CS861: Theoretical Foundations of Machine Learning

Chapter 3: Statistical Learning Theory

Kirthevasan Kandasamy
UW-Madison

Learning with respect to a hypothesis class

Consider the following setting:

- We have a data space $\mathcal{Z}$, and an unknown distribution P over $\mathcal{Z}$.
- We are given a training set $S = \{Z_i\}_{i=1}^n$ of n points drawn i.i.d from P.
- Let $\mathcal{H}$ be a hypothesis class.
- Let $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$ be the *instance loss*, where $\ell(h, z)$ is the loss of hypothesis h on instance (data point) z.

Supervised learning. In supervised learning[1], the dataspace can be written as $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is a set of inputs and $\mathcal{Y}$ is a set of labels. The hypothesis class $\mathcal{H}$ is a set of predictors from $\mathcal{X}$ to $\mathcal{Y}$. Here are some examples:

- Binary classification with $0-1$ loss: $\mathcal{Y} = \{0, 1\}$, $\ell(h, (x, y)) = \mathbb{1}(h(x) \neq y)$.
- Regression with squared loss, $\mathcal{Y} \subset \mathbb{R}$, $\ell(h, (x, y)) = (h(x) - y)^2$.

[1]The ideas extend to any learning problem (including unsupervised learning) where we can define a loss ℓ on an instance but most common use cases for this framework are in supervised learning.

Learning with respect to a hypothesis class (cont'd)

- The loss of a hypothesis h on a distribution P,
$$L(h) = \mathbb{E}_{Z\sim P}\left[\ell(h, Z)\right].$$
- An algorithm $\widehat{h}$ chooses a hypothesis in $\mathcal{H}$ based on the dataset S, *i.e.*, $\widehat{h} : \mathcal{Z}^n \to \mathcal{H}$.
- The risk of an algorithm $\widehat{h}$,
$$R(\widehat{h}) = \mathbb{E}_{S\sim P^n}\left[L(\widehat{h}(S))\right] = \mathbb{E}_{S\sim P^n}\left[\mathbb{E}_{Z\sim P}\left[\ell(\widehat{h}(S), Z)\right]\right].$$
(We will drop the dependence on P if it is clear from context.)
- The excess risk relative to the hypothesis class $\mathcal{H}$ is $R(h) - \inf_{h'\in\mathcal{H}} L(h')$.

Outline for this chapter

1. ERM and uniform convergence
2. Rademacher Complexity
3. VC dimension and Sauer's lemma
4. Dudley Entroy Integral
5. Case study: Two-layer Neural Networks
6. Approximation error, Estimation error, optimization error

Ch 3.1: ERM and uniform convergence

A simple algorithm: Empirical risk minimization

For a hypothesis $h \in \mathcal{H}$, let us define the empirical loss[2] as,

$$\widehat{L}(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(h, Z_i).$$

ERM chooses a hypothesis in $\mathcal{H}$ which minimizes the empirical loss,

$$\widehat{h} \in \operatorname*{argmin}_{h \in \mathcal{H}} \widehat{L}(h).$$

Note that the quantities $\widehat{L}, \widehat{h}$, depend on the dataset S, but we have suppressed this dependence for simplicity.

[2] To be consistent with the terminology used in this class, we should refer to this as empirical *loss* minimization. However, the acronym ERM is more commonly used.

On ERM

- Many modern ML algorithms are implementations of ERM.
- It is very common to use a regularizer in practice:

$$\widehat{h} \in \operatorname*{argmin}_{h \in \mathcal{H}} \left(\widehat{L}(h) + \lambda \mathcal{C}(h) \right),$$

where, $\mathcal{C}$ is a measure of complexity of h. Often, this is equivalent to ERM over a restricted class, *i.e.*, $\operatorname{argmin}_{h \in \mathcal{H}} \widehat{L}(h)$ subject to $\mathcal{C}(h) \leq B$.

In the remainder of the chapter, we will analyize ERM.

Uniform convergence

We will show that we can bound the excess risk of ERM under uniform convergence. Let us first define uniform convergence.

Uniform convergence. A class $\mathcal{H}$ is said to have uniform convergence if there exists a function $N : \mathbb{R}_+ \times (0,1) \to \mathbb{N}$ such that, for all $\epsilon > 0, \delta \in (0,1)$, and distributions P over $\mathcal{Z}$, when $n \geq N(\epsilon, \delta)$, we have

$$\mathbb{P}\left(\forall h \in \mathcal{H},\, |\widehat{L}(h) - L(h)| \leq \epsilon\right) \geq 1 - \delta,$$

Here, recall that $\widehat{L}$ is computed based on n i.i.d samples from P, and the probability $\mathbb{P}$ is with respect to this randomness.

Pointwise convergence. We can contrast this with the (much weaker) pointwise convergence: $\forall h \in \mathcal{H},\ \mathbb{P}\left(|\widehat{L}(h) - L(h)| \leq \epsilon\right) \geq 1 - \delta$.

Uniform convergence (cont'd)

We have other equivalent ways to state uniform convergence.

1) There exists a function $\epsilon : \mathbb{N} \times (0,1) \to \mathbb{R}_+$ which satisfies *(i)* $\epsilon(n,\delta) \to 0$ as $n \to \infty$ for any fixed δ, and *(ii)* for all $n \in \mathbb{N}, \delta \in (0,1)$, and distribution P, we have

$$\mathbb{P}\left(\forall h \in \mathcal{H},\ |\widehat{L}(h) - L(h)| \leq \epsilon(n,\delta)\right) \geq 1 - \delta, \tag{1}$$

2) There exists a function $\delta : \mathbb{N} \times \mathbb{R}_+ \to (0,1)$ which satisfies, *(i)* $\delta(n,\epsilon) \to 0$ as $n \to \infty$ for any fixed ϵ, and *(ii)* for all $n \in \mathbb{N}$, $\epsilon > 0$, and distribution P, we have

$$\mathbb{P}\left(\forall h \in \mathcal{H},\ |\widehat{L}(h) - L(h)| \leq \epsilon\right) \geq 1 - \delta(n,\epsilon), \tag{2}$$

Bounding the excess risk under uniform convergence

Theorem. Suppose uniform convergence (1) holds and let $\widehat{h}$ be the hypothesis chosen by ERM. Then, for all $\delta \in (0,1)$, we have

$$\mathbb{P}\left(L(\widehat{h}) < \inf_{h \in \mathcal{H}} L(h) + 2\epsilon(n,\delta) \right) \geq 1 - \delta.$$

Moreover, if the instance loss is bounded, *i.e.*, $\ell(h,(x,y)) \in [0,B]$ for all h, x, y, then

$$\mathbb{E}\left[L(\widehat{h}) - \inf_{h \in \mathcal{H}} L(h) \right] \leq 2\epsilon(n,\delta) + B\delta.$$

Proof

Let δ be given. Let G denote the good event that the empirical loss $\widehat{L}(h)$ of all hypotheses h is within $\epsilon(n, \delta)$ of the true loss $L(h)$. That is,

$$G = \left\{ \forall h \in \mathcal{H}; |\widehat{L}(h) - L(h)| \leq \epsilon(n, \delta) \right\}$$

As uniform convergence holds, we know $\mathbb{P}(G) \geq 1 - \delta$. It is sufficient to show that under G we have $L(\widehat{h}) - \inf_{h \in \mathcal{H}} L(h) \leq 2\epsilon(n, \delta)$.

For this, consider any $h' \in \mathcal{H}$. Then,

$$\begin{aligned} L(\widehat{h}) - L(h') &= L(\widehat{h}) - \widehat{L}(\widehat{h}) + \widehat{L}(\widehat{h}) - \widehat{L}(h') + \widehat{L}(h') - L(h') \\ &\leq L(\widehat{h}) - \widehat{L}(\widehat{h}) + \widehat{L}(h') - L(h') \quad \texttt{As } \widehat{h} \texttt{ minimizes } \widehat{L} \\ &\leq 2 \sup_{h \in \mathcal{H}} |L(h) - \widehat{L}(h)| \leq 2\epsilon(n, \delta). \end{aligned}$$

This is true for all $h' \in \mathcal{H}$. Therefore,

$$L(\widehat{h}) - \inf_{h' \in \mathcal{H}} L(h') \leq 2\epsilon(n, \delta).$$

Proof (cont'd)

To prove the second statement, we can write

$$\mathbb{E}\left[L(\widehat{h}) - \inf_{h\in\mathcal{H}} L(h)\right] = \mathbb{E}\left[L(\widehat{h}) - \inf_{h\in\mathcal{H}} L(h)\middle| G\right]\mathbb{P}(G) + \mathbb{E}\left[L(\widehat{h}) - \inf_{h\in\mathcal{H}} L(h)\middle| G^c\right]\mathbb{P}(G^c)$$
$$\leq 2\epsilon(n,\delta) + B\cdot\delta$$

Learnability

In this course, we will say a class $\mathcal{H}$ is *learnable* if there exists an algorithm $\widehat{h}$ such that $\mathbb{E}\left[L(\widehat{h}) - \inf_{h\in\mathcal{H}} L(h)\right] \to 0$ as $n \to \infty$ for all distributions P over $\mathcal{Z}$. We will say it is *learnable at rate* $f(n)$ if we can achieve $\mathbb{E}\left[L(\widehat{h}) - \inf_{h\in\mathcal{H}} L(h)\right] \in \mathcal{O}(f(n))$.

The following notion of PAC Learnability is also common in learning theory.

PAC Learnability. A hypothesis class $\mathcal{H}$ is (agnostic) PAC-Learnable if there exists a function $N : (0,1)^2 \to \mathbb{N}$ and a learning algorithm with the following property: for all $\epsilon \in (0,1)$, $\delta \in (0,1)$, and for all distributions P over $\mathcal{Z}$, when running the algorithm on $n \geq N(\epsilon,\delta)$ i.i.d samples generated from P, the algorithm returns a hypothesis $\widehat{h} \in \mathcal{H}$ such that with probability at least $1-\delta$, we have $L(\widehat{h}) \leq \inf_{h'\in\mathcal{H}} L(h') + \epsilon$,

- Some definitions of PAC learnability also require that N be at most polynomial in $1/\epsilon$ and $1/\delta$, and sometimes additionally that the algorithm run in $\mathrm{poly}(1/\epsilon, 1/\delta)$ time.

- In this class, we will focus on the first (weaker) definition for simplicity, although in all examples, we will also have PAC Learnability.

Bounding excess risk

Recall, theorem. Suppose uniform convergence (1) holds, and let $\widehat{h}$ be the hypothesis chosen by ERM. Then, for all $\delta \in (0,1)$, we have $\mathbb{P}\left(L(\widehat{h}) < \inf_{h\in\mathcal{H}} L(h) + 2\epsilon(n,\delta)\right) \geq 1-\delta$.

Moreover, if the instance loss is bounded in $[0,1]$, then $\mathbb{E}\left[L(\widehat{h}) - \inf_{h\in\mathcal{H}} L(h)\right] \leq 2\epsilon(n,\delta) + \delta$.

We will strive to achieve bounds that look like the following,

$$\mathbb{E}\left[L(\widehat{h}) - \inf_{h\in\mathcal{H}} L(h)\right] \in \tilde{\mathcal{O}}\left(\sqrt{\frac{\text{complexity}(\mathcal{H})}{n}}\right).$$

For example, this can be achieved if,

$$\delta \in \tilde{\mathcal{O}}\left(\frac{1}{\sqrt{n}}\right), \qquad \epsilon(n,\delta) \in \tilde{\mathcal{O}}\left(\sqrt{\frac{\text{complexity}(\mathcal{H})}{n}}\text{polylog}\left(\frac{1}{\delta}\right)\right)$$

Example 1: Finite hypothesis classes

Assume that the instance loss $\ell(h,(x,y))$ is bounded in $[0,1]$ for all h,x,y. Assume a finite hypothesis class, *i.e.*, $|\mathcal{H}| < \infty$. Using the union bound, and Hoeffding's inequality, we can upper bound the probability of the bad event as follows:

$$\begin{aligned}
\mathbb{P}\left(\exists\, h \in \mathcal{H},\ |\widehat{L}(h) - L(h)| \geq \epsilon\right) &\leq \sum_{h\in\mathcal{H}} \mathbb{P}\left(\ |\widehat{L}(h) - L(h)| \geq \epsilon\right) \\
&= \sum_{h\in\mathcal{H}} \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}(\ell(h,Z_i) - \mathbb{E}[\ell(h,Z)])\right| \geq \epsilon\right) \\
&\leq \sum_{h\in\mathcal{H}} 2\exp\left(\frac{-2\epsilon^2 n^2}{n}\right) = 2|\mathcal{H}|\exp(-2n\epsilon^2).
\end{aligned}$$

Letting $\delta = 2|\mathcal{H}|e^{-2n\epsilon^2}$, and solving for ϵ, we get,

$$\epsilon(n,\delta) = \sqrt{\frac{1}{2n}\log\left(\frac{2|\mathcal{H}|}{\delta}\right)}.$$

Example 1: Finite hypothesis classes (cont'd)

Recall, Theorem. Suppose uniform convergence (1) holds and let $\widehat{h}$ be the hypothesis chosen by ERM. If the instance loss is bounded in $[0, B]$, then, for all $\delta \in (0, 1)$, we have

$$\mathbb{E}\left[L(\widehat{h}) - \inf_{h \in \mathcal{H}} L(h)\right] \leq 2\epsilon(n, \delta) + B\delta.$$

By this theorem, we have

$$\mathbb{E}\left[L(\widehat{h}) - \inf_{h \in \mathcal{H}} L(h)\right] \leq 2\sqrt{\frac{1}{2n} \log\left(\frac{2|\mathcal{H}|}{\delta}\right)} + \delta.$$

By choosing $\delta = 1/\sqrt{n}$ we get,

$$\mathbb{E}\left[L(\widehat{h}) - \inf_{h \in \mathcal{H}} L(h)\right] \leq \sqrt{\frac{2}{n} \log\left(2|\mathcal{H}|\sqrt{n}\right)} + \frac{1}{\sqrt{n}} \in \tilde{\mathcal{O}}\left(\sqrt{\frac{\log(|\mathcal{H}|)}{n}}\right)$$

Here, the complexity parameter (for finite $\mathcal{H}$) is $\log(|\mathcal{H}|)$.

Example 2: Linear regression

Consider the following hypothesis class for regression,

$$\mathcal{H} = \left\{ h_\theta; h_\theta(x) = \theta^\top x; \theta \in \mathcal{B}_2(B_\theta) \right\}$$

Moreover, let us assume that the joint distribution P is such that for all $(X, Y) \sim P$, we have $\|X\|_2 \leq B_x, |Y| \leq B_y$ a.s.

Consider any $\theta, \theta' \in \mathcal{B}_2(B_\theta)$. We have,

$$\begin{aligned}
\ell(h_\theta, (x, y)) - \ell(h_{\theta'}, (x, y)) &= (\theta^\top x - y)^2 - (\theta'^\top x - y)^2 \\
&= (\theta^\top x)^2 - (\theta'^\top x)^2 - 2y(\theta^\top x - \theta'^\top x) \\
&= x^\top(\theta + \theta') \cdot x^\top(\theta - \theta') - 2yx^\top(\theta - \theta').
\end{aligned}$$

Therefore,

$$\begin{aligned}
|\ell(h_\theta, (x, y)) - \ell(h_{\theta'}, (x, y))| &\leq \left(\|x\|_2^2 \|\theta + \theta'\|_2 + 2|y| \|x\|_2 \right) \|\theta - \theta'\|_2 \\
&\leq \underbrace{\left(2B_x^2 B_\theta + 2B_y B_x \right)}_{K} \|\theta - \theta'\|_2
\end{aligned}$$

The loss $\ell(h_\theta, (x, y))$ is K-Lipschitz in θ for all x, y.

Example 2: Linear regression (cont'd)

The remainder of this example works for instance losses that are K-Lipschitz, *i.e.*,

$$\text{for all } z \in \mathcal{Z}, \quad |\ell(h_\theta, z) - \ell(h_{\theta'}, z)| \leq K\|\theta - \theta'\|_2.$$

We cannot apply a union bound over all $\theta \in \mathcal{B}_2(B_\theta)$. Instead, we will

- Apply the union bound over a finite cover of $\mathcal{B}_2(B_\theta)$.
- Use the Lipschitz property of the instance loss to show that both the population loss and empirical loss of h_θ are smooth, and therefore close to h_θ's "representative" in the cover.
- Combining the two results above, we will establish uniform convergence.

Example 2: Linear regression (cont'd)

Recall the following definitions and theorem from Chapter 0.

Covering number. Let $(\mathcal{X}, \rho)$ be a pseudo-metric space and let $A \subset \mathcal{X}$. Let $\epsilon > 0$. A set $C \subset A$ is called an ϵ-**cover** of A if, for all $x \in A$, there exists $c \in C$ such that $\rho(x, c) \leq \epsilon$.
- The ϵ-**covering number** $N(\epsilon, A, \rho)$ is the size of the smallest ϵ-cover of A.

Theorem. Let $\mathcal{X} = \mathbb{R}^d$ and let $\|\cdot\|$ be any norm. Let $B = \{x \in \mathbb{R}^d; \|x\| \leq 1\}$ be the unit ball. Suppose A is a convex set and contains ϵB. Then,

$$\left(\frac{1}{\epsilon}\right)^d \frac{\mathrm{vol}(A)}{\mathrm{vol}(B)} \leq N(\epsilon, A, \|\cdot\|) \leq M(\epsilon, A, \|\cdot\|) \leq \frac{\mathrm{vol}\left(A + \frac{\epsilon}{2}B\right)}{\mathrm{vol}\left(\frac{\epsilon}{2}B\right)} \leq \left(\frac{3}{\epsilon}\right)^d \frac{\mathrm{vol}(A)}{\mathrm{vol}(B)}.$$

Let $C_{\epsilon'}$ be a minimal ϵ'-cover of $\mathcal{B}_2(B_\theta)$. For any $\theta \in \mathcal{B}_2(B_\theta)$, let $[\theta] \in C_{\epsilon'}$ be the "representative" of θ such that $\|\theta - [\theta]\|_2 \leq \epsilon'$.

We know $|C_{\epsilon'}| \leq \left(\frac{3}{\epsilon'}\right)^d \frac{\mathrm{vol}(\mathcal{B}_2(B_\theta))}{\mathrm{vol}(\mathcal{B}_2(1))} = \left(\frac{3B_\theta}{\epsilon'}\right)^d$.

Example 2: Linear regression (cont'd)

Recall, our goal is to show uniform convergence. That is, for all $\theta \in \mathcal{B}_2(B_\theta)$, we wish to bound $\widehat{L}(h_\theta) - L(h_\theta)$.

Let us first decompose,

$$\widehat{L}(h_\theta) - L(h_\theta) = \underbrace{\widehat{L}(h_\theta) - \widehat{L}(h_{[\theta]})}_{\text{(i)}} + \underbrace{\widehat{L}(h_{[\theta]}) - L(h_{[\theta]})}_{\text{(ii)}} + \underbrace{L(h_{[\theta]}) - L(h_\theta)}_{\text{(iii)}}.$$

To bound (i) we will use the the Lipschitz property:

$$\begin{aligned} |\text{(i)}| &= \left| \frac{1}{n} \sum_{i=1}^{n} \left(\ell(h_\theta, Z_i) - \ell(h_{[\theta]}, Z_i) \right) \right| \\ &\leq \frac{1}{n} \sum_{i=1}^{n} \left| \ell(h_\theta, Z_i) - \ell(h_{[\theta]}, Z_i) \right| \leq \frac{1}{n} \sum_{i=1}^{n} K \|\theta - [\theta]\|_2 \leq K\epsilon'. \end{aligned}$$

You can similarly show $|\text{(iii)}| \leq K\epsilon'$.

Example 2: Linear regression (cont'd)

We therefore have,

$$\left|\widehat{L}(h_\theta) - L(h_\theta)\right| \leq |\underbrace{\widehat{L}(h_\theta) - \widehat{L}(h_{[\theta]})}_{\text{(i)}}| + |\underbrace{\widehat{L}(h_{[\theta]}) - L(h_{[\theta]})}_{\text{(ii)}}| + |\underbrace{L(h_{[\theta]}) - L(h_\theta)}_{\text{(iii)}}|$$
$$\leq 2K\epsilon' + |\underbrace{\widehat{L}(h_{[\theta]}) - L(h_{[\theta]})}_{\text{(ii)}}|.$$

Let $\epsilon > 0$. If we set $\epsilon' = \epsilon/(4K)$ and show $|(ii)| \leq \epsilon/2$ via the union bound, we will have $|\widehat{L}(h_\theta) - L(h_\theta)| \leq \epsilon$.

Let G be the good event where all h_θ in the cover have empirical loss close to the true loss, *i.e.*, $G = \left\{\forall\, \theta' \in C_{\epsilon'}; |\widehat{L}(h_{\theta'}) - L(h_{\theta'})| \leq \frac{\epsilon}{2}\right\}$. By the union bound, Hoeffding's, and similar calculations to example 1, we have

$$\mathbb{P}(G^c) \leq \sum_{\theta' \in C_{\epsilon'}} \mathbb{P}\left(|\widehat{L}(h_{\theta'}) - L(h_{\theta'})| > \frac{\epsilon}{2}\right) \leq |C_{\epsilon'}| \cdot 2e^{-n\epsilon^2/2}.$$

Example 2: Linear regression (cont'd)

By setting $\epsilon' = \epsilon/(4K)$, with probability at least $1 - 2|C_{\epsilon/4K}|e^{-n\epsilon^2/2}$, we have,

$$\forall\, \theta \in \mathcal{B}_2(B_\theta), \quad \left|\widehat{L}(h_\theta) - L(h_\theta)\right| \leq \epsilon.$$

Recall, we wish to achieve $\mathbb{E}\left[L(\widehat{h}) - \inf_{h\in\mathcal{H}} L(h)\right] \in \tilde{\mathcal{O}}\left(\sqrt{\frac{\text{complexity}(\mathcal{H})}{n}}\right)$. Let us try the following ϵ with a, b to be chosen later,

$$\epsilon = \sqrt{\frac{2}{n}\log\left(an^b\right)}$$

Hence, recalling that, $|C_{\epsilon'}| \leq \left(\frac{3B_\theta}{\epsilon'}\right)^d$, the probability of the bad event is,

$$\delta = 2|C_{\epsilon/4K}|e^{-n\epsilon^2/2} \leq 2\left(\frac{3B_\theta}{(\epsilon/4K)}\right)^d e^{-n\epsilon^2/2} = 2\left(\frac{12B_\theta K\cdot\sqrt{n}}{\sqrt{2\log(an^b)}}\right)^d \cdot \frac{1}{an^b}$$

Example 2: Linear regression (cont'd)

Recall, Theorem. Suppose uniform convergence holds, *i.e.*, $\mathbb{P}\left(\forall h \in \mathcal{H},\ |\widehat{L}(h) - L(h)| \leq \epsilon\right) \geq 1 - \delta$. Then, if the instance loss is bounded in $[0, B]$, for ERM, we have $\mathbb{E}\left[L(\widehat{h}) - \inf_{h \in \mathcal{H}} L(h)\right] \leq 2\epsilon + B\delta$.

Now, let us choose $a = 2(6\sqrt{2}B_\theta K)^d$ and $b = (d+1)/2$, we then have

$$\delta \leq \frac{1}{\sqrt{n}} \cdot \frac{1}{\sqrt{\log(an^b)}}.$$

If the instance loss is bounded in $[0, B]$, we have

$$\mathbb{E}\left[L(\widehat{h}) - \inf_{h \in \mathcal{H}} L(h)\right] \leq \sqrt{\frac{2}{n}\log\left(2(6\sqrt{2}B_\theta K)^d n^{(d+1)/2}\right)} + \frac{B}{\sqrt{n}} \cdot \frac{1}{\sqrt{\log(an^b)}}$$
$$\in \tilde{\mathcal{O}}\left(\sqrt{\frac{d}{n}}\right)$$

N.B. Here, the complexity of the class is the dimensionality d.

Summary so far

- Uniform convergence implies a hypothesis class $\mathcal{H}$ is learnable via ERM, *i.e.*, we can achieve $\mathbb{E}\left[L(\widehat{h}) - \inf_{h\in\mathcal{H}} L(h)\right] \to 0$ as $n \to \infty$.
- We can prove uniform convergence using *pointwise* covers if the loss is smooth.
  Recall, for the linear regression example, we showed for all x, y,
  $|\ell(h_\theta, (x, y)) - \ell(h_{\theta'}, (x, y))| \leq K\|\theta - \theta'\|_2$.
- However, this approach has limitations:
  1. The Lipschitz-constant may be very large for complex models (*e.g.*, multi-layer neural networks), and it may not be meaningful to treat it as a constant. In practice, models may perform much better than such bounds would suggest.
  2. This approach does not work for "non-smooth" losses, *e.g.*, $0-1$ loss for classification. You will explore this in HW2.

Plan for the next few lectures

- Ch 3.2: Rademacher complexity
  - Definition
  - Properties: basic properties, Massart's lemma, Contraction lemma
  - Bounded Rademacher complexity implies uniform convergence.
- Ch 3.3 onwards: Bounding Rademacher compleixty
  - VC dimension and Sauer's lemma for binary classification.
  - Using finite covers
  - Dudley entropy integral

Ch 3.2: Rademacher complexity

Rademacher random variables. $\sigma \in \{+1, -1\}$, where $\mathbb{P}(\sigma = +1) = \mathbb{P}(\sigma = -1) = 1/2$.

Empirical Rademacher complexity of a Euclidean subset. Let $A \subset \mathbb{R}^n$. Let $\sigma = (\sigma_1, \ldots, \sigma_n) \in \{-1, +1\}^n$ be n independent Rademacher random variables. Denote,

$$\widehat{\mathcal{R}}(A) \triangleq \mathbb{E}_\sigma \left[\sup_{a \in A} \frac{1}{n} \sigma^\top a \right].$$

Empirical Rademacher complexity. Let $\mathcal{Z}$ be a set and let $\mathcal{F} \subset \mathbb{R}^{\mathcal{Z}}$. Let $S = \{z_1, \ldots, z_n\} \subset \mathcal{Z}$ and $\sigma = (\sigma_1, \ldots, \sigma_n) \in \{-1, +1\}^n$ be n independent Rademacher random variables. Denote $\mathcal{F}(S) = \{[f(z_1), \ldots, f(z_n)]; f \in \mathcal{F}\} \subset \mathbb{R}^n$. Then the empirical Rademacher complexity of $\mathcal{F}$ for set S is,

$$\widehat{\mathcal{R}}(\mathcal{F}, S) \triangleq \widehat{\mathcal{R}}(\mathcal{F}(S)) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(z_i) \right].$$

Rademacher Complexity

Intuition: $\widehat{\mathcal{R}}(\mathcal{F}, S)$ measures how well the functions in $\mathcal{F}$ can correlate with random noise,

$$\widehat{\mathcal{R}}(\mathcal{F}, S) = \mathbb{E}_\sigma \left[\frac{1}{n} \sup_{\bar{f} \in \mathcal{F}(S)} \sigma^\top \bar{f} \right].$$

Rademacher complexity. The Rademacher complexity of a function class $\mathcal{F} \subset \mathbb{R}^{\mathcal{Z}}$ with respect to a distribution P with $\mathbf{supp}(P) \subset \mathcal{Z}$ is,

$$\mathcal{R}_n(\mathcal{F}, P) = \mathbb{E}_{S \sim P^n} \left[\widehat{\mathcal{R}}(\mathcal{F}, S) \right] = \mathbb{E}_{S \sim P^n} \left[\mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(z_i) \right] \right].$$

If P is clear from context, we may drop it.

Properties of Rademacher complexity

1) Let $A \subset B \subset \mathbb{R}^n$. Then $\widehat{\mathcal{R}}(A) \leq \widehat{\mathcal{R}}(B)$.
Hence, if $\mathcal{F} \subset \mathcal{G} \subset \mathbb{R}^{\mathcal{Z}}$, we have $\mathcal{R}_n(\mathcal{F}) \leq \mathcal{R}_n(\mathcal{G})$.

2) Let $\alpha \in \mathbb{R}$, and let $a \in \mathbb{R}^n$. Let $A \subset \mathbb{R}^n$. Then, $\widehat{\mathcal{R}}(\alpha A + a) = |\alpha| \widehat{\mathcal{R}}(A)$.
Hence, for any $\mathcal{F} \subset \mathbb{R}^{\mathcal{Z}}$ and $f \in \mathbb{R}^{\mathcal{Z}}$, we have $\mathcal{R}_n(\alpha \mathcal{F} + f) = |\alpha| \mathcal{R}_n(\mathcal{F})$.

3) Let $A, B \subset \mathbb{R}^n$. Then, $\widehat{\mathcal{R}}(A + B) = \widehat{\mathcal{R}}(A) + \widehat{\mathcal{R}}(B)$.
Hence, for any $\mathcal{F}, \mathcal{G} \subset \mathbb{R}^{\mathcal{Z}}$, we have $\mathcal{R}_n(\mathcal{F} + \mathcal{G}) = \mathcal{R}_n(\mathcal{F}) + \mathcal{R}_n(\mathcal{G})$.

4) Let $A \subset \mathbb{R}^n$. Let $|A| = \{|a|; a \in A\}$, where $|a| = \{|a_1|, \ldots, |a_n|\}$ is the pointwise absolute value. Then, $\widehat{\mathcal{R}}(|A|) \leq \widehat{\mathcal{R}}(A)$.
Hence, for any $\mathcal{F} \subset \mathbb{R}^{\mathcal{Z}}$, we have $\mathcal{R}_n(|\mathcal{F}|) \leq \mathcal{R}_n(\mathcal{F})$ where $|\mathcal{F}| = \{|f|; f \in \mathcal{F}\}$ and $|f|(x) = |f(x)|$.

(Proof: HW2)

Massart's lemma

Massart's Lemma. Let $A = \{a_1, \ldots, a_N\} \subset \mathbb{R}^n$ be a finite subset of $\mathbb{R}^n$. Then,

$$\widehat{\mathcal{R}}(A) \leq \left(\max_{a \in A} \|a\|_2 \right) \cdot \frac{\sqrt{2 \log(N)}}{n}$$

Hence, for a bounded function class $\mathcal{F} \subset [-B, B]^{\mathcal{Z}}$ with a finite number of elements, we have

$$\mathcal{R}_n(\mathcal{F}) \leq B \sqrt{\frac{2 \log(|\mathcal{F}|)}{n}}.$$

Recall, Maximal inequality. Let $Z_1, \ldots, Z_N$ be zero mean γ-sub-Gaussian random variables (not necessarily independent). Then, $\mathbb{E}\left[\max_{i \in [N]} Z_i\right] \leq \gamma \sqrt{2 \log(N)}$.

Massart's lemma (cont'd)

Proof. We know that a Rademacher RV is 1-sub-Gaussian as it is bounded in $[-1, 1]$. Hence, $\sigma^\top a$ is $\|a\|_2$-sub-Gaussian. Therefore, for every $a \in A$, we have $\sigma^\top a$ is $(\max_{a \in A} \|a\|_2)$-sub-Gaussian. Therefore, by the maximal inequality,

$$\widehat{\mathcal{R}}(A) = \frac{1}{n}\mathbb{E}_\sigma \left[\max_{a \in \{a_1, \ldots, a_N\}} \sigma^\top a \right] \leq \left(\max_{a \in A} \|a\|_2 \right) \frac{\sqrt{2 \log(N)}}{n}.$$

Now let $S \subset \mathcal{Z}$. We have,

$$\widehat{\mathcal{R}}(\mathcal{F}, S) = \widehat{\mathcal{R}}(\mathcal{F}(S)) \leq \left(\max_{a \in \mathcal{F}(S)} \|a\|_2 \right) \frac{\sqrt{2 \log(|\mathcal{F}(S)|)}}{n} \leq B \sqrt{\frac{2 \log(|\mathcal{F}|)}{n}}.$$

The last inequality follows from the observation that for any $a \in \mathcal{F}(S) \subset [-B, B]^n$, we have $\|a\|_2 \leq B\sqrt{n}$. Moreover, $|\mathcal{F}(S)| \leq |\mathcal{F}|$.

Contraction lemma

Talagrand's contraction lemma. Let $\phi_1, \ldots, \phi_n$, where $\phi_i : \mathbb{R} \to \mathbb{R}$, be n K-Lipschitz functions, *i.e.*, $|\phi_i(a) - \phi_i(a')| \leq K|a - a'|$ for all i. For any $A \subset \mathbb{R}^n$, let $\phi(A) = \{[\phi_1(a_1), \ldots, \phi_n(a_n)]; a \in A\}$. Then, $\widehat{\mathcal{R}}(\phi(A)) \leq K\widehat{\mathcal{R}}(A)$.

Corollary. Let $\mathcal{F} \subset \mathbb{R}^{\mathcal{Z}}$. For any K-Lipschitz $\psi : \mathbb{R} \to \mathbb{R}$, let $\psi \circ \mathcal{F} = \{\psi \circ f; f \in \mathcal{F}\}$. Then, $\mathcal{R}_n(\psi \circ \mathcal{F}) \leq K\mathcal{R}_n(\mathcal{F})$.

Proof. First, we show that the corollary follows immediately from the first result.

$$\mathcal{R}_n(\psi \circ \mathcal{F}) = \mathbb{E}_S[\widehat{\mathcal{R}}(\psi(\mathcal{F}(S))] \leq \mathbb{E}_S[K \cdot \widehat{\mathcal{R}}(\mathcal{F}(S))] = K\mathcal{R}_n(\mathcal{F}).$$

In the inequality, we have applied the first result with $\phi_i = \psi$.

Contraction lemma (cont'd)

We will next prove the first result. We have,

$$
\begin{aligned}
n\widehat{\mathcal{R}}(\phi(A)) &= \mathbb{E}_\sigma\left[\sup_{a\in A}\sigma^\top\phi(a)\right] \\
&= \mathbb{E}_\sigma\left[\sup_{a\in A}\left(\sigma_1\phi_1(a_1)+\sum_{i=2}^{n}\sigma_i\phi_i(a_i)\right)\right] \\
&= \frac{1}{2}\mathbb{E}_{\sigma_{2:n}}\left[\sup_{a\in A}\left(\phi_1(a_1)+\sum_{i=2}^{n}\sigma_i\phi_i(a_i)\right)\right] \qquad \text{As } \sigma_1 \text{ is a Rademacher RV.} \\
&\qquad + \frac{1}{2}\mathbb{E}_{\sigma_{2:n}}\left[\sup_{a\in A}\left(-\phi_1(a_1)+\sum_{i=2}^{n}\sigma_i\phi_i(a_i)\right)\right] \\
&= \frac{1}{2}\mathbb{E}_{\sigma_{2:n}}\left[\sup_{a,a'\in A}\left((\phi_1(a_1)-\phi(a'_1)+\sum_{i=2}^{n}\sigma_i\phi_i(a_i)+\sum_{i=2}^{n}\sigma_i\phi_i(a'_i)\right)\right]
\end{aligned}
$$

Contraction lemma (cont'd)

Using K-Lipschitzness of ϕ we have,

$$n\widehat{\mathcal{R}}(\phi(A)) \leq \frac{1}{2}\mathbb{E}_{\sigma_{2:n}}\left[\sup_{a,a'\in A}\left(K|a_1 - a_1'| + \sum_{i=2}^{n}\sigma_i\phi_i(a_i) + \sum_{i=2}^{n}\sigma_i\phi_i(a_i')\right)\right]$$

As we are taking the supremum, we can drop the absolute value. Therefore,

$$\begin{aligned} n\widehat{\mathcal{R}}(\phi(A)) &\leq \frac{1}{2}\mathbb{E}_{\sigma_{2:n}}\left[\sup_{a,a'\in A}\left(K(a_1 - a_1') + \sum_{i=2}^{n}\sigma_i\phi_i(a_i) + \sum_{i=2}^{n}\sigma_i\phi_i(a_i')\right)\right] \\ &= \frac{1}{2}\mathbb{E}_{\sigma_{2:n}}\left[\sup_{a\in A}\left(Ka_1 + \sum_{i=2}^{n}\sigma_i\phi_i(a_i)\right)\right] \\ &\qquad + \frac{1}{2}\mathbb{E}_{\sigma_{2:n}}\left[\sup_{a'\in A}\left(-Ka_1' + \sum_{i=2}^{n}\sigma_i\phi_i(a_i')\right)\right] \end{aligned}$$

Contraction lemma (cont'd)

We can now combine the two summations once again to write,

$$n\widehat{\mathcal{R}}(\phi(A)) \leq \mathbb{E}_\sigma \left[\sup_{a \in A} \left(K\sigma_1 a_1 + \sum_{i=2}^{n} \sigma_i \phi_i(a_i) \right) \right]$$

Repeating this argument for $i = 2, \ldots, n$ results in,

$$n\widehat{\mathcal{R}}(\phi(A)) \leq \mathbb{E}_\sigma \left[\sup_{a \in A} \left(K \sum_{i=1}^{n} \sigma_i a_i \right) \right] = nK\widehat{\mathcal{R}}(A).$$

Towards uniform convergence: symmetrization

The theorem below will allow us to bound the maximum possible deviation of the empirical losses from their expected value in terms of the Rademacher complexity.

Theorem. Let $\mathcal{F} \subset \mathbb{R}^{\mathcal{Z}}$. Let $S = \{Z_1, \ldots, Z_n\} \subset \mathcal{Z}$ be an i.i.d sample from some distribution P. Then,

$$\mathbb{E}_{S \sim P^n}\left[\sup_{f \in \mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^{n} f(Z_i) - \mathbb{E}_{Z \sim P}[f(Z)]\right)\right] \leq 2\mathcal{R}_n(\mathcal{F}),$$

$$\mathbb{E}_{S \sim P^n}\left[\sup_{f \in \mathcal{F}}\left(\mathbb{E}_{Z \sim P}[f(Z)] - \frac{1}{n}\sum_{i=1}^{n} f(Z_i)\right)\right] \leq 2\mathcal{R}_n(\mathcal{F}).$$

Proof. We will prove the first statement. Letting $S' = \{Z_1, \ldots, Z_n\} \subset \mathcal{Z}$ be (another) i.i.d sample, we have

$$\text{LHS} = \mathbb{E}_{S \sim P^n}\left[\sup_{f \in \mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^{n} f(Z_i) - \mathbb{E}_{S' \sim P^n}\left[\frac{1}{n}\sum_{i=1}^{n} f(Z_i')\right]\right)\right]$$

Symmetrization (cont'd)

Using the fact that sup is sub-additive, we have

$$\begin{aligned}\text{LHS} &\leq \mathbb{E}_{S,S'}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^{n} f(Z_i) - \frac{1}{n}\sum_{i=1}^{n} f(Z_i')\right)\right] \\ &= \frac{1}{n}\mathbb{E}_{S,S',\sigma}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\left(\sigma_i f(Z_i) - \sigma_i f(Z_i')\right)\right]\end{aligned}$$

In the previous step, we have introduced n Rademacher random variables $\sigma = (\sigma_1, \ldots, \sigma_n)$. Equality follows from symmetry of the two datasets S, S': as we are taking expectation over S and S', we have equality for every realization of σ.

Symmetrization (cont'd)

Now continuing,

$$\begin{aligned}
\text{LHS} &= \mathbb{E}_{S,S',\sigma}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(Z_i)+\frac{1}{n}\sum_{i=1}^{n}(-\sigma_i)f(Z_i')\right] \\
&\le \mathbb{E}_{S,S',\sigma}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(Z_i)+\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}(-\sigma_i)f(Z_i')\right] \\
&= \mathbb{E}_{S,\sigma}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(Z_i)\right]+\mathbb{E}_{S',\sigma}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}(-\sigma_i)f(Z_i')\right] \\
&= 2\mathcal{R}_n(\mathcal{F}).
\end{aligned}$$

Second statement. The first result implies a bound of $2\mathcal{R}_n(-\mathcal{F})$ on the LHS of the second inequality. The claim follows from the fact that $\mathcal{R}_n(-\mathcal{F}) = \mathcal{R}_n(\mathcal{F})$.

Uniform convergence via Rademacher complexity

Recall the following definitions. Instance loss $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$, where $\ell(h, z)$ is the loss of hypothesis h on instance z. For a distribution P, let $L(h) = \mathbb{E}_{Z \sim P}[\ell(h, Z)]$. For a dataset $S = \{Z_1, \ldots, Z_n\}$, let $\widehat{L}(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(h, Z_i)$.

Theorem. Let $\mathcal{H}$ be a hypothesis class and assume the loss is bounded in $[0, 1]$. Let $\mathcal{F} \triangleq \{\ell(h, \cdot); h \in \mathcal{H}\} \subset \mathbb{R}^{\mathcal{Z}}$ be the class of real-valued functions induced by applying ℓ to hypotheses $h \in \mathcal{H}$. Then, with probability at least $1 - 2e^{-2n\epsilon^2}$ we have,

$$\forall\ h \in \mathcal{H}, \quad |\widehat{L}(h) - L(h)| \leq 2\mathcal{R}_n(\mathcal{F}) + \epsilon.$$

Corollary. We can equivalently state the above result as the following:

$$\mathbb{P}\left(\forall\ h \in \mathcal{H}, \quad |\widehat{L}(h) - L(h)| \leq \epsilon(n, \delta), \right) \geq 1 - \delta,$$

where $\epsilon(n, \delta) = 2\mathcal{R}_n(\mathcal{F}) + \sqrt{\frac{1}{2n} \log\left(\frac{2}{\delta}\right)}$.

Uniform convergence via Rademacher complexity (cont'd)

Recall, McDiarmid's inequality. Let $X_1, \ldots, X_n$ be independent random variables with $X_i \in \mathcal{X}_i$. Let $f : \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \to \mathbb{R}$. Suppose there exists $\{c_i\}_{i=1}^n$ such that for all $x_1, \ldots, x_i, x_i', \ldots, x_n$, we have, $|f(x_1, \ldots, x_i, \ldots, x_n) - f(x_1, \ldots, x_i', \ldots, x_n)| \leq c_i$. Then, for all $t > 0$,

$$\Pr\left(f(X_1, \ldots, X_n) - \mathbb{E}[f(X_1, \ldots, X_n)] \geq t\right) \leq \exp\left(-\frac{2\,t^2}{\sum_{i=1}^n c_i^2}\right),$$

$$\Pr\left(f(X_1, \ldots, X_n) - \mathbb{E}[f(X_1, \ldots, X_n)] \leq -t\right) \leq \exp\left(-\frac{2\,t^2}{\sum_{i=1}^n c_i^2}\right).$$

Proof. Let us make the dependence of S on $\widehat{L}$ clear and write

$$\widehat{L}(h, S) = \frac{1}{n}\sum_{i=1}^{n} \ell(h, Z_i).$$

Now define, $g_1(S) = \sup_{h \in \mathcal{H}} \left(\widehat{L}(h, S) - L(h)\right)$, $g_2(S) = \sup_{h \in \mathcal{H}} \left(L(h) - \widehat{L}(h, S)\right)$..
We will apply McDiarmid's inequality to bound g_1 and g_2. For this denote,

$$S = \{Z_1, \ldots, Z_i, \ldots, Z_n\}, \quad \widetilde{S} = \{Z_1, \ldots, \widetilde{Z}_i, \ldots, Z_n\}, \quad S \cup \widetilde{S} = \{Z_1, \ldots, Z_i, \widetilde{Z}_i, \ldots, Z_n\},$$

Uniform convergence via Rademacher complexity (cont'd)

Let us verify the bounded difference property for $g_1(S) = \sup_{h\in\mathcal{H}} \left(\widehat{L}(h,S) - L(h)\right)$.

$$\begin{aligned}
\sup_{S\cup\widetilde{S}} |g_1(S) - g_1(\widetilde{S})| &= \sup_{S\cup\widetilde{S}} \left| \sup_{h\in\mathcal{H}} \left(\widehat{L}(h,S) - L(h)\right) - \sup_{h\in\mathcal{H}} \left(\widehat{L}(h,\widetilde{S}) - L(h)\right) \right| \\
&\leq \sup_{S\cup\widetilde{S}} \sup_{h\in\mathcal{H}} \left|\widehat{L}(h,S) - \widehat{L}(h,\widetilde{S})\right| \\
&\qquad \text{as} \quad \left|\sup_a f_1(a) - \sup_a f_2(a)\right| \leq \sup_a |f_1(a) - f_2(a)|. \\
&= \sup_{Z_i,\widetilde{Z}_i} \sup_{h\in\mathcal{H}} |\frac{1}{n}\left(\ell(h,Z_i) - \ell(h,\widetilde{Z}_i)\right)| \\
&\leq \frac{1}{n} \qquad \texttt{As losses are bounded in } [0,1].
\end{aligned}$$

Applying McDiarmdid's with $c_i = 1/n$ for all i, with probability at least $1 - e^{-2n\epsilon^2}$ we have, $g_1(S) - \mathbb{E}[g_1(S)] \leq \epsilon$.

Uniform convergence via Rademacher complexity (cont'd)

Recall that $g_1(S) = \sup_{h\in\mathcal{H}} \left(\widehat{L}(h,S) - L(h)\right)$. Therefore, with probability at least $1 - e^{-2n\epsilon^2}$, we have

$$\sup_{h\in\mathcal{H}} \left(\widehat{L}(h,S) - L(h)\right) \leq \mathbb{E}_S \left[\sup_{h\in\mathcal{H}} \left(\widehat{L}(h,S) - L(h)\right)\right] + \epsilon$$

Hence, for any $h \in \mathcal{H}$, we have

$$\widehat{L}(h,S) - L(h) \leq \mathbb{E}_S \left[\sup_{h\in\mathcal{H}} \left(\frac{1}{n}\sum_{i=1}^{n} \ell(h,Z_i) - \mathbb{E}_S[\ell(h,Z)]\right)\right] + \epsilon$$

Recall, we defined $\mathcal{F} \triangleq \{f(\cdot) = \ell(h,\cdot); h \in \mathcal{H}\}$ in the theorem statement. Hence,

$$\widehat{L}(h,S) - L(h) \leq \mathbb{E}_S \left[\sup_{f\in\mathcal{F}} \left(\frac{1}{n}\sum_{i=1}^{n} f(Z_i) - \mathbb{E}_S[f(Z)]\right)\right] + \epsilon.$$

Uniform convergence via Rademacher complexity (cont'd)

Symmetrization theorem. Let $\mathcal{F} \subset \mathbb{R}^{\mathcal{Z}}$. Let $S = \{Z_1, \ldots, Z_n\} \subset \mathcal{Z}$ be an i.i.d sample from some distribution P. Then,

$$\mathbb{E}_S \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^{n} f(Z_i) - \mathbb{E}_Z[f(Z)] \right) \right] \leq 2\mathcal{R}_n(\mathcal{F}), \quad \mathbb{E}_S \left[\mathbb{E}_Z[f(Z)] - \sup_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^{n} f(Z_i) \right) \right] \leq 2\mathcal{R}_n(\mathcal{F}).$$

By applying the symmetrization theorem we have, with probability at least $1 - e^{-2n\epsilon^2}$,

$$\forall h \in \mathcal{H}, \quad \widehat{L}(h, S) - L(h) \leq 2\mathcal{R}_n(\mathcal{F}) + \epsilon.$$

By applying McDiarmid's for g_2, we can also show, with probability at least $1 - e^{-2n\epsilon^2}$,

$$\forall h \in \mathcal{H}, \quad L(h) - \widehat{L}(h, S) \leq 2\mathcal{R}_n(\mathcal{F}) + \epsilon.$$

Hence, with probability at least $1 - 2e^{-2n\epsilon^2}$, we have

$$\forall h \in \mathcal{H}, \quad |\widehat{L}(h, S) - L(h)| \leq 2\mathcal{R}_n(\mathcal{F}) + \epsilon.$$

Example: Finite $\mathcal{H}$ revisited

Let us revisit an example from before. Assume that the instance loss $\ell(h,(x,y))$ is bounded in $[0,1]$ for all h,x,y. Assume a finite hypothesis class, *i.e.*, $|\mathcal{H}| < \infty$.

We showed, via a union bound, $\mathbb{E}\left[L(\widehat{h}) - \inf_{h\in\mathcal{H}} L(h)\right] \in \tilde{\mathcal{O}}\left(\sqrt{\frac{\log(|\mathcal{H}|)}{n}}\right)$.

What does the new bound based on Rademacher complexity give us?

Recall, Theorem (uniform convergence). Suppose uniform convergence (1) holds and let $\widehat{h}$ be the hypothesis chosen by ERM. If the instance loss is bounded in $[0,B]$, then, for all $\delta \in (0,1)$, we have

$$\mathbb{E}\left[L(\widehat{h}) - \inf_{h\in\mathcal{H}} L(h)\right] \leq 2\epsilon(n,\delta) + B\delta.$$

Recall, Theorem (Rademacher complexity). Let $\mathcal{F} = \{\ell(h,\cdot); h \in \mathcal{H}\}$. Then,

$$\mathbb{P}\left(\forall\, h \in \mathcal{H},\ |\widehat{L}(h) - L(h)| \leq \epsilon(n,\delta)\right) \geq 1-\delta, \qquad \text{where, } \epsilon(n,\delta) = 2\mathcal{R}_n(\mathcal{F}) + \sqrt{\frac{1}{2n}\log\left(\frac{2}{\delta}\right)}$$

Example: Finite $\mathcal{H}$ revisited

Recall, Massart's Lemma. For a bounded function class $\mathcal{F} \subset [-B, B]^{\mathcal{Z}}$ with a finite number of elements, we have $\mathcal{R}_n(\mathcal{F}) \leq B\sqrt{\frac{2\log(|\mathcal{F}|)}{n}}$.

From the "Rademacher complexity theorem", we know uniform convergence holds with $\epsilon(n, \delta) = 2\mathcal{R}_n(\mathcal{F}) + \sqrt{\frac{1}{2n}\log\left(\frac{2}{\delta}\right)}$. Hence, by the "uniform convergence theorem",

$$\begin{aligned}\mathbb{E}\left[L(\widehat{h}) - \inf_{h\in\mathcal{H}} L(h)\right] &\leq 2\epsilon(n,\delta) + \delta \leq 4\mathcal{R}_n(\mathcal{F}) + 2\sqrt{\frac{1}{2n}\log\left(\frac{2}{\delta}\right)} + \delta \\ &\leq 4\sqrt{\frac{2\log(|\mathcal{F}|)}{n}} + 2\sqrt{\frac{1}{2n}\log\left(\frac{2}{\delta}\right)} + \delta.\end{aligned}$$

As $\mathcal{H}$ is finite, $\mathcal{F} = \{\ell(h, \cdot); h \in \mathcal{H}\}$ is also finite with $|\mathcal{F}| \leq |\mathcal{H}|$. Hence, choosing $\delta = 1/\sqrt{n}$, we get

$$\mathbb{E}\left[L(\widehat{h}) - \inf_{h\in\mathcal{H}} L(h)\right] \in \tilde{\mathcal{O}}\left(\sqrt{\frac{\log(|\mathcal{H}|)}{n}}\right).$$

From $\mathcal{R}_n(\mathcal{F})$ to $\mathcal{R}_n(\mathcal{H})$

Recall, Theorem. Let $\mathcal{H}$ be a hypothesis class and assume the loss is bounded in $[0,1]$. Let $\mathcal{F} \triangleq \{\ell(h,\cdot); h \in \mathcal{H}\} \subset \mathbb{R}^{\mathcal{Z}}$ be the class of real-valued functions induced by applying ℓ to hypotheses $h \in \mathcal{H}$. Then, with probability at least $1 - 2e^{-2n\epsilon^2}$ we have,

$$\forall\, h \in \mathcal{H}, \quad |\widehat{L}(h) - L(h)| \leq 2\mathcal{R}_n(\mathcal{F}) + \epsilon.$$

This bound is in terms of $\mathcal{F} = \{\ell(h,\cdot); h \in \mathcal{H}\}$. Often, it is convenient to relate this to the Rademacher complexity of $\mathcal{H}$. This can often be done in supervised learning problems where $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, $\mathcal{Y} \subset \mathbb{R}$, and $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$.

We will consider two common cases:

- $0-1$ loss for binary classification, $\ell(h,(x,y)) = \mathbb{1}(h(x) \neq y)$.
- Lipschitz losses.

From $\mathcal{R}_n(\mathcal{F})$ to $\mathcal{R}_n(\mathcal{H})$: 0–1 loss for binary classification

Consider a setting where $\mathcal{Z} = \mathcal{X} \times \{0,1\}$ and the hypothesis class $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ is a set of binary predictors. We will use the 0–1 loss, $\ell(h,(x,y)) = \mathbb{1}(h(x) \neq y)$.

Theorem. Let $\mathcal{F} = \{\ell(h,\cdot); h \in \mathcal{H}\}$. For any dataset $S = \{(x_1,y_1),\ldots,(x_n,y_n)\}$, let $S_X = \{x_1,\ldots,x_n\}$ denote only the input values. Then, for all $S \subset (\mathcal{X}\times\mathcal{Y})^n$, we have $\widehat{\mathcal{R}}(\mathcal{F},S) = \widehat{\mathcal{R}}(\mathcal{H},S_X)$. Hence, $\mathcal{R}_n(\mathcal{F}) = \mathcal{R}_n(\mathcal{H})$.

Proof. First note that we can write,

$$\ell(h,(x,y)) = \mathbb{1}(h(x) \neq y) = y + h(x) - 2yh(x) = y + (1-2y)h(x).$$

Hence, for a given $S = \{(x_1,y_1),\ldots,(x_n,y_n)\}$ we can write,

$$\widehat{\mathcal{R}}(\mathcal{F},S) = \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^n \sigma_i f(x_i,y_i)\right] = \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^n \sigma_i \ell(h,(x_i,y_i))\right]$$
$$= \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^n \sigma_i\left(y_i + (1-2y_i)h(x_i)\right)\right]$$

From $\mathcal{R}_n(\mathcal{F})$ to $\mathcal{R}_n(\mathcal{H})$: 0–1 loss for binary classification (cont'd)

We therefore have,

$$\widehat{\mathcal{R}}(\mathcal{F}, S) = \frac{1}{n}\sum_{i=1}^{n} y_i \mathbb{E}[\sigma_i] + \mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}} \sum_{i=1}^{n} \sigma_i (1-2y_i) h(x_i)\right]$$

We know $\mathbb{E}[\sigma_i] = 0$. Moreover,

$$1 - 2y_i = \begin{cases} 1 & \text{if } y_i = 0, \\ -1 & \text{if } y_i = 1. \end{cases}$$

Hence, $(1-2y_i)\sigma_i$ is also a Rademacher random variable (for fixed y_i).

Therefore,

$$\widehat{\mathcal{R}}(\mathcal{F}, S) = \mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}} \sum_{i=1}^{n} \sigma_i h(x_i)\right] = \widehat{\mathcal{R}}(\mathcal{H}, S).$$

From $\mathcal{R}_n(\mathcal{F})$ to $\mathcal{R}_n(\mathcal{H})$: Lipschitz losses

Let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{Y} \subset \mathbb{R}$. Suppose the loss takes the form $\ell(h, (x, y)) = \ell'(h(x), y)$, where $\ell' : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is K-Lipschitz in the first argument. That is,

$$\text{for all } y_1, y_2, y_3 \in \mathcal{Y}, \qquad |\ell'(y_1, y_3) - \ell'(y_2, y_3)| \leq K|y_1 - y_2|.$$

You can verify that this is satisfied by $\ell'(y_1, y_2) = |y_1 - y_2|$ and $\ell'(y_1, y_2) = (y_1 - y_2)^2$ provided that $\mathcal{Y}$ is bounded. However, it is not satisfied by the $0 - 1$ loss $\ell'(y_1, y_2) = \mathbb{1}(y_1 \neq y_2)$.

Theorem. Let $\mathcal{F} = \{\ell(h, \cdot); h \in \mathcal{H}\}$. For any dataset $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, let $S_X = \{x_1, \ldots, x_n\}$ denote only the input values. Then, for all $S \subset (\mathcal{X} \times \mathcal{Y})^n$, we have $\widehat{\mathcal{R}}(\mathcal{F}, S) \leq K\widehat{\mathcal{R}}(\mathcal{H}, S_X)$. Hence, $\mathcal{R}_n(\mathcal{F}) \leq K\mathcal{R}_n(\mathcal{H})$.

Proof. Consider any fixed dataset $S = \{(x_i, y_i)\}_{i=1}^n$ of n points. We first have,

$$\begin{aligned}\mathcal{F}(S) &= \{[\ell(h, (x_1, y_1)), \ldots, \ell(h, (x_n, y_n))]; h \in \mathcal{H}\} \\ &= \{[\ell'(h(x_1), y_1), \ldots, \ell'(h(x_n), y_n)]; h \in \mathcal{H}\}.\end{aligned}$$

From $\mathcal{R}_n(\mathcal{F})$ to $\mathcal{R}_n(\mathcal{H})$: Lipschitz losses (cont'd)

Recall, Contraction lemma Let $\phi_1, \ldots, \phi_n$, where $\phi_i : \mathbb{R} \to \mathbb{R}$, be n K-Lipschitz functions, *i.e.*, $|\phi_i(a) - \phi_i(a')| \leq K|a - a'|$ for all i. For any $A \subset \mathbb{R}^n$, let $\phi(A) = \{[\phi_1(a_1), \ldots, \phi_n(a_n)]; a \in A\}$. Then, $\widehat{\mathcal{R}}(\phi(A)) \leq K\widehat{\mathcal{R}}(A)$.

Define $\phi_i(\cdot) = \ell'(\cdot, y_i)$, which is K-Lipschitz, allowing us to apply the contraction lemma. Denote $\phi : \mathbb{R}^n \to \mathbb{R}^n$ where $\phi(a) = [\phi_1(a_1), \ldots, \phi_n(a_n)]$.

We have, $\mathcal{H}(S_X) = \{[h(x_1), \ldots, h(x_n)]; h \in \mathcal{H}\}$ and hence,

$$\phi(\mathcal{H}(S_X)) = \{[\ell'(h(x_1), y_1), \ldots, \ell'(h(x_n), y_n)]; h \in \mathcal{H}\} = \mathcal{F}(S).$$

We therefore have,

$$\widehat{\mathcal{R}}(\mathcal{F}, S) = \widehat{\mathcal{R}}(\mathcal{F}(S)) = \widehat{\mathcal{R}}(\phi(\mathcal{H}(S_X))) \leq K \cdot \widehat{\mathcal{R}}(\mathcal{H}(S_X)) = K\widehat{\mathcal{R}}(\mathcal{H}, S_X).$$

Summary so far

Uniform convergence. There exists $\epsilon : \mathbb{N} \times (0,1) \to \mathbb{R}_+$ such that, for all $n \in \mathbb{N}$, $\delta \in (0,1)$, and distribution P, we have $\mathbb{P}\left(\forall h \in \mathcal{H},\ |\widehat{L}(h) - L(h)| \leq \epsilon(n,\delta)\right) \geq 1 - \delta$.

If the losses are bounded in $[0, B]$, for ERM $\widehat{h}$, we have

$$\mathbb{E}\left[L(\widehat{h}) - \inf_{h \in \mathcal{H}} L(h)\right] \leq 2\epsilon(n,\delta) + B\delta \qquad \texttt{ERM under uniform convergence}$$

$$\leq 2\left(2\mathcal{R}_n(\mathcal{F}) + \sqrt{\frac{1}{2n}\log\left(\frac{2}{\delta}\right)}\right) + \delta \qquad \texttt{Rademacher complexity}$$

$$\leq 4K\mathcal{R}_n(\mathcal{H}) + \sqrt{\frac{2}{n}\log\left(\frac{2}{\delta}\right)} + \delta.$$

Where, in the last line $K = 1$ for binary classification with the 0–1 loss and is the Lipshcitz constant for Lipschitz losses. Choosing $\delta = 1/\sqrt{n}$ we get,

$$\mathbb{E}\left[L(\widehat{h}) - \inf_{h \in \mathcal{H}} L(h)\right] \in \tilde{\mathcal{O}}\left(\mathcal{R}_n(\mathcal{H}) + n^{-1/2}\right)$$

Plan for the next two sub-chapters

In the remainder of the chapter, we will focus on bounding the Rademacher complexity of a hypothesis class $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$. We will focus on two cases:

- VC dimension, for binary classification problems
- Dudley entropy integral

Ch 3.3: VC Dimension and Sauer's Lemma

Recall the setting for binary classification, where $\mathcal{Z} = \mathcal{X} \times \{0, 1\}$ and the hypothesis class $\mathcal{H} \subset \{0, 1\}^{\mathcal{X}}$ is a set of binary predictors.

Outline for this subchapter:

- Bounding Rademacher complexity via growth function
- Shattering and VC dimension, examples
- Sauer's lemma
- VC dimension-based lower bound.

Growth function

Growth function. Let $n \in \mathbb{N}$ and let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be a hypothesis class. The growth function is defined as

$$g(n, \mathcal{H}) \triangleq \max_{A \subset \mathcal{X},\, |A|=n} |\mathcal{H}(A)|\,,$$

where, recall $\mathcal{H}(A) = \{\{h(a)\}_{a \in A};\, h \in \mathcal{H}\}$.

Interpretation: $g(n, \mathcal{H})$ the maximum number of label vectors that can be realized on a set of n points in $\mathcal{X}$ with hypotheses in $\mathcal{H}$.

Note that $g(n, \mathcal{H}) \leq 2^n$.

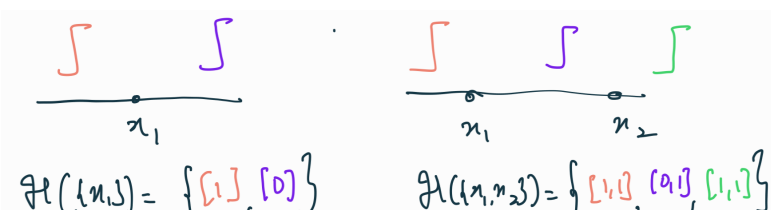
Example: One-sided threshold classifiers

Let $\mathcal{H} = \{h_a(x) = \mathbb{1}(x \geq a);\ a \in \mathbb{R}\}$ be all one-sided threshold classifiers. Let us try computing $g(n, \mathcal{H})$ for different n.

- $g(1, \mathcal{H})$: We have $\mathcal{H}(\{x_1\}) = \{[0], [1]\}$, for any $x_1 \in \mathbb{R}$ (see figure below). Hence, $g(1, \mathcal{H}) = 2 = 2^1$.

- $g(2, \mathcal{H})$: We have $\mathcal{H}(\{x_1, x_2\}) = \{[0, 0], [0, 1], [1, 1]\}$, for any $x_1 \in \mathbb{R}$. Hence, $g(2, \mathcal{H}) = 3 < 2^2$.

- You can also verify $g(n, \mathcal{H}) < 2^n$ for all $n \geq 2$.



Example: Two-sided threshold classifiers

Let $\mathcal{H} = \{h_a(x) = \mathbb{1}(x \geq a);\ a \in \mathbb{R}\} \cup \{h_a(x) = \mathbb{1}(x \leq a);\ a \in \mathbb{R}\}$ be all two-sided threshold classifiers. Let us try computing $g(n, \mathcal{H})$ for different n.

- $g(1, \mathcal{H})$: We have $\mathcal{H}(\{x_1\}) = \{[0], [1]\}$, for any $x_1 \in \mathbb{R}$. Hence, $g(1, \mathcal{H}) = 2 = 2^1$.

- $g(2, \mathcal{H})$: We have, $\mathcal{H}(\{x_1, x_2\}) = \{[0, 0], [0, 1], [1, 0], [1, 1]\}$, for any $x_1, x_2 \in \mathbb{R}$. Hence, $g(2, \mathcal{H}) = 4 = 2^2$.

- $g(3, \mathcal{H})$: We have,
$\mathcal{H}(\{x_1, x_2, x_3\}) = \{[0, 0, 0], [0, 0, 1], [0, 1, 1], [1, 1, 1], [1, 0, 0], [1, 1, 0]\}$, for any $x_1, x_2, x_3 \in \mathbb{R}$. Hence, $g(3, \mathcal{H}) = 6 < 2^2$.

- You can also verify $g(n, \mathcal{H}) < 2^n$ for all $n \geq 3$.

N.B. In both examples, $g(n, \mathcal{H}) = 2^n$ whenever n does not exceed the "degrees of freedom" in $\mathcal{H}$.

Bounding the Rademacher complexity via the growth function

Theorem. Let $A = \{X_1, \ldots, X_n\} \subset \mathcal{X}$ and let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$. Then,

$$\widehat{\mathcal{R}}(\mathcal{H}, A) \leq \sqrt{\frac{2\log(|\mathcal{H}(A)|)}{n}} \leq \sqrt{\frac{2\log(g(n,\mathcal{H}))}{n}}.$$

Hence, $\mathcal{R}_n(\mathcal{H}) \leq \sqrt{\frac{2\log(g(n,\mathcal{H}))}{n}}$.

Recall, Massart's Lemma Let A be a finite subset of $\mathbb{R}^n$. Then, $\widehat{\mathcal{R}}(A) \leq (\max_{a\in A} \|a\|_2) \cdot \frac{\sqrt{2\log(|A|)}}{n}$.

Proof. Noting that $\mathcal{H}(A) \subset \{0,1\}^n \subset \mathbb{R}^n$, we can write

$$\widehat{\mathcal{R}}(\mathcal{H}, A) = \widehat{\mathcal{R}}(\mathcal{H}(A)) \leq \underbrace{\left(\max_{a\in\mathcal{H}(A)} \|a\|_2\right)}_{\leq\sqrt{n}} \cdot \frac{\sqrt{2\log(|\mathcal{H}(A)|)}}{n}.$$

Shattering and VC dimension

We just showed $\mathcal{R}_n(\mathcal{H}) \leq \sqrt{\frac{2\log(g(n,\mathcal{H}))}{n}}$.
But with the naive bound $g(n,\mathcal{H}) \leq 2^n$, we can only get $\mathcal{R}_n(\mathcal{H}) \in \mathcal{O}(1)$.

For learnability, we want at least $g(n,\mathcal{H}) \in o(2^n)$ so that $\mathcal{R}_n(\mathcal{H}) \in o(1)$.

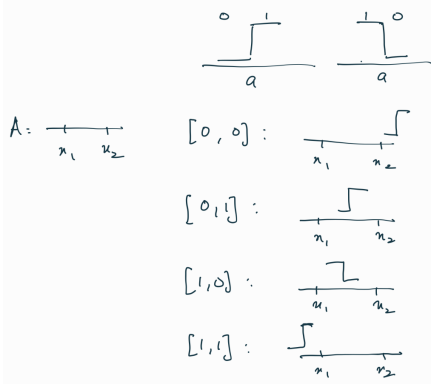
Ideally, we should achieve $g(n,\mathcal{H}) \asymp \mathrm{poly}(n)$ so that we get the $\tilde{\mathcal{O}}(n^{-1/2})$ rate.

Next, we will introduce the VC dimension. We will see that for classes with finite VC dimension, we can get $\mathcal{R}_n \in \tilde{\mathcal{O}}(n^{-1/2})$.

Shattering and VC dimension (cont'd)

Shattering. Let $A = \{x_1, \ldots, x_n\} \subset \mathcal{X}$ be a set of n points in $\mathcal{X}$. We say that A is shattered by a hypothesis class $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ if $\mathcal{H}$ can realize any label on A. That is, if $|\mathcal{H}(A)| = 2^n$.

Example 1. Two-sided threshold classifiers can shatter any set of 2 *distinct* points, but no set of 3 points.



Shattering and VC dimension (cont'd)

VC dimension. The VC dimension d of a hypothesis class $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ is the size of the largest set shattered by $\mathcal{H}$. Equivalently, $d = \max\{n;\ \ g(n, \mathcal{H}) = 2^n\}$.

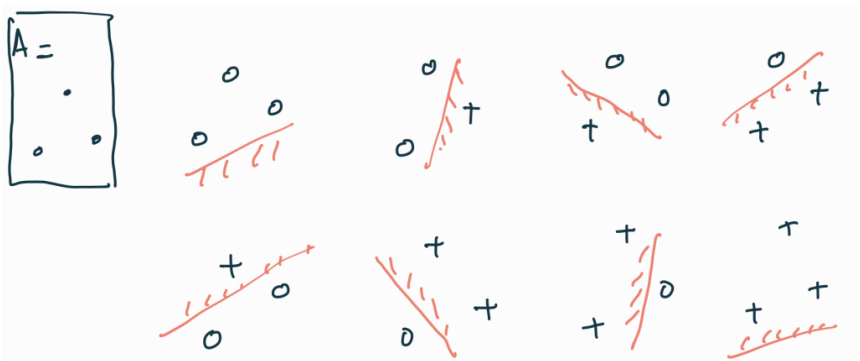
Example 1. Two-sided threshold classifiers
$\mathcal{H} = \{h_a(x) = \mathbb{1}(x \geq a);\ a \in \mathbb{R}\} \cup \{h_a(x) = \mathbb{1}(x \leq a);\ a \in \mathbb{R}\}$.
- We know it can shatter 2 points, therefore $d \geq 2$.
- But it cannot shatter any set of 3 points, therefore $d = 2$.
- There are two dofs here: a and which direction is class 1.

Example 2. One-sided threshold classifiers $\mathcal{H} = \{h_a(x) = \mathbb{1}(x \geq a);\ a \in \mathbb{R}\}$.
- We know it can shatter 1 point, therefore $d \geq 1$.
- But it cannot shatter any set of 2 points, therefore $d = 1$.
- Only one dof: only a.

Shattering and VC dimension (cont'd)

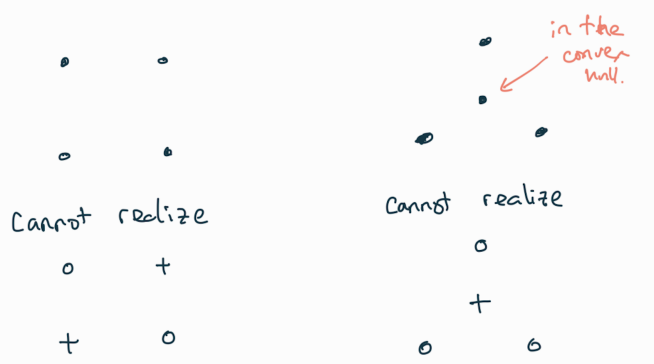
Example 3. Linear classifiers when $\mathcal{X} = \mathbb{R}^2$.

- It can shatter 3 points (see figure below), therefore $d \geq 3$.
- But it cannot shatter any set of 4 points (see next slide), therefore $d = 3$.
- There are three dofs: two parameters for the line and the direction.



Shattering and VC dimension (cont'd)

Linear classifiers in $\mathbb{R}^2$ cannot shatter 4 points.
- If three of the four points are co-linear, it cannot be shattered.
- Two cases for four non-co-linear points:



In general, for linear classifiers in $\mathbb{R}^D$, we have $d = D + 1$. (Proof in HW2)

Sauer's lemma

Sauer's Lemma. If the VC dimension of a hypothesis class $\mathcal{H}$ is d, then

$$g(n, \mathcal{H}) \leq \sum_{i=0}^{d} \binom{n}{i}.$$

As $\binom{n}{i} = 0$ if $n < i$, this means

$$g(n, \mathcal{H}) \begin{cases} = 2^n & \text{if } n \leq d, \\ \leq \sum_{i=0}^{d} \binom{n}{i} & \text{if } n > d. \end{cases}$$

Before proving Sauer's lemma, let us first use it to bound the Rademacher complexity.

Theorem. Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be a hypothesis class with VC dimension d. Then, for all $n \geq d$, we have

$$\mathcal{R}_n(\mathcal{H}) \leq \sqrt{\frac{2d}{n} \log\left(\frac{en}{d}\right)} \in \tilde{\mathcal{O}}\left(\sqrt{\frac{d}{n}}\right)$$

Sauer's lemma (cont'd)

Proof. Let us first upper bound $g(n, \mathcal{H})$ as follows. Using Sauer's lemma,

$$
\begin{aligned}
g(n, \mathcal{H}) &\leq \sum_{i=0}^{d} \binom{n}{i} && \texttt{Sauer's lemma} \\
&= \left(\frac{n}{d}\right)^d \sum_{i=0}^{d} \binom{n}{i} \left(\frac{d}{n}\right)^d \\
&\leq \left(\frac{n}{d}\right)^d \sum_{i=0}^{d} \binom{n}{i} \left(\frac{d}{n}\right)^i && \texttt{as } d \leq n \texttt{ and } i \leq d \\
&\leq \left(\frac{n}{d}\right)^d \sum_{i=0}^{n} \binom{n}{i} \left(\frac{d}{n}\right)^i && \texttt{as } d \leq n
\end{aligned}
$$

Sauer's lemma (cont'd)

Recall: By bounding Rademacher complexity by growth function we showed $\mathcal{R}_n(\mathcal{H}) \leq \sqrt{\frac{2\log(g(n,\mathcal{H}))}{n}}$.

$$
\begin{aligned}
g(n,\mathcal{H}) &\leq \left(\frac{n}{d}\right)^d \left(1+\frac{d}{n}\right)^n && \texttt{binomial expansion of } (1+d/n)^n \\
&\leq \left(\frac{n}{d}\right)^d \cdot e^d && \texttt{as } (1+x/n)^n \leq e^x \\
&= \left(\frac{en}{d}\right)^d .
\end{aligned}
$$

We therefore have,

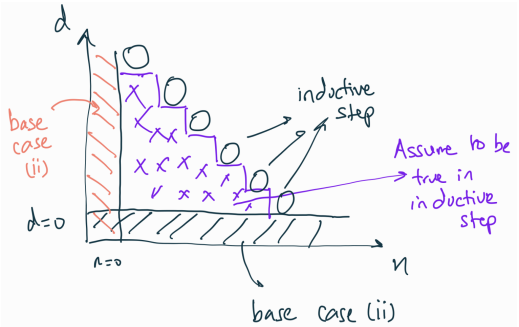
$$\mathcal{R}_n(\mathcal{H}) \leq \sqrt{\frac{2d\log(en/d)}{n}}$$

Ch 3.3.1: Proof of Sauer's lemma

We wish to show $g(n, \mathcal{H}) \leq \sum_{i=0}^{d} \binom{n}{i}$.

Plan: We will prove the lemma via induction on $k = n + d$.

- **Base case:** *(i)* $n = 0$, any d, *(ii)* $d = 0$, any n.
- **Inductive step:** Assume true for all n, d with $n + d \leq k - 1$, and prove for $n + d = k$.



Proof of Sauer's lemma (cont'd)

Denote $\Phi_d(n) = \sum_{i=0}^{d} \binom{n}{i}$.
Recall the definition of the growth function, $g(n, \mathcal{H}) = \max_{A \subset \mathcal{X},\, |A|=n} |\mathcal{H}(A)|$.

Base case.
(i) Let $n = 0$, and consider a hypothesis class $\mathcal{H}$ with VC dimension $d \in \mathbb{N}$. First note that $g(0, \mathcal{H}) = 1$ as we can label 0 points at most one way:

$$\text{If } A = \{\}, \text{ then } \mathcal{H}(A) = \{[\,]; h \in \mathcal{H}\} = \{[\,]\} \quad \Longrightarrow \quad |\mathcal{H}(A)| = 1.$$

Moreover, we have $\Phi_d(0) = \sum_{i=0}^{d} \binom{0}{i} = 1$, as $\binom{0}{0} = 1$ and $\binom{0}{i} = 0$ for $i > 0$.

(ii) For any hypothesis class $\mathcal{H}$ with VC dimension $d = 0$, we have $|\mathcal{H}(A)| = 1$ for any $A \subset \mathcal{X}$ as all hypotheses in $\mathcal{H}$ will label the points in A the same way. In particular, if any point in A can be labeled both 0 and 1 by $\mathcal{H}$, then the VC dimension would be at least 1.

Moreover, we have $\Phi_d(n) = \sum_{i=0}^{0} \binom{n}{i} = \binom{n}{0} = 1$.

Proof of Sauer's lemma (cont'd)

Inductive step. Assume the statement is true for all $d + n \leq k - 1$. Let d, n be such that $d + n = k$.

Let $\mathcal{H}$ be a hypothesis class with VC dimension $d_{\mathcal{H}} = d$.
Let $A = \{x_1, \ldots, x_n\} \subset \mathcal{X}$ be of size n.

Let $G \subset \{g : A \to \mathcal{Y}\}$ be a class of functions defined only over A as follows: for every element of $\mathcal{H}(A)$, add one element from $\mathcal{H}$ to $\mathcal{G}$.
Hence, $|\mathcal{G}| = |\mathcal{G}(A)| = |\mathcal{H}(A)|$.

We will now partition $\mathcal{G}$ into $\mathcal{G}_1, \mathcal{G}_2$ as follows: for every possible labeling of $\{x_1, \ldots, x_{n-1}\}$, add one function from $\mathcal{G}$ to $\mathcal{G}_1$. Then, let $\mathcal{G}_2 = \mathcal{G} \backslash \mathcal{G}_1$.

Proof of Sauer's lemma (cont'd)

By our construction, $\mathcal{H}(A) = \mathcal{G}(A) = \mathcal{G}_1(A) \cup \mathcal{G}_2(A)$. Hence,

$$|\mathcal{H}(A)| = |\mathcal{G}(A)| = |\mathcal{G}_1(A)| + |\mathcal{G}_2(A)|. \tag{3}$$

Moreover,

$$d_{\mathcal{G}_1} \underbrace{\leq}_{(1)} d_{\mathcal{G}} \underbrace{\leq}_{(2)} d_{\mathcal{H}} = d. \tag{4}$$

(1): as $\mathcal{G}_1 \subset \mathcal{G}$, (2): as any set shattered by G is also shattered by $\mathcal{H}$.

Claim 1: $|\mathcal{G}_1(A)| = |\mathcal{G}_1(\{x_1, \ldots, x_{n-1}\})|$.
Proof. For every labeling $\{g(x_1), \ldots, g(x_{n-1})\}$ of $\{x_1, \ldots, x_{n-1}\}$ in $\mathcal{G}_1(\{x_1, \ldots, x_{n-1}\})$, we have exactly one of $\{g(x_1), \ldots, g(x_{n-1}), 0\}$ or $\{g(x_1), \ldots, g(x_{n-1}), 1\}$ in $\mathcal{G}_1(A) = \mathcal{G}_1(\{x_1, \ldots, x_n\})$.

Proof of Sauer's lemma (cont'd)

Claim 2: $|\mathcal{G}_2(A)| = |\mathcal{G}_2(\{x_1, \ldots, x_{n-1}\})|$.
Proof: Follows a similar argument to Claim 1

To apply the inductive hypothesis, we first note that by (4),
$d_{\mathcal{G}_1} + n - 1 \leq d_{\mathcal{H}} + n - 1 = d + n - 1 = k - 1$. Therefore,

$$
\begin{aligned}
|\mathcal{G}_1(A)| &= |\mathcal{G}_1(\{x_1, \ldots, x_{n-1}\})| && \texttt{By Claim 1} \\
&\leq g(n-1, d_{\mathcal{G}_1}) && \texttt{definition of Growth function} \\
&\leq \Phi_{d_{\mathcal{G}_1}}(n-1) && \texttt{by the inductive hypothesis} \\
&\leq \Phi_d(n-1) && \texttt{as } \Phi_{d'}(n) \texttt{ is increasing in } d'
\end{aligned}
$$

Proof of Sauer's lemma (cont'd)

Claim 3. $d_{\mathcal{G}_2} \leq d - 1$.

Proof. Note that for every $g_2 \in \mathcal{G}_2$, there exists $g_1 \in \mathcal{G}_1$ which disagrees with g_2 on x_n. Therefore, if $T \subset \{x_1, \ldots, x_{n-1}\}$ is shattered by $\mathcal{G}_2$, then $T \cup \{x_n\}$ is shattered by $\mathcal{G}$. Hence,

$$|T \cup \{x_n\}| \leq d_{\mathcal{G}} \quad \Longrightarrow \quad |T| \leq d_{\mathcal{G}} - 1 \leq d_{\mathcal{H}} - 1 = d - 1.$$

Taking the maximum over all T that is shattered by $\mathcal{G}_2$, we get $d_{\mathcal{G}_2} \leq d - 1$.

Continuing with our main proof, we have

$$\begin{aligned}
|\mathcal{G}_2(A)| &= |\mathcal{G}_2(\{x_1, \ldots, x_{n-1}\})| && \text{By Claim 2} \\
&\leq g(n-1, \mathcal{G}_2) && \text{definition of Growth function} \\
&\leq \Phi_{d_{\mathcal{G}_2}}(n-1) && \text{by the inductive hypothesis} \\
&\leq \Phi_{d-1}(n-1) && \text{as } \Phi_{d'}(n) \text{ is increasing in } d' \text{ and Claim 3.}
\end{aligned}$$

Proof of Sauer's lemma (cont'd)

Recall we showed following: *(i)* $|\mathcal{H}(A)| = |\mathcal{G}(A)| = |\mathcal{G}_1(A)| + |\mathcal{G}_2(A)|$.
(ii) $|\mathcal{G}_1(A)| \leq \Phi_d(n-1)$, *(iii)* $|\mathcal{G}_2(A)| \leq \Phi_{d-1}(n-1)$.

We therefore have,

$$
\begin{aligned}
|\mathcal{H}(A)| &= |\mathcal{G}_1(A)| + |\mathcal{G}_2(A)| && \text{by } (i) \\
&\leq \Phi_d(n-1) + \Phi_{d-1}(n-1) && \text{by } (ii) \text{ and } (iii) \\
&= \sum_{i=0}^{d} \binom{n-1}{i} + \sum_{i=0}^{d-1} \binom{n-1}{i} \\
&= \binom{n-1}{0} + \sum_{i=1}^{d} \binom{n-1}{i} + \sum_{i=1}^{d} \binom{n-1}{i-1}
\end{aligned}
$$

Proof of Sauer's lemma (cont'd)

Continuing, we have

$$|\mathcal{H}(A)| \leq \binom{n-1}{0} + \sum_{i=1}^{d} \binom{n}{i} \qquad \text{as } \binom{n}{i} = \binom{n-1}{i} + \binom{n-1}{i-1}$$

$$= \sum_{i=0}^{d} \binom{n}{i} = \Phi_d(n). \qquad \text{As } \binom{n-1}{0} = \binom{n}{0} = 1\,.$$

This is true for any $A \subset \mathcal{X}$ such that $|A| = n$, hence

$$g(n, \mathcal{H}) = \max_{A \subset \mathcal{X}, |A|=n} |\mathcal{H}(A)| \leq \Phi_d(n).$$

Ch 3.3.2: Lower bounds for binary classification

Let $\mathcal{X}$ be any input space and $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be a hypothesis class with VC dimension d. We have shown that ERM achieves the following upper bound on the excess risk on all distributions P over $\mathcal{X} \times \{0,1\}$:

$$\begin{aligned} R_{\mathcal{H}}(\widehat{h}_{\mathrm{ERM}}, P) &= \mathbb{E}_{S \sim P^n}\left[L_{\mathcal{H}}(\widehat{h}_{\mathrm{ERM}}(S), P)\right] \\ &= \mathbb{E}_{S \sim P^n}\left[L\left(\widehat{h}_{\mathrm{ERM}}(S), P\right)\right] - \inf_{h \in \mathcal{H}} L(h, P) \in \tilde{\mathcal{O}}\left(\sqrt{\frac{d}{n}}\right). \end{aligned}$$

We will show that we cannot do significantly better, *i.e.*,

$$\inf_{\widehat{h}} \sup_{P} R_{\mathcal{H}}(\widehat{h}, P) \in \Omega\left(\sqrt{\frac{d}{n}}\right).$$

Lower bounds for binary classification (cont'd)

We will show this lower bound via Fano's method. Recall the following corollary:

Local Fano method. Let $\Delta(P, Q)$ be the separation as defined below,

$$\Delta(P, Q) = \sup \Big\{ \delta \geq 0; \quad L_{\mathcal{H}}(h, P) \leq \delta \implies L_{\mathcal{H}}(h, Q) \geq \delta, \quad \forall\, h \in \mathcal{H}, \\ L_{\mathcal{H}}(h, Q) \leq \delta \implies L_{\mathcal{H}}(h, P) \geq \delta, \quad \forall\, h \in \mathcal{H}, \Big\}.$$

Let S be an i.i.d dataset from some distribution $P \in \mathcal{P}$. Let $\{P_1, \ldots, P_N\} \subset \mathcal{P}$ such that $\Delta(P_j, P_k) \geq \delta$ and $\mathrm{KL}(P_j, P_k) \leq \frac{\log(N)}{4n}$ for all $j \neq k$. Suppose $N \geq 16$. Then, $R_n^\star \geq \frac{\delta}{2}$.

Step 1: Constructing alternatives

Let $\mathcal{X}_d = \{x_1, \ldots, x_d\}$ be a set of d points shattered by $\mathcal{H}$. Let $\gamma < 1/4$ be a value we will specify shortly.

For $\omega \in \{0,1\}^d$, define the distribution P_ω with pmf p_ω as follows:

$$p_\omega(x) = \frac{1}{d}\mathbb{1}\{x \in \mathcal{X}_d\}, \qquad p_\omega(Y = 1 | X = x_i) = \frac{1}{2} + (2\omega_i - 1)\gamma,$$

Let Ω_d be the Gilbert-Varshamov-pruned d-hypercube. Now define our alternatives as,

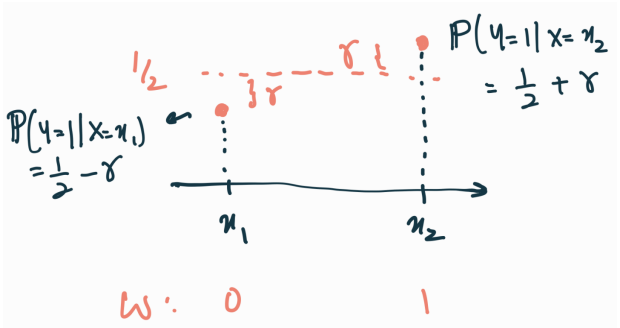
$$\mathcal{P}' = \left\{ P_\omega ; \omega \in \Omega_d \right\}$$

Step 1: Constructing alternatives (cont'd)

An example. For two-sided threshold classifiers, pick any x_1, x_2 such that $x_1 \neq x_2$. This is shattered by $\mathcal{H}$. For $\omega = (0, 1)$, the distribution P_ω would be,

$$p_\omega(X = x_1) = p_\omega(X = x_2) = \frac{1}{2},$$

$$p_\omega(Y = 1|X = x_1) = \frac{1}{2} - \gamma, \quad p_\omega(Y = 1|X = x_2) = \frac{1}{2} + \gamma.$$



Step 2: Lower bound the separation

Recall, Gilbert-Varshamov bound. In the Gilbert-Varshamov pruned m-hypercube Ω_m of $\{0,1\}^m$, we have, *(i)* $|\Omega_m| \geq 2^{m/8}$, *(ii)* $H(\omega, \omega') \geq m/8$, $\forall\, \omega, \omega' \in \Omega_m$, *(iii)* $\mathbf{0}_m \in \Omega_m$.

Claim 1. For any $\omega, \omega' \in \{0,1\}^d$, we have, $\Delta(P_\omega, P_{\omega'}) \geq \frac{\gamma}{d} H(\omega, \omega')$.
Proof. In HW3.

Then, by the Gilbert-Varshamov bound,

$$\min_{P_\omega, P_{\omega'} \in \mathcal{P}'} \Delta(P_\omega, P'_\omega) \geq \frac{\gamma}{d} \min_{\omega, \omega' \in \Omega_d} H(\omega, \omega') \geq \frac{\gamma}{d} \cdot \frac{d}{8} = \frac{\gamma}{8} \quad \triangleq \texttt{ the separation } \delta$$

Step 3: Upper bound the KL divergence

For any P_ω, P'_ω, we can bound

$$\begin{aligned}
\mathrm{KL}(P_\omega, P_{\omega'}) &= \mathbb{E}_{X,Y}\left[\log\left(\frac{p_\omega(x,y)}{p_{\omega'}(x,y)}\right)\right] \\
&= \sum_{i=1}^{d} p_\omega(x_i) \sum_{y\in\{0,1\}} p_\omega(y|x_i)\log\left(\frac{p_\omega(y|x_i)}{p_{\omega'}(y|x_i)}\right) \\
&= \sum_{i=1}^{d} \frac{1}{d}\mathbb{1}(\omega \neq \omega_i) \underbrace{\left(\left(\frac{1}{2}+\gamma\right)\log\left(\frac{1/2+\gamma}{1/2-\gamma}\right) + \left(\frac{1}{2}-\gamma\right)\log\left(\frac{1/2-\gamma}{1/2+\gamma}\right)\right)}_{\leq 16\gamma^2} \\
&\leq \frac{16\gamma^2 H(\omega,\omega')}{d}.
\end{aligned}$$

Therefore, $\max_{P_\omega, P'_\omega \in \mathcal{P}'} \mathrm{KL}(P_\omega, P_{\omega'}) \leq C_1\gamma^2$.

Step 4: Final bound

We require the maximum KL ($C_1\gamma^2$) to be smaller than $\frac{\log(N)}{4n}$.
By the Gilbert-Varshamov bound, we know $N \geq 2^{d/8}$. Therefore, it is sufficient if

$$C_1\gamma^2 \leq \frac{\log(2^{d/8})}{4n} = C_2\frac{d}{n}.$$

Therefore, choose $\gamma = C_3\sqrt{\frac{d}{n}}$.

As the separation is $\delta = \gamma/8$, we have

$$\inf_{\widehat{h}} \sup_{P} R_{\mathcal{H}}(\widehat{h}, P) \geq \frac{1}{2}\delta = C_4\sqrt{\frac{d}{n}}.$$

Ch 3.4: Dudley Entropy Integral

Recall our goal: we wish to bound the Rademacher complexity $\mathcal{R}_n(\mathcal{H})$ for a hypothesis class, where

$$\mathcal{R}_n(\mathcal{H}) = \mathbb{E}_{S\sim P^n}[\widehat{\mathcal{R}}(\mathcal{H}(S))].$$

In particular, if we can control $\widehat{\mathcal{R}}(\mathcal{H}(S))$ for all datasets S, then we can bound $\mathcal{R}_n(\mathcal{H})$.

We know that for finite $A \subset \mathbb{R}^n$, we have $\widehat{\mathcal{R}}(A) \leq (\max_{a\in A} \|a\|_2) \frac{\sqrt{2\log(|A|)}}{n}$, by Massart's lemma. However, generally speaking, $\widehat{\mathcal{R}}(\mathcal{H}(S))$ may be infinite.

An idea: If we can find an ϵ cover C_ϵ of $\mathcal{H}(S)$, then we can obtain a bound of the following form, where ϵ is the error due to approximation and the $\sqrt{\log(|C_\epsilon|)}/n$ term is from Massart's lemma.

$$\mathcal{O}\left(\epsilon + \frac{\sqrt{2\log(|C_\epsilon|)}}{n}\right).$$

A first theorem

The following theorem formalizes this idea. For $A \subset \mathbb{R}^n$, let $N_p(\epsilon, A)$ denote the ϵ covering number in the $\frac{1}{n^{1/p}}\|\cdot\|_p$ norm. That is,

$$N_p(\epsilon, A) = N\left(\epsilon, A, \frac{1}{n^{1/p}}\|\cdot\|_p\right), \text{ for } p \in [1, \infty), \qquad N_\infty(\epsilon, A) = N\left(\epsilon, A, \|\cdot\|_\infty\right).$$

Theorem. Let $p \in [1, \infty]$ and suppose $A \subset [-B, B]^n$. Then,

$$\widehat{\mathcal{R}}(A) \leq \inf_{\epsilon > 0}\left(\epsilon + B\sqrt{\frac{2\log N_p(\epsilon, A)}{n}}\right)$$

Proof. Recall, in HW0, you showed $N_p(\epsilon, A) \leq N_q(\epsilon, A)$ for $p \leq q$. Therefore, it is sufficient to prove the above theorem with $N_1(\epsilon, A)$.

A first theorem (cont'd)

Fix any $\epsilon > 0$. Let C_ϵ be a minimal ϵ-cover of A in the $\frac{1}{n}\|\cdot\|_1$ norm. For an $a \in A$, let $[a]$ denote the closest element to a in the cover, satisfying $\frac{1}{n}\|a - [a]\|_1 \leq \epsilon$.

We therefore have,

$$\begin{aligned}
\widehat{\mathcal{R}}(A) &= \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{a\in A} \sigma^\top a\right] \\
&= \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{a\in A} \sigma^\top (a - [a] + [a])\right] \\
&\leq \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{a\in A} \sigma^\top (a - [a])\right] \;+\; \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{a\in A} \sigma^\top [a]\right]
\end{aligned}$$

We now observe that by Hölder's inequality, we have

$$\sigma^\top (a - [a]) \leq \|\sigma\|_\infty \|a - [a]\|_1 \leq 1 \cdot n\epsilon.$$

Therefore, the first term of the RHS above is at most ϵ.

A first theorem (cont'd)

Massart's Lemma. Let A be a finite subset of $\mathbb{R}^n$. Then, $\widehat{\mathcal{R}}(A) \leq (\max_{a \in A} \|a\|_2) \frac{\sqrt{2\log(|A|)}}{n}$.

We will bound the second term using Massart's lemma,

$$\frac{1}{n}\mathbb{E}_\sigma\left[\sup_{a\in A} \sigma^\top [a]\right] = \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{c\in C_\epsilon} \sigma^\top c\right] = \widehat{\mathcal{R}}(C_\epsilon)$$

$$\leq \underbrace{\left(\max_{c\in C_\epsilon} \|c\|_2\right)}_{\leq B\sqrt{n}} \frac{\sqrt{2\log(|C_\epsilon|)}}{n} \leq B\sqrt{\frac{2\log(N_1(\epsilon, A))}{n}}.$$

This gives us,

$$\widehat{\mathcal{R}}(A) \leq \epsilon + B\sqrt{\frac{2\log(N_1(\epsilon, A))}{n}}.$$

As this bound is true for all ϵ, the claim follows by taking an infimum over ϵ.

Example: Linear regression in $\mathbb{R}^d$

Theorem. Let $p \in [1, \infty]$ and $A \subset [-B, B]^n$. Then, $\widehat{\mathcal{R}}(A) \leq \inf_\epsilon \left(\epsilon + B\sqrt{\frac{2 \log N_p(\epsilon, A)}{n}} \right)$

Setting. Let p, q be such that $\frac{1}{p} + \frac{1}{q} = 1$. Let $\mathcal{H} = \{h_\theta; h_\theta(x) = \theta^\top x; \theta \in \mathbb{R}^d, \|\theta\|_p \leq B_\theta\}$. Suppose that for all $X \sim P$, we have $X \in \mathbb{R}^d$ and $\|X\|_q \leq B_x$.

For any x, θ such that $\|x\|_q \leq B_x$ and $\|\theta\|_p \leq B_\theta$, by Hölder's inequality we have $|x^\top \theta| \leq B_x B_\theta$. Noting that $\mathcal{R}_n(\mathcal{H}) = \mathbb{E}_{S \sim P^n}[\widehat{\mathcal{R}}(\mathcal{H}(S))]$, we will aim to bound $\widehat{\mathcal{R}}(\mathcal{H}(S))$ as follows,

$$\widehat{\mathcal{R}}(\mathcal{H}(S)) \leq \inf_{\epsilon > 0} \left(\epsilon + B_x B_\theta \sqrt{\frac{2 \log(N_\infty(\epsilon, \mathcal{H}(S)))}{n}} \right).$$

Example: Linear regression (cont'd)

Recall, Theorem. Let $\mathcal{X} = \mathbb{R}^d$ and let $\|\cdot\|$ be any norm. Let $B = \{x \in \mathbb{R}^d; \|x\| \leq 1\}$ be the unit ball. Suppose A is a convex set and contains ϵB. Then,

$$\left(\frac{1}{\epsilon}\right)^d \frac{\mathrm{vol}(A)}{\mathrm{vol}(B)} \leq N(\epsilon, A, \|\cdot\|) \leq M(\epsilon, A, \|\cdot\|) \leq \frac{\mathrm{vol}\left(A + \frac{\epsilon}{2}B\right)}{\mathrm{vol}\left(\frac{\epsilon}{2}B\right)} \leq \left(\frac{3}{\epsilon}\right)^d \frac{\mathrm{vol}(A)}{\mathrm{vol}(B)}.$$

Note that as $|\theta^\top x_i| \leq B_x B_\theta$, we have $\mathcal{H}(S) \subset \mathcal{B}_\infty(B_x B_\theta) \subset \mathbb{R}^n$. Naively applying above theorem with this observation to bound $N_\infty(\epsilon, \mathcal{H}(S))$ leads to

$$\begin{aligned} N_\infty(\epsilon, \mathcal{H}(S)) &\leq N_\infty(\epsilon, \mathcal{B}_\infty(B_x B_\theta)) = N(\epsilon, \mathcal{B}_\infty(B_x B_\theta), \|\cdot\|_\infty) \\ &\leq \left(\frac{3}{\epsilon}\right)^n \frac{\mathrm{vol}(\mathcal{B}_\infty(B_x B_\theta))}{\mathrm{vol}(\mathcal{B}_\infty(1))} = \left(\frac{3 B_x B_\theta}{\epsilon}\right)^n \end{aligned}$$

which only gives a constant bound for $\mathcal{R}_n(\mathcal{H})$.

Instead, we note that as $\theta \in \mathbb{R}^d$, $\mathcal{H}(S)$ lies in a d-dimensional hyperplane in $\mathbb{R}^n$.

Example: Linear regression (cont'd)

Recall, Theorem. Let $\mathcal{X} = \mathbb{R}^d$ and let $\|\cdot\|$ be any norm. Let $B = \{x \in \mathbb{R}^d; \|x\| \le 1\}$ be the unit ball. Suppose A is a convex set and contains ϵB. Then,

$$\left(\frac{1}{\epsilon}\right)^d \frac{\mathrm{vol}(A)}{\mathrm{vol}(B)} \le N(\epsilon, A, \|\cdot\|) \le M(\epsilon, A, \|\cdot\|) \le \frac{\mathrm{vol}\left(A + \frac{\epsilon}{2}B\right)}{\mathrm{vol}\left(\frac{\epsilon}{2}B\right)} \le \left(\frac{3}{\epsilon}\right)^d \frac{\mathrm{vol}(A)}{\mathrm{vol}(B)}.$$

To use this observation, let us first construct an ϵ'-cover $C_{\epsilon'}$ of $\{\theta; \|\theta\|_p \le B_\theta\}$ in the $\|\cdot\|_p$ norm. Let $\mathcal{H}_{\epsilon'} = \{h_\theta; h_\theta(x) = \theta^\top x; \theta \in C_{\epsilon'}\}$. By the above theorem, we know

$$|\mathcal{H}_{\epsilon'}| = |C_{\epsilon'}| \le \left(\frac{3}{\epsilon'}\right)^d \frac{\mathrm{vol}(\mathcal{B}_p(B_\theta))}{\mathrm{vol}(\mathcal{B}_p(1)} = \left(\frac{3B_\theta}{\epsilon'}\right)^d.$$

Claim. Let $S = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$, where $\|x_i\|_q \le B_x$. Then, $\mathcal{H}_{\frac{\epsilon}{B_x}}(S)$ is an ϵ-cover of $\mathcal{H}(S)$ in the $\|\cdot\|_\infty$ norm. Moreover $|\mathcal{H}_{\frac{\epsilon}{B_x}}(S)| \le \left(\frac{3B_\theta B_x}{\epsilon}\right)^d$.

Example: Linear regression (cont'd)

Proof of claim. The statement about the size follows via,

$$|\mathcal{H}_{\frac{\epsilon}{B_x}}(S)| = |\{[h(x_1), \ldots, h(x_n)]; h \in \mathcal{H}_{\frac{\epsilon}{B_x}}\}| \leq |\mathcal{H}_{\frac{\epsilon}{B_x}}| \leq \left(\frac{3B_\theta B_x}{\epsilon}\right)^d.$$

To show the covering bound, let $[\theta]$ be the closest element to θ in $C_{\frac{\epsilon}{B_x}}$. As $\|\theta - [\theta]\|_p \leq \frac{\epsilon}{B_x}$, we know, for any $x \in \mathbb{R}^d$ such that $\|x\|_q \leq B_x$, we have

$$|\theta^\top x - [\theta]^\top x| = |(\theta - [\theta])^\top x| \leq \|\theta - [\theta]\|_p \|x\|_q \leq \epsilon.$$

Therefore, for any $v = [\theta^\top x_1, \ldots, \theta^\top x_n] \in \mathcal{H}(S)$, letting $[v] = [[\theta]^\top x_1, \ldots, [\theta]^\top x_n]$, we have $\|v - [v]\|_\infty \leq \epsilon$.

Hence, $\mathcal{H}_{\frac{\epsilon}{B_x}}(S)$ is an ϵ-cover of $\mathcal{H}(S)$ in the $\|\cdot\|_\infty$ norm.

Example: Linear regression (cont'd)

So far we have shown,

$$\widehat{\mathcal{R}}(\mathcal{H}(S)) \leq \inf_{\epsilon>0} \left(\epsilon + B_x B_\theta \sqrt{\frac{2\log(N_\infty(\epsilon, \mathcal{H}(S)))}{n}} \right),$$

$$N_\infty(\epsilon, \mathcal{H}(S)) \leq |\mathcal{H}_{\frac{\epsilon}{B_x}}(S)| \leq \left(\frac{3B_\theta B_x}{\epsilon} \right)^d.$$

Therefore,

$$\widehat{\mathcal{R}}(\mathcal{H}(S)) \leq \inf_{\epsilon>0} \left(\epsilon + B_x B_\theta \sqrt{\frac{2d}{n} \log\left(\frac{3B_x B_\theta}{\epsilon} \right)} \right).$$

Choosing $\epsilon = B_x B_\theta \sqrt{\frac{d}{n}}$, we get

$$\mathcal{R}_n(\mathcal{H}) = \mathbb{E}_{S\sim P^n}\left[\widehat{\mathcal{R}}(\mathcal{H}(S)) \right] \in \mathcal{O}\left(\sqrt{\frac{d}{n} \log\left(\frac{n}{d} \right)} \right).$$

Why may our previous theorem be loose?

Theorem. Let $p \in [1, \infty]$ and $A \subset [-B, B]^n$. Then, $\widehat{\mathcal{R}}(A) \leq \inf_{\epsilon > 0} \left(\epsilon + B\sqrt{\frac{2 \log N_p(\epsilon, A)}{n}} \right)$.

Recall the following decomposition in the proof, which we have now written in terms of the empirical Rademacher complexity of a hypothesis class $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ on a set $S \subset \mathcal{X}^n$:

$$\begin{aligned}\widehat{\mathcal{R}}(\mathcal{H}, S) = \widehat{\mathcal{R}}(\mathcal{H}(S)) &= \frac{1}{n}\mathbb{E}_\sigma \left[\sup_{a \in \mathcal{H}(S)} \sigma^\top a \right] = \frac{1}{n}\mathbb{E}_\sigma \left[\sup_{a \in \mathcal{H}(S)} \sigma^\top (a - [a] + [a]) \right] \\ &\leq \frac{1}{n}\mathbb{E}_\sigma \left[\sup_{a \in \mathcal{H}(S)} \sigma^\top (a - [a]) \right] + \frac{1}{n}\mathbb{E}_\sigma \left[\sup_{a \in \mathcal{H}(S)} \sigma^\top [a] \right]\end{aligned}$$

Here, for an $a \in \mathcal{H}(S)$, $[a]$ denotes the closest element to a in the cover, satisfying $\frac{1}{n}\|a - [a]\|_1 \leq \epsilon$. Hence, $(a - [a])$ is a vector in the L1 ball of radius $n\epsilon$ centered at $[a]$.

We bound this term using Hölder's inequality $\sigma^\top (a - [a]) \leq \|\sigma\|_\infty \|a - [a]\|_1 \leq 1 \cdot (n\epsilon)$.

Why may our previous theorem be loose?

This bound is tight only if we can find an element in the intersection of $\mathcal{B}([a], n\epsilon, \|\cdot\|_1)$ and $\mathcal{H}(S)$ that is aligned[3] with every possible Rademacher random vector $\sigma \in \{-1, +1\}^n$. This may be true, if, say $\mathcal{B}([a], n\epsilon, \|\cdot\|_1) \cap \mathcal{H}(S) \approx \mathcal{B}([a], n\epsilon, \|\cdot\|_1)$.

However, for a given hypothesis class $\mathcal{H}$, $\mathcal{B}([a], n\epsilon, \|\cdot\|_1) \cap \mathcal{H}(S)$ may be a much smaller subset of $\mathcal{B}([a], n\epsilon, \|\cdot\|_1)$ which may cause the bound to be loose.

To tighten this bound, we can view the residual term $\frac{1}{n}\mathbb{E}_\sigma \left[\sup_{a \in \mathcal{H}(S)} \sigma^\top (a - [a])\right]$ as an empirical Rademacher complexity in and of itself. Hence, we can repeat the covering argument again with a smaller ϵ, and keep repeating for even smaller values of ϵ to bound the remaining residuals.

This idea is called "chaining" and leads to the Dudley entropy integral.

[3] Recall that in Hölder's inequality $|x^\top y| \leq \|x\|_\infty \|y\|_1$, equality holds when x, y are perfectly aligned, *i.e.*, for all $i \in [n]$, $x_i = c\ \mathrm{sgn}(y_i)$, for some constant $c \geq 0$.

Dudley Entropy Integral

Theorem. Let $A \subset \mathbb{R}^n$.Then, for all $p \in [2, \infty]$, we have

$$\widehat{\mathcal{R}}(A) \leq \inf_{\epsilon \geq 0} \left(4\epsilon + \frac{12}{\sqrt{n}} \int_{\epsilon}^{\infty} \sqrt{\log(N_p(t, A))}\, \mathrm{d}t \right).$$

Proof. Recall, in HW0, you showed $N_p(\epsilon, A) \leq N_q(\epsilon, A)$ for $p \leq q$. Therefore, it is sufficient to prove the above theorem with $N_2(\epsilon, A)$.

Let $\epsilon_0 = \sup_{a \in A} \max_{i \in [n]} |a_i|$. Note that, for all $a \in A$, we have $\|a\|_2 \leq \sqrt{n}\epsilon_0$.

Let $\epsilon_j = \frac{1}{2^j}\epsilon_0$. Let C_j be a minimal ϵ_j-cover of A in the $\frac{1}{\sqrt{n}}\|\cdot\|_2$ norm. For a given $a \in A$, let $[a]_j \in C_j$ be the closest element to a in the cover satisfying $\frac{1}{\sqrt{n}}\|a - [a]_j\|_2 \leq \epsilon_j$.

For notational convenience, let $[a]_0 = \mathbf{0}$ denote the zero vector.

Proof: Dudley Entropy Integral (cont'd)

Let $\epsilon > 0$ be given. Let m be such that $\epsilon_{m+2} \leq \epsilon \leq \epsilon_{m+1}$. Let us now decompose the empirical Rademacher complexity as follows:

$$\begin{aligned}
\widehat{\mathcal{R}}(A) &= \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{a\in A}\sigma^\top a\right] \\
&= \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{a\in A}\sigma^\top(a-[a]_m) + \sum_{j=1}^{m}\sigma^\top([a]_j - [a]_{j-1})\right] \\
&\leq \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{a\in A}\sigma^\top(a-[a]_m)\right] + \sum_{j=1}^{m}\frac{1}{n}\mathbb{E}_\sigma\left[\sup_{a\in A}\sigma^\top([a]_j - [a]_{j-1})\right] \qquad (5)
\end{aligned}$$

Using the Cauchy-Schwarz inequality, we have

$$\sigma^\top(a-[a]_m) \leq \underbrace{\|\sigma\|_2}_{\leq\sqrt{n}}\underbrace{\|a-[a]_m\|_2}_{\leq\epsilon_m\sqrt{n}} \leq \epsilon_m n.$$

Therefore, the first term of the RHS can be bound by ϵ_m.

Proof: Dudley Entropy Integral (cont'd)

Recall, Massart's Lemma Let $A \subset \mathbb{R}^n$ be finite. Then, $\widehat{\mathcal{R}}(A) \leq (\max_{a \in A} \|a\|_2) \cdot \frac{\sqrt{2\log(|A|)}}{n}$.

To bound the second term, let us write each term in the sum as

$$\frac{1}{n}\mathbb{E}_\sigma \left[\sup_{a \in A} \sigma^\top ([a]_j - [a]_{j-1})\right] = \frac{1}{n}\mathbb{E}_\sigma \left[\sup_{(b,b') \in B_j} \sigma^\top (b - b')\right] = \widehat{\mathcal{R}}(B_j),$$

where,

$$B_j = \Big\{ b - b'; (b, b') \in C_j \times C_{j-1}; \\ \exists\, a \in A \text{ s.t } \frac{1}{\sqrt{n}}\|b - a\|_2 \leq \epsilon_j \text{ and } \frac{1}{\sqrt{n}}\|b' - a\|_2 \leq \epsilon_{j-1} \Big\}$$

Therefore, by Massart's lemma,

$$\widehat{\mathcal{R}}(B_j) \leq \left(\max_{b,b' \in B_j} \|b - b'\|_2\right) \cdot \frac{\sqrt{2\log(|B_j|)}}{n}.$$

Proof: Dudley Entropy Integral (cont'd)

Let us first bound $\left(\max_{b,b' \in B_j} \|b - b'\|_2\right)$. For any $(b, b') \in B_j$, letting a be the element in A satisfying the condition, we have

$$\|b - b'\|_2 \leq \|b - a\|_2 + \|a - b'\|_2 \leq \sqrt{n}\epsilon_j + \sqrt{n}\epsilon_{j-1} = 3\sqrt{n}\epsilon_j.$$

The last step uses the fact that $\epsilon_i = \epsilon_0 2^{-i}$, so $\epsilon_{j-1} = 2\epsilon_j$.

Moreover, noting that $|C_\epsilon|$ is non-decreasing in ϵ we have,

$$|B_j| \leq |C_j||C_{j-1}| \leq |C_j|^2 = \left(N_2(\epsilon_j, A)\right)^2.$$

This gives us,

$$\begin{aligned}\widehat{\mathcal{R}}(B_j) &\leq 3\sqrt{n}\epsilon_j \cdot \frac{\sqrt{2\log\left(\left(N_2(\epsilon_j, A)\right)^2\right)}}{n} = 6\epsilon_j\sqrt{\frac{\log N_2(\epsilon_j, A))}{n}} \\ &= \frac{12}{\sqrt{n}}(\epsilon_j - \epsilon_{j+1})\sqrt{\log N_2(\epsilon_j, A)},\end{aligned}$$

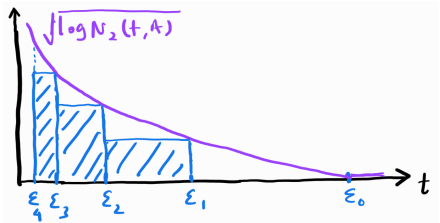
where, in the last step we have used the fact that $\epsilon_{j+1} = \epsilon_j/2$.

Proof: Dudley Entropy Integral (cont'd)

Plugging these back into (5), we obtain,

$$\begin{aligned}\widehat{\mathcal{R}}(A) &\leq \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{a\in A}\sigma^\top(a-[a]_m)\right] + \sum_{j=1}^{m}\frac{1}{n}\mathbb{E}_\sigma\left[\sup_{a\in A}\sigma^\top([a]_j-[a]_{j-1})\right] \\ &\leq \epsilon_m + \frac{12}{\sqrt{n}}\sum_{j=1}^{m}(\epsilon_j-\epsilon_{j+1})\sqrt{\log N_2(\epsilon_j,A)} \\ &\leq \epsilon_m + \frac{12}{\sqrt{n}}\int_{\epsilon_{m+1}}^{\epsilon_1}\sqrt{\log N_2(t,A))}\mathrm{d}t.\end{aligned}$$

The last step uses the fact that $\log(N_2(t,A))$ is non-increasing in t so we can bound the sum by an integral.



Proof: Dudley Entropy Integral (cont'd)

To complete the proof, we will make the following observations. Recall that we chose m so that

$$\frac{\epsilon_m}{4} = \frac{\epsilon_0}{2^{m+2}} = \epsilon_{m+2} \leq \epsilon \leq \epsilon_{m+1}.$$

By expanding the upper limit of the integral further to ∞, we get

$$\widehat{\mathcal{R}}(A) \leq 4\epsilon + \frac{12}{\sqrt{n}} \int_{\epsilon}^{\infty} \sqrt{\log N_2(t, A)} \mathrm{d}t.$$

This bound holds for any $\epsilon \geq 0$ so the theorem follows by taking the infimum.

Question: Why does this proof work only for $p \geq 2$?

Example: Linear regression revisited

Let us revisit the linear regression example and bound $\mathcal{R}_n(\mathcal{H})$ via the Dudley integral.

Setting. Let p, q be such that $\frac{1}{p} + \frac{1}{q} = 1$. Let $\mathcal{H} = \{h_\theta; h_\theta(x) = \theta^\top x; \theta \in \mathbb{R}^d, \|\theta\|_p \leq B_\theta\}$. Suppose that for all $X \sim P$, we have $X \in \mathbb{R}^d$ and $\|X\|_q \leq B_x$.

Recall that we have previously showed for any $S = \{x_1, \ldots, x_n\}$ with $\|x_i\|_q \leq B_x$, we have $\mathcal{H}(S) \subset \mathcal{B}_\infty(B_x B_\theta)$. Moreover, $N_\infty(\epsilon, \widehat{\mathcal{R}}(\mathcal{H}(S))) \leq \left(\frac{3B_x B_\theta}{\epsilon}\right)^d$ for all $\epsilon < B_x B_\theta$.

Let us apply the Dudley integral with $\epsilon = 0$ and the N_∞ covering number.

$$\begin{aligned}\widehat{\mathcal{R}}(\mathcal{H}(S)) &\leq \frac{12}{\sqrt{n}} \int_0^\infty \sqrt{\log N_\infty(t, \mathcal{H}(S))} \mathrm{d}t \\ &= \frac{12}{\sqrt{n}} \int_0^{B_x B_\theta} \sqrt{\log N_\infty(t, \mathcal{H}(S))} \mathrm{d}t\end{aligned}$$

where, we have used the fact that $N_\infty(t, \mathcal{B}_\infty(B_x B_\theta)) = 1$ for all $t \geq B_x B_\theta$ since $\mathcal{H}(S) \subset \mathcal{B}_\infty(B_x B_\theta)$.

Example: Linear regression revisited (cont'd)

Therefore,

$$\mathcal{R}_n(\mathcal{H}) \le \frac{12}{\sqrt{n}} \int_0^{B_x B_\theta} \sqrt{d \log\left(\frac{3B_x B_\theta}{t}\right)} \mathrm{d}t$$
$$\le 12\sqrt{\frac{d}{n}} \int_0^{3B_x B_\theta} \sqrt{\log\left(\frac{3B_x B_\theta}{t}\right)} \mathrm{d}t = 18\sqrt{\pi} B_x B_\theta \sqrt{\frac{d}{n}}.$$

The last step uses the fact that $\int_0^a \sqrt{\log(a/t)} \mathrm{d}t = a\sqrt{\pi}/2$ (see footnote[4]).

This result improves our previous $\sqrt{\frac{d}{n} \log\left(\frac{n}{d}\right)}$ bound by log factors. In HW3, you will see an example where the improvement will be polynomial.

[4] Let $I = \int_0^a \sqrt{\log \frac{a}{t}} \, \mathrm{d}t$. Let $u = \log \frac{a}{t}$ so $t = ae^{-u}$, $dt = -ae^{-u} du$. Then $t : 0 \to a$ gives $u : \infty \to 0$, hence $I = \int_0^a \sqrt{\log \frac{a}{t}} \, \mathrm{d}t = a \int_0^\infty e^{-u} \sqrt{u} \, \mathrm{d}u = a\,\Gamma\left(\frac{3}{2}\right) = \frac{a\sqrt{\pi}}{2}$.

Ch 3.5: Case study: Two-layer Neural Networks

Credit: Tengyu Ma

A two-layer neural network with ReLU activation can be characterized as follows:

- An input $x \in \mathbb{R}^d$.
- A hidden layer with m neurons. Let $W \in \mathbb{R}^{m \times d}$ be the weight matrix mapping inputs to the hidden layer.
- A ReLU activation function ϕ at each neuron of the hidden layer, where $\phi(x) = \max(x, 0)$, is applied pointwise.
- A linear layer at the output with parameters $\beta \in \mathbb{R}^m$.

Letting $\theta = (\beta, W)$ denote the parameters of the neural network, the output of the neural network is given by $h_\theta(x) = \beta^\top \phi(Wx)$.

We will consider the following hypothesis class,

$$\mathcal{H} = \left\{ h_\theta; h_\theta(x) = \beta^\top \phi(Wx);\ \|\beta\|_2 \le B_\theta,\ \|w_j\|_2 \le B_w\, \forall\, j \in [m] \right\}$$

Above $w_j^\top$ is the j^{th} row of W. We also assume that for $X \sim P$, we have $\|X\|_2 \le B_x$.

Two-layer neural networks (cont'd)

We will state two claims which you will prove in HW3.

Claim 1. Let $A \subset \mathbb{R}^n$ such that $\mathbf{0} \in A$. Let $\sigma \in \{-1, +1\}^n$ denote n independent Rademacher random variables. Then, $\mathbb{E}_\sigma \left[\sup_{a \in A} |\sigma^\top a|\right] \leq 2\mathbb{E}_\sigma \left[\sup_{a \in A} \sigma^\top a\right]$.

Claim 2. *(Data-dependent bounds for linear models under ℓ_2 constraints)* Let $\mathcal{H} = \{h_\theta; h_\theta(x) = \theta^\top x; \|\theta\|_2 \leq B_\theta\}$. Let $S = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ be given. Then, $\widehat{\mathcal{R}}(\mathcal{H}, S) \leq \frac{B_\theta}{n} \|X\|_F$, where $\|X\|_F^2 = \sum_{i=1}^n \sum_{j=1}^d x_{i,j}^2$ is the squared Frobenius norm of the data matrix X.

N.B. Suppose $\|X_i\|_2 \leq B_x$ for all $X_i \sim P$. Then, claim 2 may give us a tighter bound for $\mathcal{R}_n(\mathcal{H})$ than the $\sqrt{d/n}$ bound we obtained using the Dudley integral. In particular, $\|X\|_F^2 = \sum_{i=1}^n \|x_i\|_2^2 \leq B_x^2 n$. Hence, $\widehat{\mathcal{R}}(\mathcal{H}, S) \leq \frac{B_\theta B_x}{\sqrt{n}}$ which does not grow with d provided that B_x does not grow with d.

Two-layer neural networks (cont'd)

We will consider the following hypothesis class,

$$\mathcal{H} = \left\{ h_\theta; h_\theta(x) = \beta^\top \phi(Wx);\ \|\beta\|_2 \le B_\theta,\ \|w_j\|_2 \le B_w\, \forall\, j \in [m] \right\}$$

Above $w_j^\top$ is the j^{th} row of W. We also assume that for $X \sim P$, we have $\|X\|_2 \le B_x$.

Consider any $S = \{x_1, \ldots, x_n\}$ sampled from P^n. Then,

$$\begin{aligned}
\widehat{\mathcal{R}}(\mathcal{H}, S) &= \frac{1}{n}\mathbb{E}_\sigma \left[\sup_{h \in \mathcal{H}} \sum_{i=1}^n \sigma_i h_\theta(x_i) \right] = \frac{1}{n}\mathbb{E}_\sigma \left[\sup_{\beta, W} \sum_{i=1}^n \sigma_i \beta^\top \phi(Wx_i) \right] \\
&= \frac{1}{n}\mathbb{E}_\sigma \left[\sup_{\beta, W} \beta^\top \left(\sum_{i=1}^n \sigma_i \phi(Wx_i) \right) \right] \\
&= \frac{B_\beta}{n}\mathbb{E}_\sigma \left[\sup_{W} \left\| \sum_{i=1}^n \sigma_i \phi(Wx_i) \right\|_2 \right].
\end{aligned}$$

where the last step uses the fact that $\sup_{\|u\|_2 \le B} u^\top v = B\|v\|_2$.

Two-layer neural networks (cont'd)

Next, noting that $\|u\|_2 \leq \sqrt{m}\|u\|_\infty$ for any $u \in \mathbb{R}^m$, let us bound the term inside the expectation as follows:

$$\sup_{W;\|w_j\|_2 \leq B_w} \left\| \sum_{i=1}^n \sigma_i \phi(Wx_i) \right\|_2 \leq \sqrt{m} \sup_{W;\|w_j\|_2 \leq B_w} \max_{j \in [m]} \left| \sum_{i=1}^n \sigma_i \phi(w_j^\top x_i) \right|$$

$$= \sqrt{m} \sup_{\|w\|_2 \leq B_w} \left| \sum_{i=1}^n \sigma_i \phi(w^\top x_i) \right|$$

Therefore, by claim 1 (and noting that $\mathbf{0} \in \{w; \|w\|_2 \leq B_w\}$ which satisfies the conditions of the claim)

$$\widehat{\mathcal{R}}(\mathcal{H}, S) \leq \frac{\sqrt{m} B_\beta}{n} \mathbb{E}_\sigma \left[\sup_{\|w\|_2 \leq B_w} \left| \sum_{i=1}^n \sigma_i \phi(w^\top x_i) \right| \right]$$

$$\leq \frac{2\sqrt{m} B_\beta}{n} \mathbb{E}_\sigma \left[\sup_{\|w\|_2 \leq B_w} \sum_{i=1}^n \sigma_i \phi(w^\top x_i) \right] = 2B_\beta \sqrt{m} \widehat{\mathcal{R}}(\phi(\mathcal{H}_w^{\text{lin}}(S))).$$

where, $\mathcal{H}_w^{\text{lin}} = \{h_w; h_w(x) = w^\top x, \|w\|_2 \leq B_w\}$.

Two-layer neural networks (cont'd)

Recall, Claim 2. Let $\mathcal{H} = \{h_\theta; h_\theta(x) = \theta^\top x; \|\theta\|_2 \leq B_\theta\}$. Let $S = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ be given. Then, $\widehat{\mathcal{R}}(\mathcal{H}, S) \leq \frac{B_\theta}{n}\|X\|_F$, where $\|X\|_F^2 = \sum_{i=1}^n \sum_{j=1}^d x_{i,j}^2$ is the squared Frobenius norm of the data matrix X.

Recall, contraction lemma Let $S \subset \mathcal{Z}$ and $\mathcal{F} \subset \mathbb{R}^{\mathcal{Z}}$. For any K-Lipschitz $\psi : \mathbb{R} \to \mathbb{R}$, let $\psi \circ \mathcal{F} = \{\psi \circ f; f \in \mathcal{F}\}$. Then, $\widehat{\mathcal{R}}(\psi \circ \mathcal{F}(S)) \leq K\widehat{\mathcal{R}}(\mathcal{F}(S))$

We know that $\phi(x) = \max(x, 0)$ is 1-Lipschitz. Therefore, by the contraction lemma, claim 2, and noting that $\|x_i\|_2 \leq B_x$ for all $x \sim P$, we have,

$$\widehat{\mathcal{R}}(\phi(\mathcal{H}_w^{\text{lin}}(S))) \leq \widehat{\mathcal{R}}(\mathcal{H}_w^{\text{lin}}(S)) \leq \frac{B_w}{n}\|X\|_F \leq \frac{B_w B_x}{\sqrt{n}}.$$

Therefore $\widehat{\mathcal{R}}(\mathcal{H}, S) \leq 2B_\beta B_w B_x \sqrt{\frac{m}{n}}$, and

$$\mathcal{R}_n(\mathcal{H}) \leq \mathbb{E}_{S \sim P^n}\left[\widehat{\mathcal{R}}(\mathcal{H}, S)\right] \leq 2B_\beta B_w B_x \sqrt{\frac{m}{n}}.$$

Two-layer neural networks (cont'd)

In a two-layer neural network with m hidden layer neurons,

$$\mathcal{R}_n(\mathcal{H}) \leq \mathbb{E}_{S\sim P^n}\left[\widehat{\mathcal{R}}(\mathcal{H}, S)\right] \leq 2B_\beta B_w B_x \sqrt{\frac{m}{n}}.$$

This bound increases with m. But modern neural networks have very large hidden layers and their performance does not necessarily decrease with m.

In HW3, you will improve this result in two ways:

1. Prove a bound that does not explicitly depend on m.
2. Relax the bound assumption on X to $\mathbb{E}_P\left[\|X\|_2^2\right] \leq B_x^2$.

Ch 3.6: Approximation vs Estimation vs Optimization Error

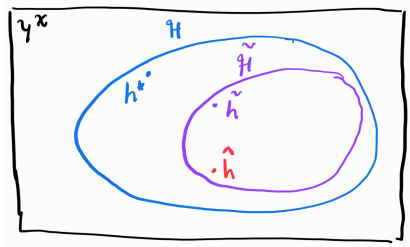
In supervised learning, we need to predict a label in $\mathcal{Y}$ from a given input $x \in \mathcal{X}$. In this chapter, we have framed this as finding a hypothesis h whose performance is competitive with the best hypothesis in a class $\mathcal{H}$ for all distributions P.
That is, we wish to design an algorithm $\widehat{h}$ to minimize,

$$\mathbb{E}_{S \sim P^n}\left[L(\widehat{h}(S)) - \inf_{h \in \mathcal{H}} L(h)\right]$$

Sometimes, instead of designing algorithms to optimize over $\mathcal{H}$, we may choose to optimize over some $\widetilde{\mathcal{H}} \subset \mathcal{H}$, which could lead to better generalization guarantees.

Example. If we perform ERM on $\widetilde{\mathcal{H}}$ instead of $\mathcal{H}$, we know $\mathcal{R}_n(\widetilde{\mathcal{H}}) \leq \mathcal{R}_n(\mathcal{H})$.

Approximation vs Estimation Error



Let $h^\star = \operatorname{argmin}_{h\in\mathcal{H}} L(h)$ and $\widetilde{h} = \operatorname{argmin}_{h\in\widetilde{\mathcal{H}}} L(h)$. We can write,

$$L(\widehat{h}) - L(h^\star) = \underbrace{L(\widehat{h}) - L(\widetilde{h})}_{\text{estimation error}} + \underbrace{L(\widetilde{h}) - L(h^\star)}_{\text{approximation error}}$$

Typically, if we choose a large $\widetilde{\mathcal{H}}$, the estimation error is large but the approximation error is small. A carefuly choice of $\widetilde{\mathcal{H}}$, possiby depending on the amount of data n, may be necessary to obtain a tight bound.

Example. For ERM, estimation error $\in \tilde{\mathcal{O}}(\mathcal{R}_n(\mathcal{H}) + n^{-1/2})$.

Optimization Error

Sometimes, our optimization algorithm to find $\widehat{h}$ may not always exactly maximize the given objective. For example, in ERM, let $\widehat{h} = \operatorname{argmin}_{h \in \widetilde{\mathcal{H}}} \widehat{L}(h)$ minimize the empirical loss, and let $\widehat{h}'$ be the hypothesis returned by our optimization algorithm. In such cases, we can decompose the error further as

$$L(\widehat{h}') - L(h^\star) = \underbrace{L(\widehat{h}') - L(\widehat{h})}_{\text{optimization error}} + \underbrace{L(\widehat{h}) - L(\widetilde{h})}_{\text{estimation error}} + \underbrace{L(\widetilde{h}) - L(h^\star)}_{\text{approximation error}}$$

Bayes' optimal predictor

Let $h^\star = \operatorname{argmin}_{h\in\mathcal{H}} L(h)$ and $\widetilde{h} = \operatorname{argmin}_{h\in\widetilde{\mathcal{H}}} L(h)$. We can write,

$$L(\widehat{h}) - L(h^\star) = \underbrace{L(\widehat{h}) - L(\widetilde{h})}_{\text{estimation error}} + \underbrace{L(\widetilde{h}) - L(h^\star)}_{\text{approximation error}}$$

When $\mathcal{H} = \mathcal{Y}^{\mathcal{X}}$, *i.e.*, $\mathcal{H}$ contains all possible predictors, then $h^\star$ is called the Bayes' optimal predictor.

Example 1. For binary classification with the $0-1$ loss, *i.e.*, $\ell(h,(x,y)) = \mathbb{1}(h(x)\neq y)$, the Bayes' optimal classifier $h^\star : \mathcal{X} \to \{0,1\}$ is given by

$$h^\star(x) = \operatorname*{argmax}_{y\in\mathcal{Y}} \mathbb{P}_{X,Y\sim P}(Y=y|X=x) = \begin{cases} 1 & \text{if } \mathbb{P}(Y=1|X=x) \geq 1/2. \\ 0 & \text{if } \mathbb{P}(Y=0|X=x) > 1/2. \end{cases}$$

Bayes' optimal predictor (cont'd)

Proof. Consider any hypothesis $h \in \{0,1\}^{\mathcal{X}}$. Then,

$$
\begin{aligned}
L(h) &= \mathbb{E}_{X,Y}\left[\mathbb{1}(h(X) \neq Y)\right] = \mathbb{E}_X[\mathbb{E}_Y\left[\mathbb{1}(h(X) \neq Y)|X\right]] \\
&= \mathbb{E}_X[\mathbb{E}_Y\left[\mathbb{1}(h(X) \neq 0) \wedge Y = 0|X\right] \; + \; \mathbb{E}_Y\left[\mathbb{1}(h(X) \neq 1) \wedge Y = 1|X\right]] \\
&= \mathbb{E}_X[\mathbb{E}_Y\big[\mathbb{1}(h(X) \neq 0)|Y = 0, X]\mathbb{P}(Y = 0|X) \\
&\qquad + \; \mathbb{E}_Y[\mathbb{1}(h(X) \neq 1)|Y = 1, X]\mathbb{P}(Y = 1|X)] \\
&= \mathbb{E}_X[\mathbb{1}(h(X) \neq 0)\mathbb{P}(Y = 0|X) \; + \; \mathbb{1}(h(X) \neq 1)\mathbb{P}(Y = 1|X)] \\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{as } h \text{ does not depend on } Y \\
&= \int_{\mathcal{X}} \underbrace{\big(\mathbb{1}(h(x) = 1)\mathbb{P}(Y = 0|X = x) + \mathbb{1}(h(x) = 0)\mathbb{P}(Y = 1|X = x)\big)}_{(*)} \mathrm{d}P_X(x)
\end{aligned}
$$

Bayes' optimal predictor (cont'd)

We see that $(*)$ is minimized pointwise by choosing

$$h(x) = 1 \quad \text{if} \quad \mathbb{P}(Y = 0|X = x) \leq \mathbb{P}(Y = 1|X = x) \iff \mathbb{P}(Y = 1|X = x) \geq 1/2,$$
$$h(x) = 0 \quad \text{if} \quad \mathbb{P}(Y = 0|X = x) \geq \mathbb{P}(Y = 1|X = x) \iff \mathbb{P}(Y = 0|X = x) \geq 1/2,$$

This is precisely $h^\star$. Therefore, $L(h) \leq L(h^\star)$.

Example 2 (try at home). For regression problems with the squared loss, *i.e.*, $\ell(h, (x, y)) = (h(x) - y)^2$, the Bayes' optimal predictor $h^\star$ is the regression function.. That is, $h^\star(\cdot) = \mathbb{E}_{X,Y \sim P}[Y|X = \cdot]$.