. So for any non-trivial amount of steps T, we have that ε is slower than the n^{-1} scaling required so that Lemma 1.54 yields $O(n^{-1/2})$ generalization bounds. On the other hand, if we use a more aggressive step size $\eta_t \asymp t^{-1}$, then we have that $\varepsilon = O(\log T/n)$ which improves the overall generalization rate, but is generally too aggressive. If we impose extra conditions on the function ℓ, specifically that it is strongly convex (Definition C.6), then it turns out we can get a stability bound which is independent of the number of iterations run.

Exercise 1.61. Suppose that for every $z \in \mathsf{Z}$, we have that $w \mapsto \ell(w,z)$ is μ-strongly-convex (cf. Definition C.6), L-Lipschitz, and β-smooth.

(a) Show that if $\eta \leqslant 2/(\mu+\beta)$, the update operator $G_{\ell,\eta}$ becomes *contractive*, i.e., for any $u, v \in \mathbb{R}^d$:

$$\forall z \in \mathsf{Z},\ \|G_{\ell,\eta}(u,z)-G_{\ell,\eta}(u,z)\| \leqslant \left(1-\frac{\eta\mu\beta}{\mu+\beta}\right)\|u-v\|.$$

Hint: you may assume the following fact about μ-strongly-convex and β-smooth functions without proof. For all $u, v \in \mathbb{R}^d$,

$$\langle \nabla f(u)-\nabla f(v), u-v\rangle \geqslant \frac{\mu\beta}{\mu+\beta}\|u-v\|^2 + \frac{1}{\mu+\beta}\|\nabla f(u)-\nabla f(v)\|^2.$$

(b) Modify the proof of Lemma 1.58 to show the following recursion, valid whenever $\eta_t \leqslant 2/(\mu+\beta)$:

$$\delta_{t+1} \leqslant (1-\eta_t m)\delta_t + 2\eta_t L\mathbf{1}\{i_t=\kappa\}, \quad m := \frac{\mu\beta}{\mu+\beta}.$$

(c) Now suppose we use a *constant* step size $\eta_t = \eta \leqslant 2/(\mu+\beta)$. Conclude that:

$$\mathbb{E}[\delta_T] \leqslant \frac{2L}{mn},$$

and hence SGD is ε-uniformly-stable for $\varepsilon \leqslant \frac{2L^2}{mn}$.

1.4.3 Stability of Gibbs ERM

We now study the stability of an idealized algorithm which returns a sample from a particular weighted Gibbs distribution. In particular, let $p_\beta(w; S)$ be the density of a distribution on $\mathbb{R}^d$ defined as:

$$p_\beta(w; S) = \exp(-\beta \hat{L}_S[w])/Z_\beta[w; S], \quad Z_\beta[w; S] = \int \exp(-\beta \hat{L}_S[w]) \, \mathrm{d}w. \tag{1.20}$$

Proposition 1.62 (Gibbs ERM is uniformly-stable)**.** *Let $A(S, \xi)$ denote the algorithm which returns a sample from the Gibbs distribution defined by Equation (1.20). Suppose that $|\ell| \leqslant M$. We have that A is ε-uniformly-stable with $\varepsilon \leqslant (\exp(4\beta M/n) - 1)M$.*

Proof. Let S, S' differ in only the i-th index. Note that:

$$\begin{aligned}
\mathbb{E}_\xi[\ell(A(S), z) - \ell(A(S'), z)] &= \int \ell(w, z) p_\beta(w; S) \, \mathrm{d}w - \int \ell(w, z) p_\beta(w; S') \, \mathrm{d}w \\
&= \int \ell(w, z) \left(\frac{p_\beta(w; S)}{p_\beta(w; S')} - 1 \right) p_\beta(w; S') \, \mathrm{d}w \\
&\leqslant M \left| \frac{p_\beta(w; S)}{p_\beta(w; S')} - 1 \right|.
\end{aligned}$$

To bound the likelihood ratio, we first observe that,

$$\frac{\exp(-\beta \hat{L}_S[w])}{\exp(-\beta \hat{L}_{S'}[w])} = \exp\left(-\frac{\beta}{n} (\ell(w, z_i) - \ell(w, z_i')) \right) \leqslant \exp(2\beta M/n).$$

On the other hand,

$$\frac{Z_\beta[w; S]}{Z_\beta[w; S']} = \frac{\int \exp(-\beta \hat{L}_S[w]) \, \mathrm{d}w}{\int \exp(-\beta \hat{L}_{S'}[w]) \, \mathrm{d}w} = \frac{\int \exp\left(-\frac{\beta}{n} (\ell(w, z_i) - \ell(w, z_i')) \right) \exp(-\beta \hat{L}_{S'}[w]) \, \mathrm{d}w}{\int \exp(-\beta \hat{L}_{S'}[w]) \, \mathrm{d}w} \leqslant \exp(2\beta M/n).$$

Similar arguments show that

$$\min \left\{ \frac{\exp(-\beta \hat{L}_S[w])}{\exp(-\beta \hat{L}_{S'}[w])}, \frac{Z_\beta[w; S]}{Z_\beta[w; S']} \right\} \geqslant \exp(-2\beta M/n).$$

Hence,

$$\begin{aligned}
\left| \frac{p_\beta(w; S)}{p_\beta(w; S')} - 1 \right| &= \left| \frac{\exp(-\beta \hat{L}_S[w])}{\exp(-\beta \hat{L}_{S'}[w])} \cdot \frac{Z_\beta[w; S']}{Z_\beta[w; S]} - 1 \right| \\
&\leqslant \max \left\{ \exp\left(\frac{4\beta M}{n} \right) - 1, 1 - \exp\left(-\frac{4\beta M}{n} \right) \right\} = \exp(4\beta M/n) - 1.
\end{aligned}$$

The last equality holds since $\cosh(x) \geqslant 1$ for all x. □

Remark 1.63. **TODO:** mention connections to Gibbs ERM and differential privacy

TODO: discuss SGLD as a way to sample from the Gibbs ERM distribution.

1.5 PAC-Bayes inequalities

We now study an alternative approach to controlling generalization error. When we studied the generalization of ERM, we essentially proved uniform convergence of training and population risk over the entire hypothesis space $\mathscr{F}$. One downside of this approach is that it uniformly weights every hypothesis in the function class. An alternative approach is to construct a prior distribution over hypothesis, and take a "weighted" union bound over the entire function class which takes into account the prior. This is what the PAC-Bayes approach does. After we derive the main PAC-Bayes deviation inequality, we will instantiate it for a few special cases, including what are referred to as compression-based bound in the literature.

Perhaps the main conceptual difference of PAC-Bayes analysis compared with ERM analysis is that the learning algorithm is now a *distribution* over hypothesis $\mathscr{F}$, which we will refer to as the posterior (hence where the "Bayes" part of the name stems from). In PAC-Bayes, we start with a prior π over $\mathscr{F}$, data $\{z_i\}_{i=1}^n$ is observed, and then a posterior distribution ρ_n over $\mathscr{F}$ is formed by incorporating the observed data. Unlike ERM analysis, the main quantity which governs the generalization error is the KL-divergence between the posterior ρ_n and the prior π.

While there are many variants of PAC-Bayes inequalities, we will study one due to Cantoni, described by Alquier [2021].

Theorem 1.64 (PAC-Bayes deviation inequality)**.** *Suppose that $|\ell| \leqslant B$. For any $\lambda > 0$ and $\delta \in (0,1)$, we have with probability at least $1-\delta$,*

$$\forall \rho \in \mathcal{P}(\mathscr{F}),\ \mathbb{E}_{f\sim\rho}[L[f]] \leqslant \mathbb{E}_{f\sim\rho}[L_n[f]] + \frac{\lambda B^2}{2n} + \frac{\mathrm{KL}(\rho \,\|\, \pi) + \log(1/\delta)}{\lambda}.$$

Here, $\mathcal{P}(\mathscr{F})$ refers to the set of probability measures over the hypothesis class $\mathscr{F}$.

Proof. The main tool used to derive the PAC-Bayes inequality is the Donsker-Varadhan representation of KL-divergence (Lemma A.3). We first set things up in order to apply Donsker-Varadhan.

We start with a fixed $f \in \mathscr{F}$. By Hoeffding's inequality (Proposition B.12),

$$\mathbb{E}_S \exp(\lambda(L_n[f] - L[f])) \leqslant \exp(\lambda^2 B^2/(2n)).$$

(Here $\mathbb{E}_S$ denotes expectation w.r.t. the data $\{z_i\}$.) Note that since this inequality holds for every fixed $f \in \mathscr{F}$, it also holds if we integrate it w.r.t. the prior π (this works because the prior *does not depend on* the data). Hence,

$$\mathbb{E}_{f\sim\pi}\mathbb{E}_S \exp(\lambda(L_n[f] - L[f])) \leqslant \exp(\lambda^2 B^2/(2n)).$$

Now by Fubini's theorem, we exchange the order of expectations on the LHS,

$$\mathbb{E}_S\mathbb{E}_{f\sim\pi} \exp(\lambda(L_n[f] - L[f])) \leqslant \exp(\lambda^2 B^2/(2n)).$$

The next step is to apply the Donsker-Varadhan lemma (Lemma A.3). Specifically, we will use the following identity which holds for any fixed g,

$$\log \mathbb{E}_\pi[\exp(g)] = \sup_{\rho\in\mathcal{P}(\mathscr{F})} \{\mathbb{E}_\rho[g] - \mathrm{KL}(\rho \,\|\, \pi)\}.$$

Applying this identity to $g = \lambda(L_n[f] - L[f])$ and taking an expectation w.r.t. S on both sides,

$$\mathbb{E}_S\mathbb{E}_\pi \exp\left(\lambda(L_n[f] - L[f])\right) = \mathbb{E}_S \exp\left(\sup_{\rho\in\mathcal{P}(\mathscr{F})} \{\lambda\mathbb{E}_\rho[L_n[f] - L[f]] - \mathrm{KL}(\rho \,\|\, \pi)\}\right) \leqslant \exp(\lambda^2 B^2/(2n)).$$

Hence for any $t > 0$, by the Laplace transform method (Proposition B.2),

$$\mathbb{P}_S\left\{\sup_{\rho\in\mathcal{P}(\mathscr{F})} \{\lambda\mathbb{E}_\rho[L_n[f] - L[f]] - \mathrm{KL}(\rho \,\|\, \pi)\} - \frac{\lambda^2 B^2}{2n} \geqslant t\right\}$$

$$\leqslant e^{-t}\mathbb{E}_S \exp\left(\sup_{\rho\in\mathcal{P}(\mathscr{F})}\{\lambda\mathbb{E}_\rho[L_n[f]-L[f]]-\mathrm{KL}(\rho\,\|\,\pi)\}-\frac{\lambda^2B^2}{2n}\right)\leqslant e^{-t}.$$

The claim now follows by setting $t=\log(1/\delta)$. □

Let us first consider to what extent PAC-Bayes generalizes the uniform convergence analysis of ERM. First, we see that the PAC-Bayes inequality implies the uniform convergence result from Proposition 1.14.

Exercise 1.65. Suppose that $\mathscr{F}$ is finite. Show that the PAC-Bayes deviation bound (Theorem 1.64) implies that with probability at least $1-\delta$,

$$\forall f\in\mathscr{F},\ L[f]\leqslant L_n[f]+B\sqrt{\frac{2\log(|\mathscr{F}|/\delta)}{n}}.$$

Let us now explore an idea from the deep learning literature, known as compression-based bounds. There is a vast literature here, so we will only cover the very basics. The intuition is that many deep learning models, while vastly overparameterized, can actually be compressed (in a lossy manner) but still retain much of its performance. The PAC-Bayes framework suggests a way to utilize this insight in order to derive sharper generalization bounds, by building a prior distribution which upweights compressible models. Let us describe a simplistic model of this. Let M denote a max bound on the hypothesis we will consider, denoting the number of bits used to represent the model (hence there are $|\mathscr{F}|=\sum_{i=1}^M 2^i=2(2^M-1)$ possible models in total). For each model $f\in\mathscr{F}$, we let $|f|\in\{1,\ldots,M\}$ denote the number of bits used to represent f. With this notation, let us consider a prior π on $\mathscr{F}$ such that:

$$\pi(f)=\frac{2^{-|f|}}{M}.$$

Now if we let ρ_f denote a point mass on model f, a quick computation yields:

$$\mathrm{KL}(\rho_f\,\|\,\pi)=|f|\log 2+\log M.$$

This prior can now be used to construct the following compression-bound.

Exercise 1.66. Suppose that $\mathscr{F}$ is as described above. Show that the PAC-Bayes deviation bound (Theorem 1.64) implies that with probability at least $1-\delta$,

$$\forall f\in\mathscr{F},\ L[f]\leqslant L_n[f]+B\sqrt{\frac{2}{n}(|f|\log 2+2\log(M/\delta))}.$$

(Obtaining these precise constants is not important. It is valid to obtain a bound with slightly different constants, just make sure to show your work.)

Applying the standard uniform convergence bound for finite hypothesis classes (Proposition 1.14) to this setting yields a bound of:

$$\max_{f\in\mathscr{F}}\mathrm{gen}[f]\leqslant B\sqrt{\frac{2\log(|\mathscr{F}|/\delta)}{n}}\leqslant cB\sqrt{\frac{M\log(1/\delta)}{n}},$$

for an absolute constant c. Compared this with Exercise 1.66 where compressible models (e.g. $|f|\ll M$) enjoy a much tighter generalization bound. Of course, in the worst case when $|f|\asymp M$, the Exercise 1.66 yields orderwise the same bound.

2 Unsupervised learning: generative modeling

We now turn our attention towards unsupervised learning. Unsupervised learning is a very broad topic. To focus our discussion, we will consider a subfield of unsupervised learning, specifically generative modeling. Let us first setup the learning problem. As before, we have covariates $x \in \mathsf{X}$ and an underlying distribution $p(x)$ over these covariates. (Again, we will typically assume Euclidean structure on X, namely $\mathsf{X} \subseteq \mathbb{R}^d$, although many of the ideas we will discuss transfer over the case when X is countable.) However, we no longer have any labels $y \in \mathsf{Y}$, making this an unsupervised learning setup. The goal then is no longer prediction, but rather sampling. Indeed, given a set of samples $x_1, \dots, x_n$, the goal is to learn a sampler $\hat{p}_n$ such that samples drawn from $\hat{p}_n$ approximate samples drawn from $p(x)$ as close as possible.

There are many ways to measure this closeness, including: asking that the distributions match in total-variation norm (i.e., $\|p - \hat{p}_n\|_{\mathrm{tv}} \leqslant \varepsilon$) or in KL-divergence (i.e., $\mathrm{KL}(p \,\|\, \hat{p}_n) \leqslant \varepsilon$). We will mostly focus on the latter, noting that by Pinsker's inequality the latter implies the former:

$$\|p - \hat{p}_n\|_{\mathrm{tv}} \leqslant \sqrt{\frac{1}{2}\mathrm{KL}(p \,\|\, \hat{p}_n)}.$$

2.1 Energy based models

Our first attempt at solving this problem will be using classic energy based models. This is a classic idea dating back to Yann LeCun in the early 2000s **TODO:** cite. Suppose that the distribution $p(x)$ over X has a density, and with slight abuse of notation we let $p(x)$ denote this density. Energy based models (EBMs) simply model the *un-normalized log density*, i.e., for a function class $\mathscr{F}$ mapping $\mathsf{X} \mapsto \mathbb{R}$,

$$p(x) \propto \exp(f(x)), \quad f \in \mathscr{F}. \tag{2.1}$$

The important property of this model is that there is no constraint placed on f that enforces normalization, i.e., generically we have that

$$Z[f] := \int \exp(f(x))\,\mathrm{d}x \neq 1.$$

In reference to the statistical physics motivations behind EBMs, the function $Z[f]$ is called the *partition function*. Of course for a given $f \in \mathscr{F}$, the resulting normalized density is hence:

$$p_f(x) = \exp(f(x))/Z[f].$$

Since computing $Z[f]$ is typically not tractable, as it involves high dimensional integration, the exact form p_f is to be thought of as a purely analytical tool. As stated before, the loss function we want to minimize is now the KL-divergence. This motivates the following maximum likelihood (MLE) procedure at the population level:

$$f_\star = \arg\min_{f \in \mathscr{F}} \mathrm{KL}(p \,\|\, p_f) = \arg\min_{f \in \mathscr{F}} \{-H[p] - \mathbb{E}_p[\log p_f]\} = \arg\min_{f \in \mathscr{F}} \{-\mathbb{E}_p[\log p_f]\} =: \arg\min_{f \in \mathscr{F}} L[f]. \tag{2.2}$$

Here, $H[p] = -\mathbb{E}_p[\log p]$ is the (differential) entropy of p. Note the loss $L[f]$ is referred to as the *cross-entropy loss*. Let us now suppose that $\mathscr{F} = \{f_\theta \mid \theta \in \Theta\}$ where Θ is a subset of Euclidean space, and compute what the gradient of the cross-entropy loss is. Slightly overloading notation with $L[\theta] = L[f_\theta]$, $p_\theta = p_{f_\theta}$, and $Z[\theta] = Z[f_\theta]$, a quick computation yields:

$$\begin{aligned}
\nabla_\theta L[\theta] &= -\nabla_\theta \int \log p_\theta(x) p(x)\,\mathrm{d}x \\
&\overset{(a)}{=} -\int \nabla_\theta \log p_\theta(x) p(x)\,\mathrm{d}x \\
&= -\int \nabla_\theta [f_\theta(x) - \log Z[\theta]] p(x)\,\mathrm{d}x
\end{aligned}$$

$$= -\mathbb{E}_p[\nabla_\theta f_\theta] + \nabla_\theta \log Z[\theta].$$

In (a), we exchanged the order of differentiation and integration.[6] Next we compute:

$$\nabla_\theta \log Z[\theta] = \frac{1}{Z[\theta]} \nabla_\theta \int \exp(f_\theta(x))\,\mathrm{d}x \overset{(a)}{=} \frac{1}{Z[\theta]} \int \nabla_\theta \exp(f_\theta(x))\,\mathrm{d}x$$
$$= \frac{1}{Z[\theta]} \int \nabla_\theta f_\theta(x) \exp(f_\theta(x))\,\mathrm{d}x = \mathbb{E}_{p_\theta}[\nabla_\theta f_\theta].$$

Again in (a) we exchanged the order of differentiation and integration. Above, the blue color p_θ is used to emphasize that the sampling is done w.r.t. the model p_θ and not the true data generating distribution p. Combining these calculations,

$$\nabla_\theta L[\theta] = -(\mathbb{E}_p[\nabla_\theta f_\theta] - \mathbb{E}_{p_\theta}[\nabla_\theta f_\theta]). \tag{2.3}$$

Thus, we see that to form a stochastic estimate of this gradient, we need to *sample* from the model p_θ. We now turn to how to generate these samples.

2.1.1 Markov Chain Monte Carlo sampling

Here, suppose that $\mathsf{X} \subseteq \mathbb{R}^d$. The standard algorithm for sampling from p_θ is Langevin sampling, which sets up a Markov chain with a stationary distribution equal to p_θ. A classic fact from statistical physics (by way of the Fokker-Planck equation) states that the following Itô stochastic differential equation (SDE), know as *Langevin dynamics*, has its stationary measure equal to p_θ:

$$\mathrm{d}X_t = \nabla_x \log p_\theta(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t. \tag{2.4}$$

Here, $(B_t)_{t\geqslant 0}$ is standard Brownian motion in $\mathbb{R}^d$. For those unfamiliar with SDEs, we can think of the process $(X_t)_{t\geqslant 0}$ described in Equation (2.4) as the continuous time limit as $\eta \to 0$ of the following discrete-time stochastic process parameterized by $\eta > 0$, which does not require any fancy machinery to describe:[7]

$$X_{t+1} = X_t + \eta \nabla_x \log p_\theta(X_t) + \sqrt{2\eta} W_t. \tag{2.5}$$

Here, $W_t \sim N(0, I)$ is drawn independently for $t \in \mathbb{N}$. Note the appearance of $\log p_\theta$ above is critical, since it allows the partition function to disappear in the gradient:

$$\nabla_x \log p_\theta(x) = \nabla_x [f_\theta(x) - \log Z[\theta]] = \nabla_x f_\theta(x).$$

The object $\nabla_x \log p_\theta(x)$ is an important enough quantity that we give it a particular name: the *score function*. Let us denote X_t from the discrete-time process (2.5) as:

$$X_t = \mathsf{MCMC}_\eta(f_\theta, X_0, t),$$

with the dependence on the random $\{W_t\}$'s made implicit.

Stochastic gradient estimate for $L[\theta]$. Recall that the population level gradient for the cross-entropy loss $L[\theta] = -\mathbb{E}_p[\log p_\theta]$ is given by $\nabla_\theta L[\theta] = -(\mathbb{E}_p[\nabla_\theta f_\theta] - \mathbb{E}_{p_\theta}[\nabla_\theta f_\theta])$. We now have the necessary notation to write down a stochastic estimator for this gradient. This estimator has two hyperparameters, $\eta > 0$ and $\kappa \in \mathbb{N}_+$. Consider the following estimator

$$\hat{g}_n := \nabla_\theta f_\theta(x_\iota) - \nabla_\theta f_\theta(\mathsf{MCMC}_\eta(f_\theta, 0, \kappa)), \quad \iota \sim \mathrm{Unif}(\{1, \ldots, n\}).$$

This estimator approximately satisfies $\mathbb{E}[\hat{g}_n] \approx \nabla_\theta L[\theta]$, although quantifying the bias of $\hat{g}_n$ is non-trivial and out of scope for this course (we will briefly discuss the sources of error in a moment). Furthermore, it is not necessary to start $X_0 = 0$ in the MCMC; more sophisticated variants use warm-starting techniques (e.g. using the last sample generated) to improve MCMC convergence. See **TODO:** cite for more advanced techniques.

[6] Technically, we should check some integrability conditions to ensure this is valid (using Lebesgue's dominated convergence theorem), but we will ignore this technicality and assume this step holds.

[7] This is known as the Euler–Maruyama discretization of the Langevin SDE.

Divergence, Laplacian, and integration by parts. For what follows, we will make use of the following notation and tools from multivariable calculus. For a vector field $v : \mathbb{R}^d \mapsto \mathbb{R}^d$, the divergence $\nabla \cdot v$ is:

$$(\nabla \cdot v)(x) = \sum_{i=1}^{d} \frac{\partial v_i(x)}{\partial x_i}.$$

Next, for a function $f : \mathbb{R}^d \mapsto \mathbb{R}$, the Laplacian Δf is:

$$(\Delta f)(x) = \sum_{i=1}^{d} \frac{\partial^2 f(x)}{\partial x_i^2}.$$

(Check the following identity that $\nabla \cdot \nabla f = \Delta f$.)

A key technical tool we will use is the classic integration by parts identity, which states that for a smooth function $f : \mathbb{R}^d \mapsto \mathbb{R}$ and smooth vector field $v : \mathbb{R}^d \mapsto \mathbb{R}^d$ whose behavior at infinity vanishes,[8]

$$\int \langle v(x), \nabla f(x) \rangle \, \mathrm{d}x = -\int \nabla \cdot v(x) f(x) \, \mathrm{d}x. \tag{2.6}$$

Stationarity of Langevin dynamics (2.4). To check that p_θ is the stationary measure of the Langevin dynamics (2.4), we will appeal to the Fokker-Planck (FP) equation, which states the following. Suppose we have a SDE of the form:

$$\mathrm{d}X_t = v(X_t) \, \mathrm{d}t + \sqrt{2}\mathrm{d}B_t.$$

Then, the probability density function ρ_t of X_t satisfies the following partial differential equation (PDE):[9]

$$\frac{\partial \rho_t}{\partial t} = -\nabla \cdot (\rho_t v) + \Delta \rho_t. \tag{2.7}$$

Now, plugging (2.4) into the FP equation (2.7), we have:

$$\begin{aligned}
\frac{\partial p_\theta}{\partial t} &= -\nabla \cdot (p_\theta \nabla \log p_\theta) + \Delta p_\theta \\
&= -\nabla \cdot (p_\theta \nabla \log p_\theta) + \nabla \cdot \nabla p_\theta && \text{since } \Delta p_\theta = \nabla \cdot \nabla p_\theta \\
&= -\nabla \cdot (p_\theta \nabla \log p_\theta) + \nabla \cdot (p_\theta \nabla \log p_\theta) && \text{since } p_\theta \nabla \log p_\theta = \nabla p_\theta \\
&= 0.
\end{aligned}$$

Therefore, tells us that the density of p_θ remains invariant over time under the SDE (2.4). In other words, if we start $X_0 \sim p_\theta$, then $\mathrm{Law}(X_t) = p_\theta$ for all $t \geqslant 0$.

Sources of MCMC error. We briefly describe the sources of error which arise in using MCMC to generate samples from p_θ. We first consider the continuous-limit SDE (2.4). Via the Fokker-Planck equation, we previous showed that p_θ is the stationary measure of the Langevin dynamics (2.4). As a consequence of this, we have that:

$$\lim_{t \to \infty} \mathrm{KL}(\mathrm{Law}(X_t) \parallel p_\theta) = 0. \tag{2.8}$$

However, for any *finite* $t > 0$, $\mathrm{KL}(\mathrm{Law}(X_t) \parallel p_\theta)$ is not necessarily zero. Furthermore, the convergence speed of (2.8), without any further assumptions on p_θ, is known to be prohibitively slow in the worst case. This happens in particular when the distribution p_θ is *multi-modal*. As an extreme case, consider a distribution with two modes that have equal probability, which are well-separated. Since both modes are equally likely,

[8] We will not be overly concerned with the technical conditions here.

[9] The fact that p_t has a well-defined density respect to the Lebesgue measure on $\mathbb{R}^d$ is a consequence of the fact that the Brownian motion driving the process $(X_t)_{t \geqslant 0}$ is full dimensional.

the Langevin dynamics should spend equal time traversing between the two modes, in order to accurately reflect the underlying distribution. However, since Langevin dynamics noisily follow the gradient of the log density (cf. Equation (2.4)), traversing across models can take a long time (exponential in the ambient dimension d). Thus, in the worst case, convergence of Langevin dynamics to the stationary distribution p_θ requires running the SDE for very large t. This presents one source of error.

The next source of error involves the relationship between the continuous-limit SDE (2.4) and the discrete-time stochastic process (2.5). To avoid confusion, let $\{Z_t\}_{t\in\mathbb{N}}$ denote the discrete-time process (2.5). Conceptually, we can think of there being a natural correspondence between $Z_k \leftrightarrow X_{k\eta}$. However, the discrepancy between Z_k and $X_{k\eta}$ grows as η increases; for any finite $\eta > 0$ there is always discretization error. This presents another source of error.

2.1.2 Convergence of Langevin dynamics

Note: the convergence of Langevin dynamics is an advanced topic and we will probably not have time to cover it in class, but it is included for readers who are interested in diving further into this topic.

In Section 2.1.1, we noted that in the worse case, without further assumptions on the distribution to be sampled from, the Langevin dynamics can be prohibitively slow to converge. On the other hand, in this section, we will consider a sufficient condition which rules out this pathology. To motivate the analysis, consider the following Langevin dynamics (here p is an arbitrary density),

$$\mathrm{d}X_t = \nabla \log p(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t.$$

Let $p_t = \mathrm{Law}(X_t)$ denote both the measure and density of the Langevin dynamics at time t. Since p is the stationary measure of the Langevin dynamics, we know that $\lim_{t\to\infty} \mathrm{KL}(p_t \parallel p) = 0$. Our goal for this section will be to quantify the rate at which this convergence happens. To study this rate, we treat $\Phi(t) := \mathrm{KL}(p_t \parallel p)$ as a *potential function* and take its time-derivative:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\Phi(t) &= \partial_t \int p_t \log \frac{p_t}{p}\,\mathrm{d}x \\
&= \int \partial_t \left(p_t \log \frac{p_t}{p}\right)\mathrm{d}x \\
&= \int (\partial_t p_t)\log\frac{p_t}{p}\,\mathrm{d}x + \int p_t \cdot \partial_t \log p_t\,\mathrm{d}x \\
&\overset{(a)}{=} \int (\partial_t p_t)\log\frac{p_t}{p}\,\mathrm{d}x \\
&\overset{(b)}{=} \int \nabla\cdot\left(p_t \nabla \log\frac{p_t}{p}\right)\log\frac{p_t}{p}\,\mathrm{d}x && \text{Fokker-Planck equation} \\
&= -\int \left\|\nabla \log\frac{p_t}{p}\right\|^2 p_t\,\mathrm{d}x && \text{integration by parts} \\
&= -\mathbb{E}_{p_t}\left\|\nabla\log\frac{p_t}{p}\right\|^2. && (2.9)
\end{aligned}$$

To see (a) above, note that:

$$\int p_t \cdot \partial_t \log p_t\,\mathrm{d}x = \int \partial_t p_t\,\mathrm{d}x = \partial_t \int p_t\,\mathrm{d}x = 0.$$

To see how (b) arises from the Fokker-Placnk equation, from (2.7),

$$\begin{aligned}
\partial_t p_t &= -\nabla\cdot(p_t\nabla\log p) + \Delta p_t \\
&= -\nabla\cdot[p_t\nabla\log p - \nabla p_t] && \text{since } \Delta p_t = \nabla\cdot\nabla p_t
\end{aligned}$$

$$= -\nabla \cdot (p_t \nabla \log p - p_t \nabla \log p_t)$$
$$= \nabla \cdot \left(p_t \nabla \log \frac{p_t}{p} \right).$$

Immediately we can draw the following conclusions from Equation (2.9):

(a) If $p_t = p$, then we see $\frac{\mathrm{d}}{\mathrm{d}t}\Phi(t) = 0$, confirming the stationarity of p for the Langevin dynamics.

(b) Regardless of what p_t is, $\frac{\mathrm{d}}{\mathrm{d}t}\Phi(t) \leqslant 0$. Consequently, $\Phi(t) \leqslant \Phi(0)$ for all $t \geqslant 0$. That is, Langevin dynamics never increases the KL-divergence of p_t from p. Note that we did not need to enforce any extra assumptions on p for this to hold.

Equation (2.9) gives us a hint as to what condition to enforce on p_t and p to ensure convergence. In particular, suppose we are assured that for some $\alpha > 0$ and all $t \geqslant 0$,

$$\mathbb{E}_{p_t} \left\| \nabla \log \frac{p_t}{p} \right\|^2 \geqslant \alpha \cdot \mathrm{KL}(p_t \parallel p). \tag{2.10}$$

Then, combining inequality (2.10) with (2.9),

$$\forall t \geqslant 0, \ \frac{\mathrm{d}}{\mathrm{d}t}\Phi(t) \leqslant -\alpha \cdot \Phi(t),$$

and therefore by the comparison lemma for ODEs,

$$\forall t \geqslant 0, \ \Phi(t) \leqslant e^{-\alpha t}\Phi(0).$$

Now the question remains, what assumption on p_t and p ensures the condition (2.10) holds? It turns out that this condition is ensured by a standard functional inequality in probability theory known as the *log-Sobolev* inequality. Log-Sobolev inequalities have a rich history in probability theory which we will not have time to discuss. We will simply state its definition and see how it implies (2.10). The first definition we need is that of *concentration entropy* (not to be confused with the information theoretic notion of entropy). For a measure μ and a non-negative random variable X, the concentration entropy $\mathrm{Ent}_\mu(X)$ (or entropy for short) of X is defined as:

$$\mathrm{Ent}_\mu(X) := \mathbb{E}_\mu[X \log X] - \mathbb{E}_\mu[X] \cdot \log \mathbb{E}_\mu[X].$$

A measure μ satisfies the *log-Sobolev inequality* with constant $C > 0$ if for all smooth functions $f : \mathsf{X} \mapsto \mathbb{R}$,

$$\mathrm{Ent}_\mu(f^2) \leqslant 2C \cdot \mathbb{E}_\mu \|\nabla f\|^2. \tag{2.11}$$

While this definition is seemingly arbitrary, it turns out that it directly implies (2.10). To see this, first let us compute the entropy of the function $f = \sqrt{p_t/p}$ under the measure p:

$$\mathrm{Ent}_p \left(\frac{p_t}{p} \right) = \mathbb{E}_p \left[\frac{p_t}{p} \log \frac{p_t}{p} \right] - \mathbb{E}_p \left[\frac{p_t}{p} \right] \log \mathbb{E}_p \left[\frac{p_t}{p} \right] = \mathbb{E}_p \left[\frac{p_t}{p} \log \frac{p_t}{p} \right] = \mathrm{KL}(p_t \parallel p).$$

Therefore, if we suppose that the stationary measure p satisfies the log-Sobolev inequality (LSI) with constant C, then applying the LSI inequality to $f = \sqrt{p_t/p}$,

$$\mathrm{Ent}_p \left(\frac{p_t}{p} \right) = \mathrm{KL}(p_t \parallel p) \leqslant 2C \cdot \mathbb{E}_p \left\| \nabla \sqrt{\frac{p_t}{p}} \right\|^2 = \frac{C}{2} \cdot \mathbb{E}_{p_t} \left\| \nabla \log \frac{p_t}{p} \right\|^2.$$

Hence, this implies (2.10) with $\alpha = 2/C$. That is, we have shown that if p satisfies the LSI inequality with constant C, then for all $t > 0$,

$$\mathrm{KL}(p_t \parallel p) \leqslant e^{-2t/C}\mathrm{KL}(p_0 \parallel p).$$

That is, if we want to ensure that $\mathrm{KL}(p_t \parallel p) \leqslant \varepsilon$, it suffices to ensure that $t \geqslant \frac{C}{2} \log \left(\frac{\mathrm{KL}(p_0 \parallel p)}{\varepsilon} \right)$.

Which distributions satisfy log-Sobolev inequalities? In general, verifying the functional inequality (2.11) for a given distribution p is not a simple task. There is, however, a very clean sufficient condition, known as the Bakry-Émery criterion. Suppose that $p = \exp(-U)$, and suppose that U is α strongly-convex. Then p satisfies the LSI inequality with constant $C = 1/\alpha$. The proof of this result is non-trivial and out of the scope of this course: see e.g., **TODO:** cite Bakry's book for a proof. Distributions of this form are known as log-concave distributions, and include e.g., Gaussian distributions.

2.2 Score matching

The MCMC sampling dynamics (2.5) motivate an alternative loss which does not rely on the the cross-entropy loss (2.2). Indeed, the idea is to *directly* learn the score function $\nabla_x \log p(x)$. It turns out that we can setup a very simple loss function, called the score matching loss, which bypasses the requirement to sample from the current model in order to implement the gradient of the cross-entropy loss (cf. Equation (2.3)).

Proposition 2.1 (Score matching loss)**.** *Let $p(x)$ denote a distribution over $\mathsf{X} \subset \mathbb{R}^d$, and suppose its density is sufficiently smooth. For any smooth vector field $f : \mathbb{R}^d \mapsto \mathbb{R}^d$ vanishing at infinity,*

$$\mathbb{E}_p \|\nabla \log p - f\|^2 = \mathbb{E}_p \left[\|f\|^2 + 2\nabla \cdot f + \|\nabla \log p\|^2 \right].$$

Proof. First, we use the identity integration by parts to derive,

$$\begin{aligned}\mathbb{E}_p[\langle \nabla \log p, f \rangle] &= \int \langle \nabla \log p(x), f(x) \rangle p(x) \, \mathrm{d}x \\ &= \int \langle \nabla p(x), f(x) \rangle \, \mathrm{d}x = -\int \nabla \cdot f(x) p(x) \, \mathrm{d}x = -\mathbb{E}_p[\nabla \cdot f].\end{aligned}$$

We now complete the proof as follows:

$$\mathbb{E}_p \|\nabla \log p - f\|^2 = \mathbb{E}_p \left[\|f\|^2 - 2\langle \nabla \log p, f \rangle + \|\nabla \log p\|^2 \right] = \mathbb{E}_p \left[\|f\|^2 + 2\nabla \cdot f + \|\nabla \log p\|^2 \right].$$

□

An immediate consequence of this result is that:

$$\arg\min_{f \in \mathscr{F}} \mathbb{E}_p \|\nabla \log p - f\|^2 = \arg\min_{f \in \mathscr{F}} L[f] := \mathbb{E}_p \left[\|f\|^2 + 2\nabla \cdot f \right].$$

The loss $L[f]$ is called the (population) score matching loss. Given samples $x_1, \ldots, x_n \sim p(x)$ iid, we can form the empirical score matching loss:

$$\hat{L}_n[f] := \frac{1}{n} \sum_{i=1}^{n} \left[\|f(x_i)\|^2 + 2\nabla \cdot f(x_i) \right]. \tag{2.12}$$

Note that unlike the cross-entropy loss, computing gradients of the score matching loss is straightforward given a reasonable auto-differentiation library.

Curl-free vector fields. One issue which arises from directly representing the vector field corresponding to the score function, is that in general it is difficult to directly ensure that a vector field is *curl-free*, meaning that the Jacobian of the vector field is a symmetric matrix. However, we do know that the true score function has this property, since

$$\frac{\mathrm{d}\nabla \log p(x)}{\mathrm{d}x} = \nabla^2 \log p(x),$$

and Hessian matrices are symmetric. One simple way to enforce this symmetry condition is to indirectly parameterize a vector field through the gradient of a scalar potential, e.g. using a function class of the form:

$$\mathscr{F} = \{ x \mapsto \nabla \phi(x) \mid \phi : \mathsf{X} \mapsto \mathbb{R}, \ \phi \in \mathscr{G} \}, \tag{2.13}$$

where $\mathscr{G}$ is a function class of smooth potential functions (functions mapping $\mathsf{X} \mapsto \mathbb{R}$). Note that this parameterization also comes with the advantage that it directly searches for an energy model (2.1), showing that we can actually think of the score matching loss as an alternative loss for learning an energy model (e.g., score matching and EBMs are not incompatible).

While the parameterization (2.13) does enforce the curl-free condition, it is not often used in practice since it comes with additional computational overhead. Specifically, we see that the empirical loss $\hat{L}_n[f]$ (cf. Equation (2.12)) now involves the Laplacian of ϕ. Hence, taking the gradient of $\hat{L}_n[f]$ requires taking *three* derivatives, which can become a computational bottleneck. Therefore in practice the lack of curl-free vector fields is often simply ignored, and it does not seem to result in any performance loss.

Sliced score matching. Despite the empirical score matching loss $\hat{L}_n[f]$ not requiring MCMC sampling to differentiate, computing the divergence of f does introduce some computational overhead. However, we can utilize randomized estimators to mitigate this issue, as shown in the following exercise.

Exercise 2.2. Let $f : \mathbb{R}^d \mapsto \mathbb{R}$ be a smooth function, and let $\partial f(x) \in \mathbb{R}^{d \times d}$ denote its Jacobian matrix evaluated at x.

(a) Show that for every x,
$$\mathbb{E}_v \langle v, \partial f(x) \cdot v \rangle = \nabla \cdot f(x), \quad v \sim N(0, I).$$

(b) Can you come up with an alternative distribution on v such that the identity above still hold?

(c) Use this identity to propose a modification to the empirical score matching loss $\hat{L}_n[f]$, and argue why your proposed loss is computationally less burdensome than the original empirical loss $\hat{L}_n[f]$.

2.2.1 Score matching with Gaussians

Let us now work through a specific example of score matching for intuition, instantiating it on Gaussian distributions. Suppose that $p(x) \sim N(\mu, I)$, where $\mu \in \mathbb{R}^d$ is unknown (but the covariance I is known). A quick calculation yields that the score function of $p(x)$ is the following affine function:

$$\nabla \log p(x) = -(x - \mu).$$

Suppose our family of score functions $\mathscr{F}$ is the set of Gaussian score functions corresponding to different means, e.g.,

$$\mathscr{F} = \{x \mapsto -(x - \theta) \mid \theta \in \mathbb{R}^d\}.$$

Then we have that:

$$L[\theta] = \mathbb{E}_{x \sim N(\mu, I)} \left[\|x - \theta\|^2 - 2\Delta \cdot (x - \theta) \right] = \mathbb{E}_{x \sim N(\mu, I)} \|x - \theta\|^2 - 2d = \|\theta - \mu\|^2 - d.$$

Clearly at the population level, $\arg\min_{\theta \in \mathbb{R}^d} L[\theta] = \mu$, which is correct. Now at the empirical level, we have:

$$\hat{L}_n[\theta] = \frac{1}{n} \sum_{i=1}^{n} \|x_i - \theta\|^2 - 2d.$$

Therefore, the empirical risk minimizer of $\hat{L}_n[\theta]$ is the sample mean (check this!):

$$\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

2.2.2 Score matching with exponential families

We can generalize beyond the Gaussian example we saw previously. Suppose that $p(x)$ has the form of an exponential family:

$$p(x) = \exp(\langle \phi(x), \theta_\star \rangle)/Z[\theta_\star], \quad Z[\theta] = \int \exp(\langle \phi(x), \theta \rangle)\, \mathrm{d}x, \quad \theta \in \Theta \subseteq \mathbb{R}^p.$$

Here, we assume the sufficient statistic $\phi : \mathsf{X} \mapsto \mathbb{R}^p$ is known. Let the component-wise functions of ϕ be denoted as $\phi = (\phi_1, \ldots, \phi_p)$, and let $J_\phi(x) \in \mathbb{R}^{p \times d}$ denote the Jacobian of ϕ, i.e., $J_\phi = \frac{\mathrm{d}\phi}{\mathrm{d}x}$. A quick calculation yields:

$$\nabla \log p(x) = J_\phi^\mathsf{T}(x)\theta_\star.$$

As we did in the realizable case, let us consider the following function class:

$$\mathscr{F} = \{x \mapsto J_\phi^\mathsf{T}(x)\theta \mid \theta \in \Theta\}.$$

The score matching loss has the following least-squares form given as follows:

$$L[\theta] = \mathbb{E}_p \left[\|J_\phi^\mathsf{T}(x)\theta\|^2 + 2\langle h_\phi(x), \theta \rangle \right], \quad h_\phi(x) = (\Delta\phi_1(x), \ldots, \Delta\phi_p(x)).$$

To see this, we observe that $J_\phi^\mathsf{T}(x)\theta = \sum_{i=1}^p \nabla\phi_i(x)\theta_i$, and therefore:

$$\nabla \cdot J_\phi^\mathsf{T}(x)\theta = \sum_{i=1}^p \nabla \cdot \nabla\phi_i(x)\theta_i = \sum_{i=1}^p \Delta\phi_i(x)\theta_i = \langle h_\phi(x), \theta \rangle.$$

Hence the empirical loss $\hat{L}_n[\theta]$ has the following form:

$$\hat{L}_n[\theta] = \frac{1}{n}\sum_{i=1}^n \left[\|J_\phi^\mathsf{T}(x_i)\theta\|^2 + 2\langle h_\phi(x_i), \theta \rangle \right].$$

This is a least-squares problem with one solution given by:

$$\hat{\theta}_n = -\left(\sum_{i=1}^n J_\phi(x_i) J_\phi^\mathsf{T}(x_i) \right)^\dagger \left(\sum_{i=1}^n h_\phi(x_i) \right).$$

Here, $(\cdot)^\dagger$ denotes the pseudo-inverse of a matrix.

2.2.3 Disadvantages of score matching

While the score matching loss is an improvement over the cross-entropy loss, there are a few disadvantages:

(a) The divergence calculation in the score matching loss (2.12) adds non-negligible computational overhead. Indeed, computing a gradient of $\hat{L}_n[f]$ requires at least two derivatives (three if using the scalar potential parameterization (2.13)). While the score matching loss is a clear improvement over the cross-entropy loss, there is still some computational burden.

(b) Score matching requires that the underlying distribution $p(x)$ to be learned is absolutely continuous w.r.t. the Lebesgue measure. This means that $p(x)$ is not allowed to be supported on a lower dimensional manifold. However, when the event space X corresponds to say, RGB images, this may not be a realistic assumption. Note that this disadvantage is also present with EBMs, which also starts by postulating the existence of a density on X.

(c) Despite the improve score matching loss over the cross-entropy loss, sampling from a learned score function still involves MCMC sampling (cf. Section 2.1.1), and therefore comes with all the disadvantages of MCMC sampling.

Our goal in the subsequent sections will be to address these concerns. Foreshadowing a bit, it turns out denoising diffusion models actually address all these shortcomings (quite cleverly actually). However, in order to fully appreciate the ingenuity behind diffusion models, we will proceed more or less in chronological order and consider the solutions which came before in the literature (which all contain fundamental ideas in generative modeling which are still extremely relevant).

2.3 Denoising score matching

TODO: Switch the notation in this section: $z \mapsto x$ and $x \mapsto \tilde{x}$.

One important step towards addressing the issues brought up in Section 2.2.3 regarding score matching is the idea of *denoising score matching*, which we will cover here. The idea is quite straightforward and intuitive. Regardless of what properties the distribution $p(x)$, the noised distribution $p_\sigma = p \star N(0, \sigma^2 I)$, where $\star$ denotes convolution, has the following properties:

(a) For any $\sigma > 0$, p_σ is fully supported on $\mathbb{R}^d$, meaning that it has a well-defined density function on $\mathbb{R}^d$. This holds even if $p(x)$ does not.

(b) As long as σ is not too large, then p and p_σ should not be too different in some sense.

Given these observations, denoising score matching simply applies score matching to p_σ instead of p. However, because of the special structure of p_σ, the resulting loss function simplifies even more compared to the vanilla score matching loss. To see this, we let $q_\sigma(x \mid z)$ denote the density of the distribution $N(z, \sigma^2 I)$, so that we can write the density of p_σ as:

$$p_\sigma(x) = \mathbb{E}_{z \sim p}[q_\sigma(x \mid z)].$$

With this notation, letting $(z, x) \sim p \times q_\sigma$ denote sampling from the joint distribution with $z \sim p$ and $x \sim q_\sigma(\cdot \mid z)$, observe that for any $f : \mathsf{X} \mapsto \mathbb{R}$,

$$\begin{aligned}
\mathbb{E}_{p_\sigma} \langle f, \nabla \log p_\sigma \rangle &= \int \langle f(x), \nabla \log p_\sigma(x) \rangle p_\sigma(x) \, \mathrm{d}x \\
&= \int \langle f(x), \nabla p_\sigma(x) \rangle \, \mathrm{d}x \\
&= \int \langle f(x), \nabla \mathbb{E}_{z \sim p}[q_\sigma(x \mid z)] \rangle \, \mathrm{d}x \\
&= \int \langle f(x), \mathbb{E}_{z \sim p}[\nabla_x q_\sigma(x \mid z)] \rangle \, \mathrm{d}x \\
&= \mathbb{E}_{z \sim p} \int \langle f(x), \nabla_x q_\sigma(x \mid z) \rangle \, \mathrm{d}x \\
&= \mathbb{E}_{z \sim p} \int \langle f(x), \nabla_x \log q_\sigma(x \mid z) \rangle q_\sigma(x \mid z) \, \mathrm{d}x \\
&= \mathbb{E}_{(z,x) \sim p \times q_\sigma} \langle f(x), \nabla_x \log q_\sigma(x \mid z) \rangle. \qquad (2.14)
\end{aligned}$$

With this calculation, letting $(z, x) \sim p \times q_\sigma$, the score matching loss on p_σ simplifies to:

$$\begin{aligned}
\mathbb{E}_{p_\sigma} \|f - \nabla \log p_\sigma\|^2 &= \mathbb{E}_{p_\sigma}[\|f\|^2 + \|\nabla \log p_\sigma\|^2] - 2\mathbb{E}_{p_\sigma} \langle f, \nabla \log p_\sigma \rangle \\
&= \mathbb{E}_{p_\sigma}[\|f\|^2 + \|\nabla \log p_\sigma\|^2] - 2\mathbb{E}_{(z,x)} \langle f(x), \nabla_x \log q_\sigma(x \mid z) \rangle \qquad \text{using (2.14)} \\
&= \mathbb{E}_{p_\sigma}[\|f\|^2 + \|\nabla \log p_\sigma\|^2] - 2\mathbb{E}_{(z,x)} \langle f(x), \nabla_x \log q_\sigma(x \mid z) \rangle \\
&\quad + \mathbb{E}_{(z,x)}[\|\nabla_x \log q_\sigma(x \mid z)\|^2 - \|\nabla_x \log q_\sigma(x \mid z)\|^2] \\
&= \mathbb{E}_{(z,x)} \|f(x) - \nabla_x \log q_\sigma(x \mid z)\|^2 \qquad \text{Law}(x) = p_\sigma \\
&\quad + \mathbb{E}_{(z,x)}[\|\nabla \log p_\sigma(x)\|^2 - \|\nabla_x \log q_\sigma(x \mid z)\|^2].
\end{aligned}$$

Hence for any function class $\mathscr{F}$,

$$\arg\min_{f\in\mathscr{F}} \mathbb{E}_{p_\sigma}\|f - \nabla \log p_\theta\|^2 = \arg\min_{f\in\mathscr{F}} L[f] := \mathbb{E}_{(z,x)}\|f(x) - \nabla_x \log q_\sigma(x \mid z)\|^2. \tag{2.15}$$

The loss $L[f]$ is known as the *denoising loss*. Its empirical counterpart $\hat{L}_n[f]$ can be constructed by taking $\{z_i\}_{i=1}^n$ drawn i.i.d. from $p(x)$, taking $\{w_i\}_{i=1}^n$ drawn i.i.d. from $N(0, I)$ (and independent of the z_i's), and then setting:

$$\hat{L}_n[f] = \frac{1}{n}\sum_{i=1}^n \|f(z_i + \sigma w_i) - \nabla_x \log q_\sigma(z_i + \sigma w_i \mid z_i)\|^2.$$

Observe that the denoising loss resolves disadvantages (a) and (b) described in Section 2.2.3. Indeed, (a) is resolved since the denoising loss is a standard least-squares loss; no divergence calculations are needed anymore. Furthermore, (b) is resolved since adding Gaussian noise makes p_σ full dimensional even with p is not.

Let us see where the term "denoising" comes from. A quick calculation yields that

$$\nabla_x \log q_\sigma(x \mid z) = -\frac{1}{\sigma^2}(x - z),$$

which is proportional to the (negative) noise vector added to make the noised sample x from the pure sample z. Thus, we can interpret the denoising loss as trying to, given a noised sample x, predict the direction $f(x)$ such that $x + \sigma^2 f(x) \approx z$ is a denoised version of x.

Exercise 2.3. Use the identity established in Equation (2.15) to prove *Tweedie's formula*:

$$\nabla \log p_\sigma(x) = \frac{\mathbb{E}[z \mid x] - x}{\sigma^2}.$$

To be clear, here $\mathbb{E}[z \mid x]$ is understood to be the function $x \mapsto \mathbb{E}[Z \mid X = x]$ with $(Z, X) \sim p \times q_\sigma$.

References

Pierre Alquier. User-friendly introduction to pac-bayes bounds. *arXiv preprint arXiv:2110.11216*, 2021.

Moritz Hardt and Benjamin Recht. *Patterns, predictions, and actions: Foundations of machine learning*. Princeton University Press, 2022.

Tengyu Ma. Lecture notes for machine learning theory (CS229M/STATS214). `https://raw.githubusercontent.com/tengyuma/cs229m_notes/main/master.pdf`, 2022.

A Basic properties of probability divergences

We first state the definition of the *Kullback–Leibler Divergence*. We state it in a general measure-theoretic form, but we will instantiate the definition for some special cases as follows.

Definition A.1 (KL-Divergence)**.** *Let μ, ν be two probability measures on the same measure space. Suppose that μ is absolutely continuous w.r.t. ν, i.e., $\mu \ll \nu$, and let $\frac{\mathrm{d}\mu}{\mathrm{d}\nu}$ denote the Radon-Nikodym derivative of μ w.r.t. ν. The KL-divergence* $\mathrm{KL}(\mu \,\|\, \nu)$ *is defined as:*

$$\mathrm{KL}(\mu \,\|\, \nu) = \mathbb{E}_\mu\left[\log \frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right].$$

Proposition A.2. *Suppose that $\mu \ll \nu$. We have that* $\mathrm{KL}(\mu \,\|\, \nu) \geqslant 0$.

Proof. This proof uses the f-divergence representation of KL-divergence to deal with the case when $\nu \ll \mu$ does not hold. Let $f(t) = t \log t$. One can check that this function is convex on $\mathbb{R}_+$. Furthermore, by a change of measure, and Jensen's inequality

$$\mathrm{KL}(\mu \parallel \nu) = \int \log \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \, \mathrm{d}\mu = \int \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \log \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \, \mathrm{d}\nu = \int f\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right) \mathrm{d}\nu \geqslant f\left(\int \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \, \mathrm{d}\nu\right) = f(1) = 0.$$

□

Lemma A.3 (Donsker-Varadhan)**.** *Let p, q be two probability measures on the same measure space, and suppose that p is absolutely continuous w.r.t. q. We have that:*

$$\mathrm{KL}(p \parallel q) = \sup_g \left\{ \mathbb{E}_p[g] - \log \mathbb{E}_q[\exp(g)] \right\},$$

where the supremum is taken over all measurable $g : \Theta \to \mathbb{R}$ satisfying $\mathbb{E}_q[\exp(g)] < \infty$.

Equivalently, let q be a probability measure and let g be a measurable function satisfying $\mathbb{E}_q[\exp(g)] < \infty$. We have that:

$$\log \mathbb{E}_q[\exp(g)] = \sup_p \left\{ \mathbb{E}_p[g] - \mathrm{KL}(p \parallel q) \right\},$$

where the supremum is taken over probability measures p which are absolutely continuous w.r.t. q.

Proof. Let us define the Gibbs measure with density:

$$\frac{\mathrm{d}\pi}{\mathrm{d}q}(\theta) = \frac{\exp(g(\theta))}{\mathbb{E}_q[\exp(g)]}.$$

Computing the KL-divergence between p and π,

$$\begin{aligned}
\mathrm{KL}(p \parallel \pi) &= \int \log \left[\frac{\mathrm{d}p}{\mathrm{d}\pi}\right] \mathrm{d}p = \int \log \left[\frac{\mathrm{d}p}{\mathrm{d}q} \frac{\mathrm{d}q}{\mathrm{d}\pi}\right] \mathrm{d}p \\
&= \mathrm{KL}(p \parallel q) + \int \log \left[\frac{\mathbb{E}_q[\exp(g)]}{\exp(g)}\right] \mathrm{d}p \\
&= \mathrm{KL}(p \parallel q) + \log \mathbb{E}_q[\exp(g)] - \mathbb{E}_p[g].
\end{aligned}$$

Since $\mathrm{KL}(p \parallel \pi) \geqslant 0$, this shows:

$$\mathrm{KL}(p \parallel q) \geqslant \mathbb{E}_p[g] - \log \mathbb{E}_q[\exp(g)] \iff \log \mathbb{E}_q[\exp(g)] \geqslant \mathbb{E}_p[g] - \mathrm{KL}(p \parallel q).$$

Furthermore, equality is achieved when $p = \pi$, i.e., when $g = \log \frac{\mathrm{d}p}{\mathrm{d}q}$. □

Another often used probability divergence is the total-variation distance.

Definition A.4. *Let μ, ν be two probability measures on the same measure space $(\Omega, \mathcal{B})$. The total-variation distance (TV-distance) is defined as:*[10]

$$\|\mu - \nu\|_{\mathrm{tv}} = \sup_{A \in \mathcal{B}} |\mu(A) - \nu(A)|.$$

[10] For those unfamiliar with measure theory, you can think of the supremum as being taken over all subsets of Ω.

B Concentration inequalities

We give a brief primer on concentration equalities needed for this course. The starting point is Markov's inequality.

Proposition B.1 (Markov's inequality)**.** *Let X be a non-negative random variable. Then,*

$$\forall t > 0,\ \mathbb{P}\{X \geqslant t\} \leqslant \frac{\mathbb{E}[X]}{t}.$$

Proof. Using the fact that X is non-negative,

$$\mathbb{E}[X] = \mathbb{E}[X\mathbf{1}\{X \geqslant t\}] + \mathbb{E}[X\mathbf{1}\{X < t\}] \geqslant \mathbb{E}[t\mathbf{1}\{X \geqslant t\}] = t\mathbb{P}\{X \geqslant t\}.$$

□

Markov's inequality is often used in conjunction with a monotonically increasing bijection, i.e., $\{X \geqslant t\} = \{\phi(X) \geqslant \phi(t)\}$ for any monotonically increasing bijection ϕ. A very useful transform is $\phi(x) = \exp(\lambda x)$ for any $\lambda > 0$:

Proposition B.2 (Laplace transform)**.** *Let X be a random variable. Then,*

$$\forall t > 0,\ \mathbb{P}\{X \geqslant t\} \leqslant \inf_{\lambda > 0} e^{-\lambda t}\mathbb{E}\exp(\lambda X).$$

Proof. For any $\lambda > 0$, $\{X \geqslant t\} = \{\exp(\lambda X) \geqslant \exp(\lambda t)\}$. Now apply Markov's inequality to the RHS event. Since $\lambda > 0$ is arbitrary, take the infimum over all $\lambda > 0$ to conclude. □

The quantity $\mathbb{E}\exp(\lambda X)$ is referred to as the *moment generating function* (MGF) of X. It has a very useful property that it "tensorizes" in the following away for independent sums. If $S_n = X_1 + \cdots + X_n$ where $X_i \perp X_j$ for $i \neq j$, then

$$\mathbb{E}\exp(\lambda S_n) = \prod_{i=1}^{n} \mathbb{E}\exp(\lambda X_i).$$

Notice that the RHS is the product of the individual MGFs, a property we will soon utilize.

Our next order of business will be to obtain control of MGFs. Without any assumptions, the MGF does not even need to be finite. So, we will need to impose some conditions on our random variables so that the MGFs do exist. The starting point of study will be the Gaussian distribution. Suppose that $X \sim N(0, \sigma^2)$. Then, one can show that:

$$\forall \lambda > 0,\ \mathbb{E}\exp(\lambda X) = \exp(\lambda^2\sigma^2/2). \tag{B.1}$$

Exercise B.3. Show that (B.1) is true.
Hint: write out $\mathbb{E}\exp(\lambda X)$ in integral form and complete the square.

The Gaussian MGF motivates the following definition of a random variable, which we will utilize quite extensively in this course.

Definition B.4 (sub-Gaussian random variable)**.** *Let X be a random variable with finite expectation. The random variable X is σ-sub-Gaussian if:*

$$\forall \lambda > 0,\ \mathbb{E}\exp(\lambda(X - \mathbb{E}[X])) \leqslant \exp(\lambda^2\sigma^2/2).$$

Clearly, a $N(\mu, \sigma^2)$ Gaussian random variable is σ-sub-Gaussian by definition. Also, it is not hard to see that constant scalings of sub-Gaussian random variables remain sub-Gaussian. That is, if $a \in \mathbb{R}$ is fixed and X is σ-sub-Gaussian, then aX is $|a|\sigma$-sub-Gaussian, since:

$$\mathbb{E}\exp(\lambda(aX - \mathbb{E}[aX])) \leqslant \exp(\lambda^2 a^2 \sigma^2/2) = \exp(\lambda^2 (a\sigma)^2/2),$$

Next, we will show that bounded random variables are also sub-Gaussian. To do this, we start with a definition of a *Rademacher random variable*, which will play a critical role in this class.

Definition B.5 (Rademacher random variable)**.** *A Rademacher random variable ε satisfies:*

$$\mathbb{P}\{\varepsilon = +1\} = \mathbb{P}\{\varepsilon = -1\} = 1/2.$$

Proposition B.6. *A Rademacher random variable is* 1*-sub-Gaussian.*

Proof. Let ε be a Rademacher random variable, and let $\lambda \in \mathbb{R}$ be fixed. We have:

$$\begin{aligned}
\mathbb{E}\exp(\lambda\varepsilon) &= \frac{1}{2}(\exp(\lambda) + \exp(-\lambda)) \\
&= \frac{1}{2}\left[\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} + \sum_{k=0}^{\infty} \frac{(-\lambda)^k}{k!}\right] && \text{Taylor series of } \exp(x) \\
&= \sum_{k=0}^{\infty} \frac{\lambda^{2k}}{(2k)!} && \text{odd terms cancel} \\
&\leqslant \sum_{k=0}^{\infty} \frac{\lambda^{2k}}{2^k k!} && \text{since } (2k)! \geqslant 2^k k! \\
&= \sum_{k=0}^{\infty} \frac{(\lambda^2/2)^k}{k!} = \exp(\lambda^2/2).
\end{aligned}$$

□

We can now show that bounded random variables are sub-Gaussian.

Proposition B.7 (Hoeffding's lemma)**.** *Let X be a random variable satisfying $X \in [-1, 1]$ a.s. Then, X is* 1*-sub-Gaussian.*

Proof. Note: we will prove this result with a slightly worse constant. Specifically, we will prove that X is 2-sub-Gaussian. However, the proof technique we will used illustrates a lot of the key ideas in studying concentration inequalities. The proof for the sharper constant is somewhat less intuitive.

This is our first example of a symmetrization argument, which will be a technique we return to many times. Let ε be a Rademacher random variable, and let X' be a copy of X (all of X, X', ε are mutually independent). The key observation is that $(X - X')$ has the same distribution as $\varepsilon(X - X')$, due to symmetry. We now let $\lambda \in \mathbb{R}$ be fixed, and proceed as follows:

$$\begin{aligned}
\mathbb{E}\exp(\lambda(X - \mathbb{E}[X])) &= \mathbb{E}\exp(\lambda(X - \mathbb{E}[X'])) && \text{since } X \overset{(d)}{=} X' \\
&\leqslant \mathbb{E}\exp(\lambda(X - X')) && \text{Jensen's inequality} \\
&= \mathbb{E}\exp(\lambda\varepsilon(X - X')) && \text{since } X - X' \overset{(d)}{=} \varepsilon(X - X') \\
&= \mathbb{E}[\mathbb{E}[\exp(\lambda\varepsilon(X - X')) \mid X, X']] && \text{tower property} \\
&= \mathbb{E}\exp(\lambda^2 (X - X')^2/2) && \text{since } \varepsilon \text{ is 1-sub-Gaussian} \\
&\leqslant \exp(\lambda^2 2^2/2) && \text{since } X \in [-1, 1] \text{ a.s.}
\end{aligned}$$

□

Exercise B.8. Let X be a random variable satisfying $X \in [-a, a]$ a.s. for some $a > 0$. Show that X is a-sub-Gaussian.

Exercise B.9. Let $X_1, \dots, X_n$ be independent random variables, and suppose that X_i is σ_i-sub-Gaussian for $i = 1, \dots, n$. Show that $S_n = \sum_{i=1}^n X_i$ is $\sqrt{\sum_{i=1}^n \sigma_i^2}$-sub-Gaussian.

We are now in a position to prove our first concentration inequality.

Proposition B.10 (sub-Gaussian tail bound)**.** *Let X be a σ-sub-Gaussian random variable. We have that:*

$$\mathbb{P}\{X - \mathbb{E}[X] \geqslant t\} \leqslant \exp(-t^2/(2\sigma^2)).$$

Consequently,

$$\mathbb{P}\{|X - \mathbb{E}[X]| \geqslant t\} \leqslant 2\exp(-t^2/(2\sigma^2)).$$

Proof. This is a simple consequence of the Laplace transform in addition to the sub-Gaussian MGF bound. For any $\lambda > 0$,

$$\begin{aligned}\mathbb{P}\{X - \mathbb{E}[X] \geqslant t\} &\leqslant e^{-\lambda t}\mathbb{E}\exp(\lambda(X - \mathbb{E}[X])) \\ &\leqslant \exp\left\{-\lambda t + \lambda^2\sigma^2/2\right\}.\end{aligned}$$

Now choosing $\lambda = t/\sigma^2$ to minimize the quadratic form inside the exponential, we obtain $\mathbb{P}\{X - \mathbb{E}[X] \geqslant t\} \leqslant \exp(-t^2/(2\sigma^2))$. Applying the same argument to the random variable $-X$ yields $\mathbb{P}\{X - \mathbb{E}[X] \leqslant -t\} \leqslant \exp(-t^2/(2\sigma^2))$ (why is $-X$ also σ-sub-Gaussian?). Now we conclude by a union bound:

$$\begin{aligned}\mathbb{P}\{|X - \mathbb{E}[X]| \geqslant t\} &= \mathbb{P}\{\{X - \mathbb{E}[X] \geqslant t\} \cup \{X - \mathbb{E}[X] \leqslant -t\}\} \\ &\leqslant \mathbb{P}\{X - \mathbb{E}[X] \geqslant t\} + \mathbb{P}\{X - \mathbb{E}[X] \leqslant -t\} \leqslant 2\exp(-t^2/(2\sigma^2)).\end{aligned}$$

□

Exercise B.11. Let X be a zero-mean, 1-sub-Gaussian random variable. Show that there exists a universal constant $c > 0$ such that for all $p \geqslant 1$, $\|X\|_{L_p} := (\mathbb{E}|X|^p)^{1/p} \leqslant c\sqrt{p}$.

Hint: You will need to utilize the following facts (which you can use without proof):

(a) For a non-negative random variable X, $\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X > t)\,\mathrm{d}t$.

(b) For all $s \geqslant 1/2$, $\Gamma(s) \leqslant 3s^s$, where $\Gamma(s) := \int_0^\infty t^{s-1}e^{-t}\,\mathrm{d}t$ is the Gamma function.

A nearly immediate consequence is Hoeffding's inequality.

Proposition B.12 (Hoeffding's inequality)**.** *Let $X_1, \dots, X_n$ be independent random variables with $X_i \in [-a_i, a_i]$ a.s. Put $S_n = \sum_{i=1}^n X_i$. We have that:*

$$\mathbb{P}\{S_n - \mathbb{E}[S_n] \geqslant t\} \leqslant \exp\left(-\frac{t^2}{2\sum_{i=1}^n a_i^2}\right).$$

Consequently,

$$\mathbb{P}\{|S_n - \mathbb{E}[S_n]| \geqslant t\} \leqslant 2\exp\left(-\frac{t^2}{2\sum_{i=1}^n a_i^2}\right).$$

Proof. By Exercise B.8, we have that each X_i is a_i-sub-Gaussian. Furthermore, by Exercise B.9, we have that S_n is $\sqrt{\sum_{i=1}^n a_i^2}$-sub-Gaussian. Hence by the sub-Gaussian tail bound Proposition B.10, we have:

$$\mathbb{P}\{S_n - \mathbb{E}[S_n] \geqslant t\} \leqslant \exp\left(-\frac{t^2}{2\sum_{i=1}^n a_i^2}\right).$$

To control $\mathbb{P}\{|S_n - \mathbb{E}[S_n]| \geqslant t\}$, use the same argument in Proposition B.10. □

Our next result will be the well known bounded differences inequality. To do this, we will need the concept of a martingale.

Definition B.13. *Let $\{X_t\}_{t\geqslant 0}$ be a stochastic process adapted to the filtration $\{\mathcal{F}_t\}_{t\geqslant 0}$. The process $\{X_t\}$ is a martingale if*

$$\forall t \geqslant 0,\ \mathbb{E}[X_{t+1} \mid \mathcal{F}_t] = X_t.$$

Definition B.14. *Let $\{X_t\}_{t\geqslant 0}$ be a martingale. We say that $\{X_t\}_{t\geqslant 0}$ is a $\{\sigma_t\}_{t\geqslant 0}$-sub-Gaussian martingale if for all $t \geqslant 0$, the random variable $(X_{t+1} - X_t) \mid \mathcal{F}_t$ is σ_t-sub-Gaussian almost surely, i.e.,*

$$\forall t \geqslant 0, \lambda \in \mathbb{R},\ \mathbb{E}[\exp(\lambda(X_{t+1} - X_t)) \mid \mathcal{F}_t] \leqslant \exp(\lambda^2\sigma_t^2/2) \text{ a.s.}$$

Proposition B.15 (sub-Gaussian martingale tail inequality)**.** *Let $\{X_t\}_{t\geqslant 0}$ be a $\{\sigma_t\}_{t\geqslant 0}$-sub-Gaussian martingale. We have that:*

$$\forall t > 0,\ \mathbb{P}\{X_n - X_0 \geqslant t\} \leqslant \exp\left(-t^2/\left(2\sum_{i=0}^{n-1}\sigma_t^2\right)\right).$$

Consequently,

$$\forall t > 0,\ \mathbb{P}\{|X_n - X_0| \geqslant t\} \leqslant 2\exp\left(-t^2/\left(2\sum_{i=0}^{n-1}\sigma_t^2\right)\right).$$

Proof. By the Laplace transform method (cf. Proposition B.2) and the tower property, for any $\lambda > 0$:

$$\begin{aligned}
\mathbb{P}\{X_n - X_0 \geqslant t\} &\leqslant e^{\lambda t}\mathbb{E}\exp(\lambda(X_n - X_0)) \\
&= e^{-\lambda t}\mathbb{E}\exp\left(\sum_{i=0}^{n-1}\lambda(X_{i+1} - X_i)\right) \\
&= e^{-\lambda t}\mathbb{E}\left[\exp\left(\sum_{i=0}^{n-2}\lambda(X_{i+1} - X_i)\right)\mathbb{E}[\exp(\lambda(X_n - X_{n-1})) \mid \mathcal{F}_{n-1}]\right] \\
&\leqslant \exp\{-\lambda t + \lambda^2\sigma_{n-1}^2/2\}\mathbb{E}\exp\left(\sum_{i=0}^{n-2}\lambda(X_{i+1} - X_i)\right) \\
&\vdots \\
&\leqslant \exp\left\{-\lambda t + \lambda^2\left(\sum_{i=0}^{n-1}\sigma_i^2\right)/2\right\}.
\end{aligned}$$

Now choose $\lambda = t/\sum_{i=0}^{n-1}\sigma_i^2$, from which the claim follows. □

B.1 Bounded differences inequality

Proposition B.16 (Bounded differences inequality)**.** *Let $f : \mathsf{X}^n \mapsto \mathbb{R}$ satisfy the following bounded differences property:*

$$\forall i \in \{1, \ldots, n\},\ \sup_{x_1,\ldots,x_n,x' \in \mathsf{X}} |f(x_{1:i-1}, x_i, x_{i+1:n}) - f(x_{1:i-1}, x_i', x_{i+1:n})| \leqslant c_i.$$

Let $x_1, \ldots, x_n \in \mathsf{X}$ be independent random variables. Then,

$$\mathbb{P}\{f(x_1, \ldots, x_n) - \mathbb{E}[f(x_1, \ldots, x_n)] \geqslant t\} \leqslant \exp\left(-t^2/\left(2\sum_{i=1}^{n}c_i^2\right)\right).$$

Consequently,

$$\mathbb{P}\{|f(x_1,\dots,x_n)-\mathbb{E}[f(x_1,\dots,x_n)]|\geqslant t\}\leqslant 2\exp\left(-t^2/\left(2\sum_{i=1}^n c_i^2\right)\right).$$

Proof. The idea here is to construct a martingale, known as the Doob martingale. Let $\mathcal{F}_i=\sigma(x_{1:i})$ denote the minimal σ-algebra for $x_{1:i}$,[11] and set $Z_i=\mathbb{E}[f(x_{1:n})\mid\mathcal{F}_i]$. Let us first check this is a martingale. We have by the tower property of conditional expectations:

$$\mathbb{E}[Z_{i+1}\mid\mathcal{F}_i]=\mathbb{E}[\mathbb{E}[f(x_{1:n})\mid\mathcal{F}_{i+1}]\mid\mathcal{F}_i]=\mathbb{E}[f(x_{1:n})\mid\mathcal{F}_i]=Z_i.$$

Second, observe that:

$$Z_n=f(x_{1:n}),\ Z_0=\mathbb{E}[f(x_{1:n})]\implies Z_n-Z_0=f(x_{1:n})-\mathbb{E}[f(x_{1:n})].$$

Now let us bound, conditioned on $\mathcal{F}_{i-1}$,

$$\begin{aligned}
|Z_i-Z_{i-1}|&=\left|\int f(x_{1:n})\,\mathrm{d}x_{i+1:n}-\int f(x_{1:n})\,\mathrm{d}x_{i:n}\right|\\
&=\left|\int f(x_{1:i-1},x_i,x_{i+1:n})\,\mathrm{d}x_{i+1:n}-\int f(x_{i:i-1},x_i',x_{i+1:n})\,\mathrm{d}x_i'\mathrm{d}x_{i+1:n}\right|\\
&=\left|\int[f(x_{1:i-1},x_i,x_{i+1:n})-f(x_{i:i-1},x_i',x_{i+1:n})]\,\mathrm{d}x_i'\mathrm{d}x_{i+1:n}\right|\\
&\leqslant\int|f(x_{1:i-1},x_i,x_{i+1:n})-f(x_{i:i-1},x_i',x_{i+1:n})|\,\mathrm{d}x_i'\mathrm{d}x_{i+1:n} && \text{Jensen's inequality}\\
&\leqslant c_i.
\end{aligned}$$

(Check your understanding: where did we use independence of the x_i's above?) Hence, we have that for all $i\in\mathbb{N}_+$, that $(Z_i-Z_{i-1})\mid\mathcal{F}_{i-1}$ is a c_i-sub-Gaussian random variable almost surely (cf. Exercise B.8). Therefore, the Doob martingale $\{Z_i\}$ is a $\{c_i\}$-sub-Gaussian martingale. By Proposition B.15, we conclude that for all $t>0$,

$$\mathbb{P}\{f(x_{1:n})-\mathbb{E}[f(x_{1:n})]\geqslant t\}=\mathbb{P}\{Z_n-Z_0\geqslant t\}\leqslant\exp\left(-t^2/\left(2\sum_{i=1}^n c_i^2\right)\right).$$

□

B.2 Maximal inequalities

Proposition B.17 (sub-Gaussian maximal inequality)**.** *Let X_i, $i=1,\dots,n$ be zero-mean σ-sub-Gaussian random variables (not necessarily independent). Then,*

$$\mathbb{E}\max_{i=1,\dots,n}X_i\leqslant\sigma\sqrt{2\log n}.$$

Proof. Fix any $\lambda>0$. We have:

$$\begin{aligned}
\mathbb{E}\max_{i=1,\dots,n}X_i&=\lambda^{-1}\mathbb{E}\max_{i=1,\dots,n}\lambda X_i && \text{since } \lambda>0\\
&=\lambda^{-1}\mathbb{E}\log\exp\left(\max_{i=1,\dots,n}\lambda X_i\right)\\
&\leqslant\lambda^{-1}\log\mathbb{E}\exp\left(\max_{i=1,\dots,n}\lambda X_i\right) && \text{Jensen's inequality}
\end{aligned}$$

[11] If you are unfamiliar with this terminology, just think of $\mathcal{F}_i$ as encoding the information available in the first i datapoints.

$$
\begin{aligned}
&= \lambda^{-1} \log \mathbb{E} \max_{i=1,\dots,n} \exp(\lambda X_i) && \exp \text{ is monotonically increasing} \\
&\leqslant \lambda^{-1} \log \left(\sum_{i=1}^{n} \mathbb{E} \exp(\lambda X_i) \right) && \max_i a_i \leqslant \sum_i a_i \text{ if } \forall i,\ a_i \geqslant 0 \\
&\leqslant \lambda^{-1} \log \left(n \exp(\lambda^2 \sigma^2 / 2) \right) && \text{each } X_i \text{ is } \sigma\text{-sub-Gaussian} \\
&= \frac{\log n}{\lambda} + \frac{\sigma^2 \lambda}{2}.
\end{aligned}
$$

Since the inequality holds for any $\lambda > 0$, we can optimize over the bound by choosing $\lambda = \sqrt{2 \log n / \sigma^2}$. □

Note that this result is actually sharp: if $X_1, X_2, \dots$ are i.i.d. $N(0,1)$, then one can show **TODO:** cite:

$$
\lim_{n \to \infty} \frac{\mathbb{E} \max_{i=1,\dots,n} X_i}{\sqrt{2 \log n}} = 1.
$$

Exercise B.18. Let X_i, $i = 1, \dots, n$ be zero-mean σ-sub-Gaussian random variables. Show that

$$
\mathbb{E} \max_{i=1,\dots,n} |X_i| \leqslant \sigma \sqrt{2 \log(2n)}.
$$

C Convex functions

We give a very brief primer on basic properties of convex functions.

Definition C.1. *A set D is convex if for all $x, y \in D$, $\lambda x + (1-\lambda) y \in D$ for all $\lambda \in [0,1]$.*

Definition C.2. *A function $f : D \mapsto \mathbb{R}$ is convex if the domain D is convex and the functions satisfies for all $x, y \in D$ and $\theta \in [0,1]$,*

$$
f(\theta x + (1-\theta) y) \leqslant \theta f(x) + (1-\theta) f(y)
$$

Definition C.3. *A differentiable function $f : D \mapsto \mathbb{R}$ has L-Lipschitz gradients (alternatively, L-smooth), if ∇f is L-Lipschitz,*

$$
\|\nabla f(x) - \nabla f(y)\| \leqslant L \|x - y\| \ \forall x, y \in D.
$$

Proposition C.4. *Let $f : \mathbb{R}^n \mapsto \mathbb{R}$ be L-smooth. Then for all $x \in \mathbb{R}^n$,*

$$
\inf_{x \in \mathbb{R}^n} f(x) := f_\star \leqslant f(x) - \frac{1}{2L} \|\nabla f(x)\|^2.
$$

Proof. Taking a second order Taylor series expansion, we have that for all x, y:

$$
f(y) \leqslant f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|x - y\|^2.
$$

Choosing $y = x - \frac{1}{L} \nabla f(x)$ (which minimizes the RHS over $y \in \mathbb{R}^n$), we obtain:

$$
f(y) \leqslant f(x) - \frac{1}{2L} \|f(x)\|^2.
$$

By definition, $f_\star \leqslant f(y)$, which yields the claim. □

Proposition C.5 (Co-coercive gradients)**.** *Let $f : \mathbb{R}^n \mapsto \mathbb{R}$ be convex and L-smooth. Then,*

$$
\langle \nabla f(x) - \nabla f(y), x - y \rangle \geqslant \frac{1}{L} \|\nabla f(x) - \nabla f(y)\|^2 \ \forall x, y \in \mathbb{R}^n.
$$

Proof. Fix an $x \in \mathbb{R}^n$, and define the function:

$$f_x(z) = f(z) - \langle \nabla f(x), z \rangle.$$

Since f is convex, so is $f_x(z)$ (why?). Hence, we can study its stationary points $\nabla f_x(z) = 0$ to understand its minimizers. Taking $\nabla f_x(z) = \nabla f(z) - \nabla f(x)$, we find that $z = x$ minimizes $f_x(z)$. Furthermore, since f is L-smooth, so is $f_x(z)$ (why?). Therefore, applying Proposition C.4, we conclude that for any $y \in \mathbb{R}^n$,

$$f(y) - f(x) - \langle \nabla f(x), y - x \rangle = f_x(y) - f_x(x) \geqslant \frac{1}{2L} \|\nabla f_x(y)\|^2 = \frac{1}{2L} \|\nabla f(y) - \nabla f(x)\|^2.$$

Next, we can flip the role of x and y in the above argument and conclude

$$f(x) - f(y) - \langle \nabla f(y), x - y \rangle \geqslant \frac{1}{2L} \|\nabla f(y) - \nabla f(x)\|^2.$$

Combining these two inequalities yields the claim. □

Definition C.6 (Strong convexity)**.** *A function $f : D \mapsto \mathbb{R}$ is* μ-strongly-convex *if the domain D is convex and the following holds for all $\theta \in [0, 1]$:*

$$f(\theta x + (1 - \theta) y) \leqslant \theta f(x) + (1 - \theta) f(y) - \frac{\theta(1 - \theta)\mu}{2} \|x - y\|^2 \ \forall x, y \in D.$$