1 — Mathematical Foundations of Learning

Daniel Khashabi [1]
KHASHAB2@ILLINOIS.EDU

1.1 Introduction

The *Statistical Learning Theory*, is somewhat least theoretical part of Machine Learning, which is most about the theoretical guarantees in learning concepts, in different conditions and scenarios. These guarantees are usually expressed in the form of probabilistic concentration of some measure, around some optimal value which is unknown and need to discovered. These bounds are functions of problem specifications, for example,

- The number of samples: the more samples we have, there is a better chance of learning.
- The easiness of the underlying distribution need to be learnt.
- The generalization power (or flexibility) of the family of the functions which is being used to approximate the target distribution.

1.2 Concentration Inequalities

To measure up how the modelling using samples is close to the original (unknown) system, researchers have developed various. One of the ways to quantify this is to use *probabilistic* approach for modelling the *concentration* (closeness) of the model to the original system. Such analysis need using various probabilistic inequalities that could bound the results.

1.2.1 Markov's inequality; the most basic bound

[1] This is part of my notes; to find the complete list of notes visit `http://web.engr.illinois.edu/~khashab2/learn.html`. This document is updated on December 25, 2013.

Theorem 1.1 If X is a non-negative random variable, for any $t > 0$ we have

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}X}{t}$$

Proof.

$$\begin{aligned}
\mathbb{E}X &= \int_{X\in\mathcal{X}} P_X(dx) \\
&= \int_{X\in\mathcal{X}, X<t} XP_X(dx) + \int_{X\in\mathcal{X}, X\geq t} XP_X(dx) \\
&\geq \int_{X\in\mathcal{X}, X\geq t} XP_X(dx) \\
&\geq \int_{X\in\mathcal{X}, X\geq t} tP_X(dx) = t\int_{X\in\mathcal{X}, X\geq t} P_X(dx) = t\mathbb{P}(X \geq t)
\end{aligned}$$

■

1.2.2 Chebyshev's inequality; generalizing the Markov's inequality

Theorem 1.2

$$\mathbb{P}\left(|X - \mathbb{E}X| \geq t\right) \leq \frac{\mathrm{Var}X}{t^2}$$

Proof. There is a simple proof using Markov's inequality. Based Markov's inequality we know,

$$\mathbb{P}(X \geq t) = \mathbb{P}(\phi(X) \geq \phi(t)) \leq \frac{\mathbb{E}\phi(X)}{\phi(t)}$$

Now assume $\phi(x) = x^2 (x \geq 0)$, which gives the desired result. In other words, using the Markov's inequality

$$\begin{aligned}
\mathbb{P}\left(|X - \mathbb{E}X| \geq t\right) &= \mathbb{P}\left(|X - \mathbb{E}X|^2 \geq t^2\right) \\
&\leq \frac{\mathbb{E}\left|X - \mathbb{E}X\right|^2}{t^2} = \frac{\mathbb{E}X^2 - (\mathbb{E}X)^2}{t^2} = \frac{\mathrm{Var}X}{t^2}
\end{aligned}$$

■

R In general $\phi(x) = x^q (x \geq 0)$, for any positive q we have,

$$\mathbb{P}\left(|X - \mathbb{E}X| \geq t\right) \leq \frac{\mathbb{E}\left|X - \mathbb{E}X\right|^q}{t^q}$$

Because the goal is to find tighter upper-bounds one can minimize the right-side with respect to q.

1.2.3 Chernoff's Trick

For any $s > 0$,

$$\mathbb{P}(X > t) = \mathbb{P}(e^{sX} > e^{st}) \leq e^{-st}\mathbb{E}\left[e^{sX}\right]$$

To make the bound as tight as possible,

$$\mathbb{P}(X > t) \leq \inf_{s>0} e^{-st}\mathbb{E}\left[e^{sX}\right]$$

1.2.4 Hoeffding's inequality; a special case

■ **Lemma 1.1** If $X \in \mathcal{X}$ is a random variable with $\mathbb{E}X = 0$, and $\exists a, b \in \mathcal{X}$, s.t. $\mathbb{P}(a \leq X \leq b) = 1$, then for any $s > 0$

$$\mathbb{E}\left[e^{sX}\right] \leq e^{\frac{1}{8}s^2(b-a)^2}$$

Proof. Assume any point $x \in [a, b]$, which can be represented as $x = \beta.b + (1 - \beta).a, 0 \leq \beta \leq 1$. Let us assume a function $\phi(x) = e^{sx}$ for any $s > 0$ which is a convex function on $\mathcal{X} = \mathbb{R}$. Then we have,

$$e^{sx} \leq \beta e^{sb} + (1 - \beta)e^{sa}$$

By replacing β with $\frac{x-a}{b-a}$, and since $\mathbb{E}X = 0$, we have,

$$\mathbb{E}\left[e^{sx}\right] \leq \frac{b}{b-a}e^{sa} - \frac{a}{b-a}e^{sb}$$

Now set $p = \frac{a}{a-b}$, and thus $1 - p = \frac{b}{b-a}$. We continue the previous expression:

$$\begin{aligned}\mathbb{E}\left[e^{sx}\right] &\leq \frac{b}{b-a}e^{sa} - \frac{a}{b-a}e^{sb} \\ &\leq (1-p)e^{sa} + pe^{sb} = e^{sa}\left(1 - p + pe^{s(b-a)}\right)\end{aligned}$$

Define $u \triangleq s(b-a)$ and define $\Phi(u) \triangleq= -pu + \log(1 - p + pe^u)$. Then we have

$$\mathbb{E}e^{sX} \leq e^{\Phi(u)}$$

which holds for any value of u, for $p \in [0, 1]$. One show that the $Phi(u)$ is upper-bounded by $\frac{1}{8}u^2 = \frac{1}{8}s^2(b-a)^2$, which proves the desired result. ■

Theorem 1.3 — Hoeffding's inequality. Let $X_1, \ldots, X_n$ be independent random variables, and for any X_i, $\exists a_i, b_i \in \mathcal{X} = \mathbb{R}$, s.t. $\mathbb{P}(a_i \leq X \leq b_i) = 1$. Let $S_n \triangleq \sum_{i=1}^n X_i$ then for any $s > 0$,

$$\mathbb{P}\left(|S_n - \mathbb{E}S_n| \geq t\right) \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

Proof. We can divide the inequality into two parts,

$$\mathbb{P}\left(S_n - \mathbb{E}S_n \geq t\right) \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right) \tag{1.1}$$

$$\mathbb{P}\left(S_n - \mathbb{E}S_n \leq -t\right) \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right) \tag{1.2}$$

To use the Lemma 1.1, we can replace each variable $X_i \to X_i - \mathbb{E}X_i$, so that $\mathbb{E}X_i = 0$; then we have, $\mathbb{E}\left[e^{sX_i}\right] \leq e^{s^2(b_i-a_i)^2/8}$. Using Chernoff's trick,

$$\mathbb{P}\left(S_n \geq t\right) = \mathbb{P}\left(e^{S_n} \geq e^t\right) \leq e^{-st}\mathbb{E}\left[e^{sS_n}\right]$$

And since X_i's are independent,

$$\mathbb{E}\left[e^{sS_n}\right] = \mathbb{E}\left[e^{s(X_1+\ldots+X_n)}\right] = \mathbb{E}\left[\prod_{i=1}^{n} e^{sX_i}\right] = \prod_{i=1}^{n} \mathbb{E}\left[e^{sX_i}\right]$$

$$\Rightarrow \mathbb{P}\left(S_n \geq t\right) \leq e^{-st} \prod_{i=1}^{n} e^{s^2(b_i-a_i)^2/8} = \exp\left\{-st + \frac{s^2}{8}\sum_{i=1}^{n}(b_i-a_i)^2\right\}$$

To minimize the right hand-side with respect to s we can choose it to be $s = \frac{4t}{\sum_{i=1}^{n}(b_i-a_i)^2}$. This proves the Equation 1.1. The proof for the Equation 1.2 is similar. ■

1.2.5 McDiarmid's inequality; generalizing the bounded variables to bounded differences

Definition 1.1 — A function with bounded differences. A function is said to have bounded differences, if changing only one variable, and keep everything the save, the absolute value of the difference is always bounded. In other words, if we have a function $f : \mathcal{X}^n \to \mathbb{R}$, then,

$$\exists c_i \in \mathbb{R}, \text{ s.t. } \forall x_1, \ldots, x_n, x_i' \in \mathbb{X}, \sup\left|f(x_1, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)\right| \leq c_i$$

Theorem 1.4 — McDiarmid's inequality. Let $X = (X_1, \ldots, X_n) \in \mathcal{X}^n$, and $f : \mathcal{X}^n \to \mathbb{R}$ has bounded differences. Then for any $t > 0$,

$$\mathbb{P}\left(\left|f(X) - \mathbb{E}f(X)\right| \geq t\right) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n} c_i^2}\right)$$

TBW. ■

■ **Example 1.1** 1. Given a real-valued random variable Z, such that,

$$\log \mathbb{E}\left[e^{sZ}\right] \leq \frac{vs^2}{2\left(1 - cs\right)}, \tag{1.3}$$

for some $v, c \in \mathbb{R}^+$ and every $s \in [0, \frac{1}{c}]$.
We know $\{X_i\}_{i=1}^n$ are i.i.d. and $X_i \sim \mathcal{N}(0, 1)$. Then $U = \sum_{i=1}^n X_i^2$ with $U \sim \chi_n^2$, we want to prove that

$$\mathbb{P}\left(Z \geq \sqrt{2vt} + ct\right) \leq e^{-t} \tag{1.4}$$

for all $t > 0$.

Proof. Based on the Markov inequality we know that we have,

$$\mathbb{P}\left(X \geq t\right) \leq \frac{\mathbb{E}\left[X\right]}{t}.$$

We apply the Chernoff trick to find better (smaller=tighter) upper bound

for our desired upperbound:

$$\mathbb{P}\left(X \geq a\right) = \mathbb{P}\left(e^{sX} \geq e^{sa}\right) \leq \frac{\mathbb{E}\left[e^{sX}\right]}{e^{sa}} = e^{-sa}\mathbb{E}\left[e^{sX}\right], \quad \forall s > 0. \tag{1.5}$$

$$\mathbb{P}\left(X \geq a\right) \leq \inf_{s\in\mathbb{R}^+} e^{-sa}\mathbb{E}\left[e^{sX}\right]. \tag{1.6}$$

From the Equation 1.4, and combining it with the Equation 1.6 we have,

$$\begin{aligned}\mathbb{P}\left(X \geq a\right) \leq \inf_{s\in\mathbb{R}^+} e^{-sa}\mathbb{E}\left[e^{sX}\right] &\leq \inf_{s\in\mathbb{R}^+}\left\{e^{-sa}\exp\left[\frac{vs^2}{2\left(1-cs\right)}\right]\right\} \\ &= \inf_{s\in\mathbb{R}^+}\left\{\exp\left[-sa+\frac{vs^2}{2\left(1-cs\right)}\right]\right\}.\end{aligned}$$

The minimizer of the right term could be found:

$$\frac{d}{ds}\left[-sa+\frac{vs^2}{2\left(1-cs\right)}\right] = -a+\frac{2vs\left(1-cs\right)+cvs^2}{2\left(1-cs\right)^2} = 0$$

$$\Rightarrow \mathbb{P}\left(X \geq a\right) \leq e^{-sa}\mathbb{E}\left[e^{sX}\right] \leq \left\{e^{-sa}\exp\left[\frac{vs^2}{2\left(1-cs\right)}\right]\right\}.$$

Since the minimization could be cumbersome, we do reverse engineering using the modified objective,

$$\begin{aligned}\mathbb{P}\left(Z \geq \sqrt{2vt}+ct\right) &= \mathbb{P}\left(e^{sZ} \geq e^{s\left(\sqrt{2vt}+ct\right)}\right) \\ &\leq \inf_{s\in\mathbb{R}^+}\frac{\mathbb{E}e^{sZ}}{e^{s\left(\sqrt{2vt}+ct\right)}} \\ &\leq \inf_{s\in\mathbb{R}^+}\exp\left\{\frac{vs^2}{2\left(1-cs\right)}-s\left(\sqrt{2vt}+ct\right)\right\}\end{aligned}$$

Now we just need to show that,

$$\inf_{s\in\mathbb{R}^+}\exp\left\{\frac{vs^2}{2\left(1-cs\right)}-s\left(\sqrt{2vt}+ct\right)\right\} \leq e^{-t}$$

We can make it looser and show that,

$$\exists s \in \mathbb{R}^+ \quad s.t. \quad \exp\left\{\frac{vs^2}{2\left(1-cs\right)}-s\left(\sqrt{2vt}+ct\right)\right\} = e^{-t}$$

$$\Rightarrow \exists s \in \mathbb{R}^+ \quad s.t. \quad \exp\left\{\frac{vs^2}{2\left(1-cs\right)}-s\left(\sqrt{2vt}+ct\right)+1\right\} = 1$$

$$\Rightarrow \exists s \in \mathbb{R}^+ \quad s.t. \quad \frac{vs^2}{2\left(1-cs\right)}-s\left(\sqrt{2vt}+ct\right)+1 = 0$$

The expression above could be turned into a nice closed form:

$$\begin{aligned}\frac{vs^2}{2\left(1-cs\right)}-s\left(\sqrt{2vt}+ct\right)+1 &= \frac{vs^2-2s\left(\sqrt{2vt}+ct\right)\left(1-cs\right)+2\left(1-cs\right)}{2\left(1-cs\right)} \\ &= \frac{vs^2-2s\sqrt{2vt}\left(1-cs\right)+2\left(1-cs\right)^2}{2\left(1-cs\right)} \\ &= \frac{\left(s\sqrt{v}-\left(1-cs\right)\sqrt{2t}\right)^2}{2\left(1-cs\right)} = 0\end{aligned}$$

$$\Rightarrow s = \frac{\sqrt{2t}}{c\sqrt{2t} - \sqrt{v}}$$

This holds only when $s \in [0, \frac{1}{c}]$, since we used Equation 1.3. Since $v, t \in \mathbb{R}^+$, for any $t \geq \frac{v}{2c^2}$, s belongs to $[0, \frac{1}{c}]$ and this bound holds. ■

2. Now we want to prove that,

$$\mathbb{P}\left(U - n \geq 2\sqrt{nt} + 2t\right) \leq e^{-t}.$$

Before proving the desired claim, we first a lemma which is will become handy in the proof.

■ **Lemma 1.2** The following holds for any $s \in (0, 1/2)$

$$-s - \frac{1}{2}\log(1-2s) \leq \frac{s^2}{1-2s}, \quad 0 < s < 1/2 \tag{1.7}$$

Proof. Let's label each side of the inequality,

$$\begin{cases} A(s) = -s - \frac{1}{2}\log(1-2s) \\ B(s) = \frac{s^2}{1-2s} \end{cases}$$

We prove this inequality in two steps,

- $A(s) = B(s)$, for $s = 0$:
  This is easy to show that $A(0) = B(0) = 0$
- $\frac{dA(s)}{ds} \leq \frac{dB(s)}{ds}, \forall s \in (0, 1/2)$:

$$\frac{dA(s)}{ds} = -1 - \frac{1}{2}\frac{-2}{1-2s} = \frac{2s}{1-2s}$$
$$\frac{dB(s)}{ds} = \frac{2s(1-s)}{(1-2s)^2}$$
$$\frac{dB(s)}{ds} - \frac{dA(s)}{ds} = \frac{2s(1-s)}{(1-2s)^2} - \frac{2s}{1-2s} = \frac{2s^2}{(1-2s)^2} \geq 0$$

This gives us the desired result. ■

Proof. We first prove that the inequality in Equation 1.3 holds, when we

choose $Z = U - n = \sum_{i=1}^{n} X_i^2 - n$,

$$\begin{aligned}
\mathbb{E}\left[e^{sZ}\right] &= \int_{\mathbf{x}\in\mathcal{X}} e^{s\left(\sum_{i=1}^{n} x_i^2 - n\right)} (2\pi)^{-k/2} \left|\mathbf{\Sigma}\right|^{-1/2} e^{-\frac{1}{2}(\mathbf{x}-\mu)^\top \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)} d\mathbf{x} \\
&= e^{-sn} \prod_{i=1}^{n} \int_{x_j} e^{sx_i^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{x_i^2}{2}} dx_i \\
&= e^{-sn} \prod_{i=1}^{n} \int_{x_j} \frac{1}{\sqrt{2\pi}} e^{-x_i^2\left(\frac{1}{2}-s\right)} dx_i \\
&= e^{-sn} \prod_{i=1}^{n} \sqrt{\frac{1}{1-2s}} \\
&= e^{-sn} \left(1-2s\right)^{-n/2}
\end{aligned}$$

Taking logarithm of both sides, and using the result of Lemma 1.2,

$$\log \mathbb{E}\left[e^{sZ}\right] = n\left(-s - \frac{1}{2}\left(1-2s\right)\right) \leq n \times \frac{s^2}{1-2s}, \quad 0 < s < 1/2$$

Which is of the form given in the equation 1.3, when $v = n$ and $c = 2$. Given this, we can use the bound we found in part (a), with replacing the values $v = n$ and $c = 2$, which gives us the desired result. ■

■

■ **Example 1.2** We consider we have a set of i.i.d. random variables $\{x_i\}_{i=1}^n$, and $x_i \sim \text{Binomial}(\theta)$. Clearly if $S \triangleq \sum_{i=1}^{n} x_i$,

$$S \sim \text{B}(n,\theta) \Rightarrow \begin{cases} \mathbb{E}[S] = n\theta, \\ \mathbb{V}[S] = n\theta(1-\theta) \end{cases}$$

Now we prove each part.

1. First we prove the following holds

$$\mathbb{P}(S \geq n\alpha) \leq e^{-n \times d(\alpha||\theta)}, \ \forall \alpha \in [\theta, 1] \tag{1.8}$$

where, $d(\alpha||\theta) = \alpha \log \frac{\alpha}{\theta} + (1-\alpha) \log \frac{(1-\alpha)}{(1-\theta)}$ is the KL-divergence between α and θ, two Bernoulli random variables.

Proof. Based on the Markov inequality we know that,

$$\mathbb{P}(S \geq n\alpha) \leq \inf_{\lambda \in \mathbb{R}^+} \left[e^{-\lambda n\alpha} \mathbb{E}_{S\sim \text{B}(n,\theta)} e^{S\lambda}\right] \tag{1.9}$$

It is easy to show that the Moment Generating Function (MGF) for a binomial distribution is as following,

$$M_X(t) = \mathbb{E}_{X\sim \text{Binomial}(n,\theta)} e^{tX} = (1 - \theta + \theta e^t)^n$$

Using MGF formula, we can simplify our bound in Equation 1.9.

$$\mathbb{P}(S \geq n\alpha) \leq \inf_{\lambda \in \mathbb{R}^+} \left[e^{-\lambda n\alpha}(1-\theta+\theta e^{\lambda})^n\right] = \inf_{\lambda \in \mathbb{R}^+} \left[\left((1-\theta)e^{-\lambda\alpha} + \theta e^{\lambda(1-\alpha)}\right)^n\right]$$
$$= \left(\inf_{\lambda \in \mathbb{R}^+} \left[(1-\theta)e^{-\lambda\alpha} + \theta e^{\lambda(1-\alpha)}\right]\right)^n$$

We define $A(\lambda) = (1-\theta)e^{-\lambda\alpha} + \theta e^{\lambda(1-\alpha)}$, and find minimizer of $A(\lambda)$,

$$\frac{\partial A}{\partial \lambda} = -\alpha(1-\theta)e^{-\lambda\alpha} + (1-\alpha)\theta e^{\lambda(1-\alpha)} = 0 \Rightarrow \lambda = \ln\frac{\alpha(1-\theta)}{(1-\alpha)\theta}$$

$$\Rightarrow A(\lambda) = (1-\theta)e^{-\lambda\alpha} + \theta e^{\lambda(1-\alpha)} = (1-\theta)e^{-\lambda\alpha}\left(1 + \frac{\theta}{1-\theta}e^{\lambda}\right)$$

$$\Rightarrow A(\lambda)\big|_{\lambda=\ln\frac{\alpha(1-\theta)}{(1-\alpha)\theta}} = (1-\theta)e^{-\lambda\alpha}\left(1 + \frac{\theta}{1-\theta}e^{\lambda}\right)\Big|_{\lambda=\ln\frac{\alpha(1-\theta)}{(1-\alpha)\theta}}$$
$$= \left(\frac{\theta}{\alpha}\right)^{\alpha}\left(\frac{1-\theta}{1-\alpha}\right)^{1-\alpha}$$
$$= -d(\alpha||\theta)$$

$$\Rightarrow \mathbb{P}(S \geq n\alpha) \leq e^{-nd(\alpha||\theta)}$$

■

2. Now we prove the bound in Equation 1.8 is indeed tighter than,

$$\mathbb{P}(S \geq n\alpha) \leq e^{-2n(\alpha-\theta)^2},\ \forall\alpha \in [\theta, 1]$$

Proof. We want to show,

$$e^{-n\times d(\alpha||\theta)} \leq e^{-2n(\alpha-\theta)^2},\ \forall\alpha \in [\theta, 1]$$

Or we want to show,

$$d(\alpha||\theta) \geq 2(\alpha-\theta)^2,\ \forall\alpha \in [\theta, 1]$$

First note that for $\alpha = \theta$ both bounds are the same, as they are both zero. Then if we show that the derivative of $d(\alpha||\theta)$ is always greater than $2(\alpha-\theta)^2$, this would imply that $d(\alpha||\theta) \geq 2(\alpha-\theta)^2$, for all $\alpha \in [\theta, 1]$. Equivalently, we show that $\frac{\partial\lambda}{\partial\alpha} \geq 0$ for $\theta \geq \alpha \geq 1$, where

$$\lambda = d(\alpha||\theta) - 2(\alpha-\theta)^2$$

$$\frac{\partial\lambda}{\partial\alpha} = \ln\frac{\alpha}{\theta} + 1 + \ln\frac{1-\alpha}{1-\theta} - 1$$

To show that $\frac{\partial\lambda}{\partial\alpha} \geq 0$, we do the same trick; since $\frac{\partial\lambda}{\partial\alpha}\big|_{\alpha=\theta} = 0$, we just need to show that, $\frac{\partial^2\lambda}{\partial\alpha^2} \geq 0$. Since $\frac{\partial^2\lambda}{\partial\alpha^2} = \frac{1}{\alpha(1-\alpha)} - 4$, and using the Arithmetic-Geometric inequality[a], we have

$$\alpha(1-\alpha) \geq 4$$

Then $\frac{\partial^2\lambda}{\partial\alpha^2} \geq 0$ which finishes our proof. ■

■

[a] $\frac{x+y}{2} \geq \sqrt{xy} \Rightarrow \frac{\alpha+(1-\alpha)}{2} \geq \sqrt{(1-\alpha)\alpha} \Rightarrow \alpha(1-\alpha) \geq 4$

■ **Example 1.3** We assume having i.i.d. random variables, $\{Y_i\}_{i=1}^k$. For a given real number $y \in \mathbb{R}$,

$$N = |\{1 \leq j \leq k : Y_j \geq y\}| . \tag{1.10}$$

1. Considering, $\max_{1 \leq j \leq k} \mathbb{P}(Y_j \geq y) \leq p$, we want to show that there is exists a pair $(\tilde{N}, \tilde{S})$ of jointly distributed integer-valued random variables such that
   - $\tilde{N}$ has the same distribution as N.
   - $\tilde{S}$ has the Binomial(k, p).
   - $\tilde{N} \leq \tilde{S}$.

   Proof. In practice the original random variables $\{Y_i\}_{i=1}^k$, could be distributed with any arbitrary distribution. But we can easily replace them with any arbitrary distribution, and come up with the same distribution for Equation 1.10, as long as it satisfies certain conditions. To make everything simple, we use uniform distribution, and we define the set of uniform i.i.d. random variables $\{U_i\}_{i=1}^k$, such that for a fixed y,

   $$\mathbb{P}(U_j \geq y) = \mathbb{P}(Y_j \geq y), \quad \forall j \in \{1, \ldots, k\} \tag{1.11}$$

   Now we define a modified version of the Equation 1.10 for the U_j random variables,

   $$\tilde{N} = |\{1 \leq j \leq k : U_j \geq y\}| . \tag{1.12}$$

   Referring to the conditions in the Equation 1.11, it is easy to see that $\tilde{N}$ and N have the same distributions, or in other words,

   $$\mathbb{P}(\tilde{N} \geq t) = \mathbb{P}(N \geq t), \ \forall t \in \mathbb{N} \cup \{0\}$$

   In a similar way, we define the set of i.i.d. random variables $\{V_i\}_{i=1}^k$, with the condition that

   $$\mathbb{P}(V_j \geq y) = p, \quad \forall j \in \{1, \ldots, k\} \tag{1.13}$$

   Similar to $\tilde{N}$ we define a modified version of the Equation 1.10 for the V_j random variables,

   $$\tilde{S} = |\{1 \leq j \leq k : V_j \geq y\}| . \tag{1.14}$$

   Based on the condition in Equation 1.13, since

   $$\mathbb{P}(Y_j \geq y) = \mathbb{P}(U_j \geq y) \leq p = \mathbb{P}(V_j \geq y)$$

   it is clear that,

   $$\mathbb{P}(\tilde{S} \geq t) \geq \mathbb{P}(\tilde{N} \geq t), \ \forall t \in \mathbb{N} \cup \{0\} \tag{1.15}$$

   Also it is easy to see that $\tilde{S} \sim$ Binomial(k, p). To show this, we define

   $$\tilde{S}_j = \mathbf{1}\{V_j \geq y\} .$$

Each $\tilde{S}_j$ is distributed with the Bernoulli, with $\mathbb{P}(\tilde{S}_j = 1) = \mathbb{P}(V_j \geq y) = p$. Clearly $\tilde{S} = \sum_{j=1}^{k} \tilde{S}_j$, and since each $\tilde{S}_j \sim \text{Bernoulli}(p)$,

$$S \sim \text{Binomial}(k, p) \tag{1.16}$$

■

2. Assuming that $T_{k,p}(t)$ is the probability that Binomial(k, p) is a random variable greater or equal to t, we want to show that

$$\mathbb{P}(N \geq t) \leq T_{k,p}(t) \tag{1.17}$$

Proof. Referring back to Equation 1.15, we have,

$$\mathbb{P}(\tilde{S} \geq t) \geq \mathbb{P}(\tilde{N} \geq t), \ \forall t \in \mathbb{N} \cup \{0\}$$

And based on the result in Equation 1.16, it gives us the desired answer. ■

■

1.3 Rademacher Averages

Here we define Rademacher random variables which we will use in measuring complexities of class of functions.

Definition 1.2 — Rademacher Average. If $\mathcal{F} \subset \{f : \mathcal{X} \to \mathbb{R}\}$ be a class of functions we are exploring defined on domain $X \subset \mathcal{X}$, and $\{X_i\}_i^n$ be the set of samples generated by some unknown distribution $\mathbb{P}$ on the same domain X. Define σ_i to be uniform random variable on ± 1, for any i. The empirical Rademacher average or complexity is defined as following:

$$\hat{R}_n(\mathcal{F}) = \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(X_i) \right| \Big| \{X_i\}_i^n \right]$$

and the expectation of the above measure, with respect to the random samples, is called the Rademacher average or complexity:

$$R_n(\mathcal{F}) = \mathbb{E}\hat{R}_n(\mathcal{F}) = \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(X_i) \right| \right]$$

■ **Lemma 1.3** Given real-valued CDF function, and $\mathcal{F}$ being class of indicator functions on half-intervals, and

$$F(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq x\}}$$

we can show that,

$$\mathbb{E}\left[\sup_{x\in\mathbb{R}}\left|\hat{F}_n(x)-F_n(x)\right|\right]\leq 2\mathbb{E}R_n(\mathcal{F}(Z^n)),\qquad \text{for some } C>0$$

Proof.

$$\begin{aligned}\sup_{x\in\mathbb{R}}\left|\hat{F}_n(x)-F(x)\right| &= \sup_{x\in\mathbb{R}}\left|\hat{F}_n(x)-\bar{\mathbb{E}}\bar{F}(x)\right| \\ &\leq \sup_{x\in\mathbb{R}}\bar{\mathbb{E}}\left|\hat{F}_n(x)-\bar{F}(x)\right| \\ &\leq \bar{\mathbb{E}}\sup_{x\in\mathbb{R}}\left|\hat{F}_n(x)-\bar{F}(x)\right|\end{aligned}$$

$$\mathbb{E}\left[\sup_{x\in\mathbb{R}}\left|\hat{F}_n(x)-F(x)\right|\right]\leq \bar{\mathbb{E}}\sup_{x\in\mathbb{R}}\left|\hat{F}_n(x)-\bar{F}(x)\right|$$

We know,

$$\frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{1}_{\{X_i\leq x\}}-\mathbf{1}_{\left\{\bar{X}_i\leq x\right\}}\right]\overset{d}{=}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\left[\mathbf{1}_{\{X_i\leq x\}}-\mathbf{1}_{\left\{\bar{X}_i\leq x\right\}}\right]$$

$$\begin{aligned}\mathbb{E}\left[\sup_{x\in\mathbb{R}}\left|\hat{F}_n(x)-F(x)\right|\right] &\leq \bar{\mathbb{E}}\sup_{x\in\mathbb{R}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\left[\mathbf{1}_{\{X_i\leq x\}}-\mathbf{1}_{\left\{\bar{X}_i\leq x\right\}}\right]\right| \\ &\leq 2\bar{\mathbb{E}}\sup_{x\in\mathbb{R}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\mathbf{1}_{\{X_i\leq x\}}\right| = 2\mathbb{E}R_n(\mathcal{F}),\ \text{for } \mathcal{F}=\text{half-intervals}\end{aligned}$$

■

The following bounding technique could be generalized to any risk function.

■ **Lemma 1.4 — Symmetrization trick for bounding general risk with Rademacher average.** Given a class functions $\mathcal{F}\subset\{f:\mathcal{X}\to\mathbb{R}\}$ defined on domain $X\subset\mathcal{X}$, we have the following general bound on the Rademacher average:

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\mathbb{E}f-\frac{1}{n}\sum_{i=1}^{n}f(X_i)\right|\right]\leq 2R_n(\mathcal{F})$$

where

$$R_n(\mathcal{F})=\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(X_i)\right|\right]$$

Proof. The steps for the previous proof hold for this proof, with some minor changes. ■

■ **Example 1.4** Let $f : \mathsf{X} \to \{0, 1\}$, and let $(X, Y) \in \mathsf{X} \times \{0, 1\}$ be n random i.i.d. samplings from the joint distribution P_{XY}. Consider the empirical risk defined as,

$$L_n(f) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{f(X_i) \neq Y_i\}$$

1. Prove that for any $f \in \mathcal{F}$,

$$L(f) \leq L_n(f) + \sqrt{\frac{2L(f)\log(1/\delta)}{n}} + \frac{2\log(1/\delta)}{3n} \tag{1.18}$$

probability at least $1 - \delta$. Hint: Use Berinstein's inequality.

■ **Lemma 1.5 — Berinstein's inequality.** If $U_1, ..., U_n$ are n i.i.d. Bernoulli random variables with parameter p, then,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} U_i < p - \epsilon\right) \leq \exp\left(-\frac{n\epsilon^2}{2p + 2\epsilon/3}\right) \tag{1.19}$$

Answer: We first use the Bernstein's inequality and simplify it. Consider the Equation 1.19 and take $\delta = \exp\left(-\frac{n\epsilon^2}{2p+2\epsilon/3}\right)$. Then,

$$\Rightarrow n\epsilon^2 - \left(\frac{2}{3}\ln\frac{1}{\delta}\right)\epsilon - 2p\ln\frac{1}{\delta} = 0$$

$$\Rightarrow \epsilon = \frac{\frac{2}{3}\ln\frac{1}{\delta} \pm \sqrt{\left(\frac{2}{3}\ln\frac{1}{\delta}\right)^2 + 8np\ln\frac{1}{\delta}}}{2n}$$

Based on the assumption of the inequality the $\epsilon \geq 0$ and we can choose the value with the $+$ sign in the about equation. Using this simplification, we can rewrite the Bernstein inequality in the following equivalent form:

$$EU \leq \frac{1}{n}\sum_{i=1}^{n} U_i + \frac{\frac{2}{3}\ln\frac{1}{\delta} + \sqrt{\left(\frac{2}{3}\ln\frac{1}{\delta}\right)^2 + 8np\ln\frac{1}{\delta}}}{2n}, \quad \text{probability at least } 1-\delta$$

Now, for a specific $f \in \mathcal{F}$, we can consider $U_i = \mathbf{1}\{y_i \neq f(x_i)\}$ as a Bernoulli distribution, with the probability of success defined by $p = EU = L(f)$. The empirical estimation is the Bernoulli distribution is

$$\frac{1}{n}\sum_{i=1}^{n} U_i = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{Y_i \neq f(X_i)\} = L_n(f).$$

This we can rewrite the bound as:

$$L(f) \leq L_n(f) + \frac{\ln\frac{1}{\delta}}{3n} + \frac{\sqrt{\left(\frac{2}{3}\ln\frac{1}{\delta}\right)^2 + 8nL(f)\ln\frac{1}{\delta}}}{2n}, \quad \text{probability at least } 1-\delta$$

Now we use the fact that, $\sqrt{a}+\sqrt{b} \geq \sqrt{a+b}$

$$\begin{aligned} L(f) &\leq L_n(f) + \frac{\ln \frac{1}{\delta}}{3n} + \frac{\sqrt{\left(\frac{2}{3}\ln\frac{1}{\delta}\right)^2 + 8nL(f)\ln\frac{1}{\delta}}}{2n} \\ &\leq L_n(f) + \frac{\ln \frac{1}{\delta}}{3n} + \frac{\sqrt{\left(\frac{2}{3}\ln\frac{1}{\delta}\right)^2} + \sqrt{8nL(f)\ln\frac{1}{\delta}}}{2n} \\ &\leq L_n(f) + \frac{2\ln \frac{1}{\delta}}{3n} + \sqrt{\frac{2L(f)\ln\frac{1}{\delta}}{n}}, \qquad \text{probability at least } 1-\delta \end{aligned}$$

Which proves the desired result.

2. Use the result of the previous part to show that, for any $f \in \mathcal{F}$,

$$L(f) \leq L_n(f) + \sqrt{\frac{2L_n(f)\log(1/\delta)}{n}} + \frac{4\log(1/\delta)}{n}$$

with probability at least $1-\delta$. Use this to prove that if the ERM solution predicts every test data correctly, i.e., if $L_n(\hat{f}_n) = 0$, then,

$$L(\hat{f}_n) \leq \frac{4\log(|\mathcal{F}|/\delta)}{n}$$

with probability at least $1-\delta$. This bound also holds with the relationship between X and Y is deterministic. **Hint:** Use the fact that, for any $a, b, c \in \mathbb{R}^+$ and $a \leq b + c\sqrt{a}$, then we have $a \leq b + c^2 + c\sqrt{b}$. **Answer:** We use the hint on the bound which we found in the previous part, in Equation 1.18, with the following definitions:

$$a = L(f), \quad b = L_n(f) + \frac{2\log(1/\delta)}{3n}, \quad c = \sqrt{\frac{2\log(1/\delta)}{n}}$$

This would imply the following inequality:

$$L(f) \leq L_n(f) + \frac{2\log(1/\delta)}{3n} + \left(\sqrt{\frac{2\log(1/\delta)}{n}}\right)^2 + \left(\sqrt{\frac{2\log(1/\delta)}{n}}\right)\sqrt{L_n(f) + \frac{2\log(1/\delta)}{3n}}$$

$$\Rightarrow L(f) \leq L_n(f) + \frac{8\log(1/\delta)}{3n} + \sqrt{\frac{2L_n(f)\log(1/\delta)}{n} + \frac{4}{3}\left(\frac{\log(1/\delta)}{n}\right)^2}$$

We use the inequality $\sqrt{a}+\sqrt{b} \geq \sqrt{a+b}$,

$$\begin{aligned}
L(f) \leq & L_n(f) + \frac{8\log(1/\delta)}{3n} + \sqrt{\frac{2L_n(f)\log(1/\delta)}{n} + \frac{4}{3}\left(\frac{\log(1/\delta)}{n}\right)^2} \\
\leq & L_n(f) + \frac{8\log(1/\delta)}{3n} + \sqrt{\frac{2L_n(f)\log(1/\delta)}{n}} + \sqrt{\frac{4}{3}\left(\frac{\log(1/\delta)}{n}\right)^2} \\
\leq & L_n(f) + \left(\frac{2}{\sqrt{3}} + \frac{8}{3}\right)\frac{\log(1/\delta)}{n} + \sqrt{\frac{2L_n(f)\log(1/\delta)}{n}} \\
= & L_n(f) + \frac{3.83\log(1/\delta)}{n} + \sqrt{\frac{2L_n(f)\log(1/\delta)}{n}} \\
\leq & L_n(f) + \frac{4\log(1/\delta)}{n} + \sqrt{\frac{2L_n(f)\log(1/\delta)}{n}}
\end{aligned}$$

Which proves the desired result. Now using this bound, we prove the last part of the question. Before that we state the union bound for risk.

Since this bound holds for any $f \in \mathcal{F}$, this also holds for $\hat{f} \in \mathcal{F}$. Based on the assumption of the question, the risk for this function is zero. For a fixed $\hat{f} \in \mathcal{F}$, if we have $L_n(\hat{f}) = 0$,

$$L(f) \leq \frac{4\log(1/\delta)}{n}$$

since $\hat{f}$ is not known a priori and it can any function in the class of functions $\mathcal{F}$, we need to use the union bound, as in Equation 1.20:

$$L(f) \leq \frac{4\log(|\mathcal{F}|/\delta)}{n}$$

■

There are similar bounds hold for general functions, on the risk function:

■ **Lemma 1.6** Let $\mathcal{F}$ be a class of functions, defined on domain X and mapping to $[0,1]$. For some $\delta \in (0,1)$, and for any $f \in \mathcal{F}$:

$$\mathbb{E}f(X) \leq \mathbb{E}_n f(X) + 2R_n(\mathcal{F}) + \sqrt{\frac{\log 1/\delta}{2n}}, \text{ with probability at least } 1-\delta$$

Also for any $f \in \mathcal{F}$:

$$\mathbb{E}f(X) \leq \mathbb{E}_n f(X) + 2\hat{R}_n(\mathcal{F}) + 5\sqrt{\frac{\log 2/\delta}{2n}}, \text{ with probability at least } 1-\delta$$

Proof. Proof with McDiarmid's bound. ■

1.3.1 Rademacher averages for Lipchitz functions

■ **Lemma 1.7 — Ledoux-Talagrand contraction.** Let $f : \mathbb{R}_+ \to \mathbb{R}_+$ be a convex and increasing function. Also let $\phi_i(x) : \mathbb{R} \to \mathbb{R}$, s.t. it satisfies $\phi_i(0) = 0$ with Lipchitz constant L (for any $x, y \in \mathbb{R} \Rightarrow |\phi_i(x) - \phi_i(y)| \leq L\,|x - y|$). For any $T \subset \mathbb{R}^n$,

$$\mathbb{E}f\left(\frac{1}{2}\sup_{t\in T}\left|\sum_{i=1}^{n}\sigma_i\phi_i(t_i)\right|\right) \leq \mathbb{E}f\left(L.\sup_{t\in T}\left|\sum_{i=1}^{n}\sigma_i t_i\right|\right)$$

Proof. Proof with definition of Rademacher average and properties of convex functions. ■

The above lemma will result the following bound:

Corollary 1.1 Let $\mathcal{F}$ be a class of functions with domain X and $\phi(.)$ be a L-Lipchitz map from $\mathbb{R}$ to $\mathbb{R}$ with $\phi(0) = 0$. The composition of the map on the functions is defined as $\phi \circ \mathcal{F} = \{\phi \circ f | f \in \mathcal{F}\}$. Then

$$R_n(\phi \circ \mathcal{F}) \leq 2LR_n(\mathcal{F})$$

Proof. In the previous lemma, take the convex increasing function be the identity function. ■

1.4 Glivenko-Cantelli Theorem

The Glivenko-Cantelli guarantees uniform convergence bounds on empirical risk of the distributions. Our characterization of GC is based on Rademacher and Finite Class lemma, though this is not the only way to derive these results. First we introduce the finite class lemma which is a tool for bounding Rademacher averages.

■ **Lemma 1.8 — Finite Class Lemma (Massart).** Let $\mathcal{A}$ be some finite subset of $\mathbb{R}^n$ and $\{\sigma_i\}_{i=1}^m$ independent Rademacher random variables, and $L = \sup_{a\in A}\mathcal{A}$,

$$R_n(\mathcal{A}) \leq \frac{2L\sqrt{\log|\mathcal{A}|}}{n}$$

Proof. Define,

$$\mu = \mathbb{E}\left[\sup_{a\in\mathcal{A}}\sum_{i=1}^{m}\sigma_i x_i\right] = m \times R_n(\mathcal{A})$$

For any $\lambda \in \mathbb{R}^+$,

$$\begin{aligned}
e^{\lambda\mu} &\leq \mathbb{E}\left[\exp\left(\lambda \sup_{a\in\mathcal{A}} \sum_{i=1}^{m} \sigma x_i\right)\right] \\
&= \mathbb{E}\left[\sup_{a\in\mathcal{A}} \exp\left(\lambda \sum_{i=1}^{m} \sigma x_i\right)\right] \\
&\leq \mathbb{E}\left[\sum_{a\in\mathcal{A}} \exp\left(\lambda \sum_{i=1}^{m} \sigma x_i\right)\right] \\
&= \sum_{a\in\mathcal{A}} \mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{m} \sigma x_i\right)\right] \\
&= \sum_{a\in\mathcal{A}} \prod_{i=1}^{m} \mathbb{E}\left[\exp\left(\lambda \sigma x_i\right)\right] \\
&= \sum_{a\in\mathcal{A}} \prod_{i=1}^{m} \frac{\exp\left(-\lambda x_i\right) + \exp\left(\lambda x_i\right)}{2} \\
&\leq \sum_{a\in\mathcal{A}} \prod_{i=1}^{m} \exp\left(\lambda^2 {x_i}^2/2\right) \\
&\leq \sum_{a\in\mathcal{A}} \prod_{i=1}^{m} \exp\left(\lambda^2 L^2/2\right) \\
&\leq |\mathcal{A}| \prod_{i=1}^{m} \exp\left(\lambda^2 L^2/2\right)
\end{aligned}$$

■

$$\Rightarrow \mu \leq \frac{\ln|\mathcal{A}|}{\lambda} + \frac{\lambda L^2}{2}.$$

Set $\lambda = \sqrt{2\frac{\ln|\mathcal{A}|}{L^2}}$, and we will have, $\mu \leq L\sqrt{2\ln|\mathcal{A}|}$

[More details: TBW]

The finite class lemma could be generalized to the class of binary-valued functions. Now define $\mathcal{F}$ be class of binary valued functions,

$$\mathcal{F} = \{f : \mathsf{Z} \to \{0, 1\}\}.$$

In other words, given random samples $\{Z_i\}_{i=1}^n$, and $\mathcal{F}(Z^n) \triangleq \{(f(Z_1), \ldots, f(Z_n)) : f \in \mathcal{F}\}$,

We generalize the bound using the Rademacher bound for this class of functions,

■ **Lemma 1.9 — Rademacher bound for binary-valued functions.** For class of binary-valued functions $\mathcal{F}$,

$$R_n\left(\mathcal{F}\left(Z^n\right)\right) \leq 2\sqrt{\frac{\log\left|\mathcal{F}(Z^n)\right|}{n}}$$

Proof. Since each f is a binary-valued function, $\mathcal{F} \subset \{0,1\}^n$. For any set of samples $\{Z_i\}_{i=1}^n$, and any function $f \in \mathcal{F}$, we know,

$$\sqrt{\sum_{i=1}^{n} |f(Z_i)|} \leq \sqrt{\sum_{i=1}^{n} 1} = \sqrt{n}$$

For a fixed set of random samples,$\{Z_i\}_{i=1}^n$, the set $\mathcal{F}(Z^n) \triangleq \{(f(Z_1),\ldots,f(Z_n)) : f \in \mathcal{F}\}$ is equivalent to the set $\mathcal{A}$, in Lemma **??**, as $N = |\mathcal{F}(Z^n)| \leq 2^n$ and $L = \sqrt{n}$. As such,

$$R_n\left(\mathcal{F}\left(Z^n\right)\right) \leq 2\sqrt{\frac{\log\left|\mathcal{F}(Z^n)\right|}{n}}$$

■

Theorem 1.5 — Glivenko-Cantelli. Let,

$$F_n(x) \triangleq \sum_{i=1}^{n} \mathbf{1}\left\{X_i \leq x\right\}$$

if $n \to \infty$, then

$$\sup_x |F_n(x) - F(x)| \overset{a.s}{\to} 0$$

for n big enough.

Proof. The proof consists of two main pars. First using the Rademacher for bounding the risk, and the second, using the Finite-Class lemma for bounding the Rademacher average. [More details for later] ■

1.5 VC-dimension

Let $\mathcal{F}$ be class of functions defined from $\mathcal{X}$ to $\{-1,1\}$ [2]. Let $X = (X_1,\ldots,X_n)$ be a set of samples. Given $\mathcal{F}$, define the following:

$$S_{\mathcal{F}}(X) \triangleq \{h(X_1),\ldots,h(X_n)\}$$

Definition 1.3 — Growth Function.

$$\mathcal{G}_{\mathcal{F}}(n) \triangleq \max_{\mathcal{S}:|\mathcal{S}|=n} |S_{\mathcal{F}}(X)|$$

[2] The output being mapped to ± 1 is just for simplicity and holds for any binary functions.

By the above definitions we know,

$$\mathcal{G}_{\mathcal{F}}(n) \leq |\mathcal{F}|$$
$$\mathcal{G}_{\mathcal{F}}(n) \leq |2^n|$$

Definition 1.4 — Shattering. A hypothesis class $\mathcal{F}$ shatters a finite set $S \subset \mathcal{X}$, iff $|S_{\mathcal{F}}(S)| = 2^{|S|}$

If the function class shatters the finite set of samples, it means that for any single labelling of the set of the samples, there is a function.

We want to bound Rademacher average using a function of VC-dimension.

■ **Lemma 1.10** Let $\mathcal{F}$ be a class of functions defined on X to $\{+1, -1\}$, then,

$$R_n(\mathcal{F}) \leq \sqrt{\frac{2 \log 2\mathcal{G}_{\mathbf{F}}(n)}{n}}$$

If the function space is symmetric, i.e. given $f \in \mathcal{F}$ then $-f \in \mathcal{F}$:

$$R_n(\mathcal{F}) \leq \sqrt{\frac{2 \log \mathcal{G}_{\mathbf{F}}(n)}{n}}$$

Proof. Proof with finite class lemma. ■

Definition 1.5 — VC-dimension. The **Vapnik-Chervonenkis dimension** of a class $\mathcal{F}$ on a set X, is the cardinality of the largest set shattered by $\mathcal{F}$, that is, the largest n such that there exists a set $S \subset X$, and $|S| = n$ that $\mathcal{F}$ shatters the set S. We denote VC-dimension with $d_{VC}(\mathcal{F})$.

We will use the following lemma to find another bound:

■ **Lemma 1.11** For any $d \leq n$, we have

$$\sum_{i=1}^{k} \binom{n}{i} \leq \left(\frac{en}{k}\right)^k$$

Proof.

$$\begin{aligned}
\left(\frac{k}{n}\right)^k \sum_{i=1}^{k} \binom{n}{i} &\leq \sum_{i=1}^{k} \left(\frac{k}{n}\right)^i \binom{n}{i} \\
&\leq \sum_{i=1}^{n} \binom{n}{i} \left(\frac{k}{n}\right)^i \times 1^{n-i} \\
&\leq \left(1 + \frac{k}{n}\right)^n \leq e^k
\end{aligned}$$

$$\Rightarrow \sum_{i=1}^{k} \binom{n}{i} \leq \left(\frac{en}{k}\right)^k$$

■

Now using the above lemma, we find another bound:

■ **Lemma 1.12 — Souer's lemma.** Let F be a class of functions, mapping from X to a binary space, with $d_{VC}(\mathcal{F}) = d$. Then,

$$\mathcal{G}_{\mathcal{F}}(n) \leq \sum_{i=0}^{d} \binom{n}{i}$$

in addition, for any $n \geq d$

$$\mathcal{G}_{\mathcal{F}}(n) \leq \left(\frac{en}{d}\right)^{d}$$

Proof. The second part of the lemma is trivial based Lemma 1.11. The proof of the first part is by induction on $d + n$. ■

■ **Example 1.5** For a class of real-valued functions $\mathcal{F}$ on $\mathbb{R}$, we define

$$R_n(\mathcal{F}) \triangleq \sup_{z^n \in Z^n} R_n(\mathcal{F}(z^n))$$

For each of the following function classes, prove the Rademacher averages, without relying on the VC-theory,

- $\mathcal{F}_1$ the collection of indicators of all semi-infinite intervals of the form $(-\infty, t], t \in \mathbb{R}$.

$$R_n(\mathcal{F}_1) \leq 2\sqrt{\frac{\log(n+1)}{n}}, \quad \forall n$$

- $\mathcal{F}_2$ is the collection of indicators of all closed intervals of the form $[s, t]$ for $-\infty < s < t < +\infty$.

$$R_n(\mathcal{F}_2) \leq 2\sqrt{\frac{2\log(n) + \log 2}{n}}, \quad \forall n$$

- $\mathcal{F}_3$ is the collection of indicators of all subsets of R that can be represented as unions of no more than k disjoint intervals from $\mathcal{F}_2$.

$$R_n(\mathcal{F}_3) \leq 2\sqrt{\frac{k\log(ne/k)}{n}}, \quad \forall n \geq k$$

Answer: To answer this we use generalization of *the finite class lemma*. We use Lemma 1.9 to prove each of the following cases. In fact, the only thing that we need to do, is to count the number of the distinct values in $\mathcal{F}(Z^n)$.

1. For the class of functions of the form $\mathbf{1}\{X \geq t\}$, the possible configuration of the values is shown

$$n+1 \text{ cases}: \begin{cases} 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & 0 & \dots & 1 & 1 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & 1 & 1 & \dots & 1 & 1 \end{cases}$$

Just plugging-in in the Lemma 1.9, it gives us,

$$R_n(\mathcal{F}_1) \leq 2\sqrt{\frac{\log|\mathcal{F}(Z^n)|}{n}} = 2\sqrt{\frac{\log(n+1)}{n}}, \quad \forall n$$

2. Now consider the class of the functions of the form, $\mathbf{1}\{s \geq X \geq t\}$. Now, we want to count the number of distinct values could be produced by this function. This class consists of binary strings of lengths n, with consecutive sequence of 1's, and the rest being zero. For example,

$$\underbrace{0,0,0,\dots,0,0,0,\overbrace{1,1,1,\dots,1,1,1}^{k \text{ consecutive 1's}},0,0,0,\dots,0,0,0}_{\text{length } n}$$

The number of distinct such binary sequences could easily be counted. The count equals to the number of the ways we can put two separating partitions between object (and before the first object, and after the last object), which equals to $\binom{n}{2}$, plus one for everything being zero. Then,

$$|\mathcal{F}(Z^n)| = \binom{n}{2} + 1 \Rightarrow \log|\mathcal{F}(Z^n)| = \frac{n(n+1)}{2} + 1 \leq 2n^2$$

$$R_n(\mathcal{F}_2) \leq 2\sqrt{\frac{\log(2n^2)}{n}} \leq 2\sqrt{\frac{2\log(n)+\log 2}{n}}$$

3. We use the Sour's lemma. The number of distinct subsets of size at most k elements, among n elements, equal to,

$$\sum_{i=1}^{k} \binom{n}{i}$$

which, based on Lemma 1.11, is upper-bounded by $\left(\frac{en}{k}\right)^k$. Using the finite class lemma, this gives us the following Rademacher bound,

$$R_n(\mathcal{F}_3) \leq 2\sqrt{\frac{k\log(ne/k)}{n}}, \quad \forall n \geq k$$

■

■ **Example 1.6** The *sup-norm* for any space of functions is defined as

$$||f|| \triangleq \sup_{z \in Z} |f(z)|$$

Given $\mathcal{F}$ the class of positive-valued functions on Z and $\epsilon > 0$, the $\epsilon - net$ of $\mathcal{F}$ with respect to the *sup-norm* is any $f \in \mathcal{F}$, such that,

$$||f - f_j||_\infty = \sup_{z \in Z} |f(z) - f_j(z)| \leq \epsilon,$$

for at least one of the functions in $\mathcal{F}' = \{f_1, \ldots, f_k\}$, which are not necessarily in $\mathcal{F}$. The ϵ-*covering number* of $\mathcal{F}$ w.r.t. to the *sup-norm*, or the cardinality of a minimal $\epsilon - net$ of $\mathcal{F}$, is denoted by $N_\infty(\mathcal{F}, \epsilon)$. If $\mathcal{F}$ does not accept ϵ-net $N_\infty(\mathcal{F}, \epsilon) = \infty$. The logarithm of the ϵ-covering number, is usually called ϵ-number of $\mathcal{F}$ and denoted by $H_\infty(\mathcal{F}, \epsilon)$.

1. For $\mathcal{F}$ the family of uniformly-bounded functions (i.e. $\exists L > 0 \quad s.t. \quad \forall f \in \mathcal{F} \Rightarrow ||f||_\infty \leq L$). Show that,

$$R_n(\mathcal{F}) \leq \inf_{\epsilon > 0} \left(\epsilon + 2L \sqrt{\frac{\log N_\infty(\mathcal{F}, \epsilon)}{n}} \right)$$

Answer: We start with the definition of the Rademacher complexity, and bound it from above. We use the property given in the problem that, for any function $f \in \mathcal{F}$, there exits a function $f_j \in \mathcal{F}'$ such that $\sup_{z \in Z} |f(z) - f_j(z)| \leq \epsilon$. Given an arbitrary $f \in \mathcal{F}$,

$$\begin{aligned}
\left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(Z_i) \right| &= \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \left(f(Z_i) - f_j(Z_i) + f_j(Z_i) \right) \right| \\
&\leq \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \left(f(Z_i) - f_j(Z_i) \right) \right| + \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i f_j(Z_i) \right| \\
&\leq \epsilon + \max_{1 \leq j \leq k} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i f_j(Z_i) \right|
\end{aligned}$$

For a given set of samples Z^n,

$$\begin{aligned}
R_n(\mathcal{F}(Z^n)) &= \mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(Z_i) \right| \\
&\leq \mathbb{E} \left\{ \epsilon + \max_{1 \leq j \leq k} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i f_j(Z_i) \right| \right\} \\
&= \epsilon + \mathbb{E} \max_{1 \leq j \leq k} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i f_j(Z_i) \right| \\
&\leq \epsilon + 2L \sqrt{\frac{\log N_\infty(\mathcal{F}(Z^n), \epsilon)}{n}}
\end{aligned}$$

To be more accurate, the last bound above is, $\epsilon + 2(L + \epsilon) \sqrt{\frac{\log N_\infty(\mathcal{F}(Z^n), \epsilon)}{n}}$. Since the covering functions aren't necessarily bounded by L (but bounded

by $L+\epsilon$ instead). But for any set of covering function f_j which is $|f_j(z)| > L$, we can limit it to $\pm L$, and it will still be a covering:

$$f'(z) = \begin{cases} L & f_j(z) > L \\ -L & f_j(z) < -L \\ 0 & \text{otherwise} \end{cases}$$

This will bound the covering functions to L and will give the desired bound. The Rademacher average over the whole class,

$$R_n(\mathcal{F}) = \sup_{z^n \in \mathsf{Z}^n} R_n(\mathcal{F}(z^n)) \leq \epsilon + 2L\sqrt{\frac{\log N_\infty(\mathcal{F}, \epsilon)}{n}}$$

Now we can tighten the bound for an arbitrary value of $\epsilon > 0s$,

$$R_n(\mathcal{F}) \leq \inf_{\epsilon \in \mathbb{R}^+} \left\{ \epsilon + 2L\sqrt{\frac{\log N_\infty(\mathcal{F}, \epsilon)}{n}} \right\}$$

2. Let

$$Z = \left\{ \left(z^{(1)}, \ldots, z^{(d)}\right) \in \mathbb{R}^d : ||z||_1 = \sum_{j=1}^{d} z^{(j)} \leq 1 \right\},$$

and $\mathcal{F}$ consisting of functions of the form $f(z) = f_w(z) = \langle w, z \rangle$, for all $w \in \mathbb{R}^d$ with $||w||_\infty = \max_{1 \leq j \leq d} \left|w^{(j)}\right| \leq 1$.
Show that

$$N_\infty(\mathcal{F}, \epsilon) \leq \left(\frac{2}{\epsilon}\right)^d$$

and prove that,

$$R_n(\mathcal{F}) = O\left(\sqrt{\frac{d \log n}{n}}\right).$$

Answer: Before starting the proof, we state Holder's inequality without proof. We will use this theorem during the proof.

> ■ **Lemma 1.13 — Holder's inequality.** Consider f and g are two measurable real-valued functions defined on a measurable space. Let $p, q \in [1, +\infty]$, and $1/p + 1/q = 1$. Then,
>
> $$||fg||_1 \leq ||f||_p ||g||_q$$
>
> The theorem also holds in the extremal cases, when $p = \infty$, $q = 1$.

We show that one can find a family of functions $\mathcal{G}$ with size $(2/\epsilon)^n$ such that for any functions $f \in \mathcal{F}$, there exists $g \in \mathcal{G}$, and $||f - g||_\infty \leq \epsilon$. To show this, it is enough to show it for special case of $\mathcal{G}$, though there might be better answers. For that, we define the following class functions:

$$\mathcal{G}(\epsilon) = \left\{ \langle w.z \rangle | z \in \mathsf{Z}, w \in \mathcal{W}(\epsilon)^d \right\},$$

in which $\mathcal{W}(\epsilon) = \{\pm\epsilon k | k \in [1, ..., 1/\epsilon]\}$. Based on the above definition,

$$\forall w \in [-1,1], \exists w' \in \mathcal{W},\ s.t.\ |w - w'| \leq \epsilon$$

$$\Rightarrow \forall w \in [-1,1]^d, \exists w' \in \mathcal{W}^d,\ s.t.\ ||w - w'||_\infty = \max_{1 \leq j \leq d} |w - w'| \leq \epsilon$$

Also note that, $|\mathcal{G}(\epsilon)| = (2/\epsilon)^d$. Now, for any arbitrary function $f \in \mathcal{F}$, we choose the function $g \in \mathcal{G}$ which has smallest $||f - g||_\infty$. For this function, for a given $z \in \mathsf{Z}$,

$$\begin{aligned} \forall z \in \mathsf{Z} \quad |f(z) - g(z)| &\leq |w.z - w'.z| \\ &\leq |w - w'|_\infty |z|_1 \quad \text{(Holder's inequality)} \\ &\leq \epsilon \times 1 \Rightarrow ||f - g||_\infty = \sup_{z \in \mathsf{Z}} |f(z) - g(z)| \leq \epsilon \end{aligned}$$

$$\Rightarrow \forall f \in \mathcal{F}, \exists g \in \mathcal{G},\ s.t.\ ||f - g||_\infty \leq \epsilon$$

This end our proof that, $N_\infty(\mathcal{F}, \epsilon) \leq (2/\epsilon)^d$.

Now we use the result of the previous part, and plug-in $N_\infty(\mathcal{F}, \epsilon)$,

$$\begin{aligned} R_n(\mathcal{F}) &\leq \inf_{\epsilon \in \mathbb{R}^+} \left\{ \epsilon + 2\sqrt{\frac{\log N_\infty(\mathcal{F}, \epsilon)}{n}} \right\} \\ &\leq \inf_{\epsilon \in \mathbb{R}^+} \left\{ \epsilon + 2\sqrt{\frac{d \ln \frac{2}{\epsilon}}{n}} \right\} \end{aligned}$$

If we choose $\epsilon = \frac{2}{n}$,

$$R_n(\mathcal{F}) \leq \epsilon + 2\sqrt{\frac{d \ln \frac{2}{\epsilon}}{n}} = \frac{2}{n} + 2L\sqrt{\frac{d \ln n}{n}}$$

Thus,

$$R_n(\mathcal{F}) \leq \frac{2}{n} + 2\sqrt{\frac{d \ln n}{n}} \leq C_2 \sqrt{\frac{d \ln n}{n}},\ \text{for } C_2 \text{ big enough.}$$

It can be shown that $C_2 = 4$ satisfies the above property. Now we prove this. For any $d \geq 1$, and for any $n \geq 1$, we have

$$\frac{2}{n} \leq 2\sqrt{\frac{1}{n}} \leq 2\sqrt{\frac{d \ln n}{n}} \Rightarrow \frac{2}{n} + 2\sqrt{\frac{d \ln n}{n}} \leq 4\sqrt{\frac{d \ln n}{n}},$$

$$\Rightarrow R_n(\mathcal{F}) \leq 4\sqrt{\frac{d \ln n}{n}}.$$

Which proves,

$$R_n(\mathcal{F}) = O\left(\sqrt{\frac{d \ln n}{n}}\right)$$

3. Suppose $\mathcal{F}$ is such that $H_\infty(\mathcal{F}, \epsilon) \leq C\epsilon^{-\frac{1}{\alpha}}$ for some constant $C > 0$ and $\alpha > 0$. For example,
   - the class of functions $\mathcal{F}$ all differentiable $f : [0,1] \to [0,1]$ with $|f'| \leq 1$, then the above bound holds with $\alpha = 1$.

   Prove that
$$R_n(\mathcal{F}) \leq Cn^{-\frac{\alpha}{2\alpha+1}}, \quad C > 0$$

   Answer: Given the assumption, we know,
$$\ln N_\infty(\mathcal{F}, \epsilon) \leq C\epsilon^{-\frac{1}{\alpha}}.$$

   By plugging this into the result of the first part,
$$\begin{aligned} R_n(\mathcal{F}) &\leq \inf_{\epsilon \in \mathbb{R}^+} \left\{ \epsilon + 2L\sqrt{\frac{\log N_\infty(\mathcal{F}, \epsilon)}{n}} \right\} \\ &\leq \inf_{\epsilon \in \mathbb{R}^+} \left\{ \epsilon + 2L\sqrt{\frac{C\epsilon^{-\frac{1}{\alpha}}}{n}} \right\} \end{aligned}$$

   Now, choosing $\epsilon = n^{-\frac{\alpha}{2\alpha+1}}$,
$$\epsilon + 2L\sqrt{\frac{C\epsilon^{-\frac{1}{\alpha}}}{n}} = n^{-\frac{\alpha}{2\alpha+1}} + 2L\sqrt{C}n^{-\frac{\alpha}{2\alpha+1}} = \left(1 + 2L\sqrt{C}\right) n^{-\frac{\alpha}{2\alpha+1}}$$
$$\Rightarrow R_n(\mathcal{F}) \leq C' n^{-\frac{\alpha}{2\alpha+1}} \text{ for } C' \geq 1 + 2L\sqrt{C}$$

■

■ **Lemma 1.14** For non-negative random variable Z, if we know,
$$\mathbb{P}(Z \geq t) \leq Ce^{-2nt^2},$$
for some universal constant $C > 0$ and $C < +\infty$, one can show that,
$$\mathbb{E}[Z] \leq \sqrt{\frac{\ln(Ce)}{2n}}$$

Proof. To prove this we use the fact that, the variance of a non-negative random variable is non-negative. Thus,
$$\mathbb{E}[Z]^2 \leq \mathbb{E}\left[Z^2\right].$$

Using this fact,

$$\begin{aligned}\mathbb{E}\left[Z\right]^2 \leq \mathbb{E}\left[Z^2\right] &= \int_0^{+\infty} \mathbb{P}\left(Z^2 \geq t\right) dt \\ &\leq \int_0^{z} 1dt + \int_z^{+\infty} \mathbb{P}\left(Z \geq \sqrt{t}\right) dt \\ &\leq z + \int_z^{+\infty} Ce^{-2nt} dt \\ &= z + \frac{C}{2n} e^{-2nz}\end{aligned}$$

$$\Rightarrow \mathbb{E}\left[Z\right] \leq \sqrt{\inf_{z \in \mathbb{R}^+} \left\{ z + \frac{C}{2n} e^{-2nz} \right\}}$$

Now, since this bound holds for any $z \in \mathbb{R}^+$, we minimize it with respect to z to find a tighter bound.

$$\frac{\partial}{\partial z}\left(z + \frac{C}{2n} e^{-2nz}\right) = 1 + \frac{C}{2n}(-2n)e^{-2nz} = 0$$

$$\Rightarrow z = \frac{1}{2n} \ln C \Rightarrow \mathbb{E}\left[Z\right] \leq \sqrt{\frac{\ln e \times C}{2n}}$$

■

■ **Example 1.7** Let X be real-valued random variable with CDF $F(x) = \mathbb{P}\left(X \leq x\right)$. If $X_1, \ldots, X_n$ are i.i.d. copies of X, the empirical CDF is,

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq x\}}.$$

1. Using the Rademacher complexity techniques prove that

$$\mathbb{E}\left[\sup_{x \in \mathbb{R}} \left| \hat{F}_n(x) - F_n(x) \right| \right] \leq \frac{C}{\sqrt{n}}, \quad C > 0$$

Answer:
To prove this bound, we use the notion of the VC-dimension. We know,

$$\mathbb{E}R_n(\mathcal{F}(Z^n)) \leq C\sqrt{\frac{V(\mathcal{F})}{n}}$$

We define $\mathcal{F}$ to be set of indicators on half-interval. It is easy to show that, the VC-dimension for a class of functions consisting of half-intervals is one. Also, we can treat the CDF function, as empirical risk for a set of half-interval functions. By this assumption, we know that, using the symmetrization trick we can find the following bound (proof in Lemma

1.3):

$$\mathbb{E}\left[\sup_{x\in\mathbb{R}}\left|\hat{F}_n(x)-F_n(x)\right|\right]\leq 2\mathbb{E}R_n(\mathcal{F}(Z^n))$$

Using these facts, we can find the following bound,

$$\mathbb{E}\left[\sup_{x\in\mathbb{R}}\left|\hat{F}_n(x)-F_n(x)\right|\right]\leq 2\mathbb{E}R_n(\mathcal{F}(Z^n))\leq\frac{C}{\sqrt{n}},\qquad\text{for some } C>0$$

2. If $C=1$, prove *Massart*'s inequality,

$$\mathbb{P}\left(\sup_{x\in\mathbb{R}}\left|\hat{F}_n(x)-F_n(x)\right|>t\right)\leq 2e^{-2nt^2},\quad\forall t>0.$$

Note: it can be shown $C=1$ is optimal.
Answer:
To prove this, we first prove a lemma.
We use the Lemma 1.15. Defining $Z=\sup_{x\in\mathbb{R}}\left|\hat{F}_n(x)-F_n(x)\right|$, we know that,

$$\mathbb{E}\left[Z\right]\leq\sqrt{\frac{\ln\left(e\times 2\right)}{2n}}\leq\sqrt{\frac{2\ln\left(e\right)}{2n}}=\frac{1}{\sqrt{n}}.$$

■

■ **Lemma 1.15** For non-negative random variable Z, if we know,

$$\mathbb{P}(Z\geq t)\leq Ce^{-2nt^2},$$

for some universal constant $C>0$ and $C<+\infty$, one can show that,

$$\mathbb{E}\left[Z\right]\leq\sqrt{\frac{\ln(Ce)}{2n}}$$

Proof. To prove this we use the fact that, the variance of a non-negative random variable is non-negative. Thus,

$$\mathbb{E}\left[Z\right]^2\leq\mathbb{E}\left[Z^2\right].$$

Using this fact,

$$\begin{aligned}\mathbb{E}\left[Z\right]^2\leq\mathbb{E}\left[Z^2\right]&=\int_0^{+\infty}\mathbb{P}\left(Z^2\geq t\right)dt\\&\leq\int_0^{z}1dt+\int_z^{+\infty}\mathbb{P}\left(Z\geq\sqrt{t}\right)dt\\&\leq z+\int_z^{+\infty}Ce^{-2nt}dt\\&=z+\frac{C}{2n}e^{-2nz}\end{aligned}$$

$$\Rightarrow \mathbb{E}[Z] \leq \sqrt{\inf_{z \in \mathbb{R}^+} \left\{ z + \frac{C}{2n} e^{-2nz} \right\}}$$

Now, since this bound holds for any $z \in \mathbb{R}^+$, we minimize it with respect to z to find a tighter bound.

$$\frac{\partial}{\partial z}\left(z + \frac{C}{2n} e^{-2nz} \right) = 1 + \frac{C}{2n}(-2n)e^{-2nz} = 0$$

$$\Rightarrow z = \frac{1}{2n} \ln C \Rightarrow \mathbb{E}[Z] \leq \sqrt{\frac{\ln e \times C}{2n}}$$

■

■ **Example 1.8** Let X be real-valued random variable with CDF $F(x) = \mathbb{P}(X \leq x)$. If $X_1, \ldots, X_n$ are i.i.d. copies of X, the empirical CDF is,

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq x\}}.$$

1. Using the Rademacher complexity techniques prove that

$$\mathbb{E}\left[\sup_{x \in \mathbb{R}} \left| \hat{F}_n(x) - F_n(x) \right| \right] \leq \frac{C}{\sqrt{n}}, \quad C > 0$$

Answer: To prove this bound, we use the notion of the VC-dimension. We know,

$$\mathbb{E}R_n(\mathcal{F}(Z^n)) \leq C\sqrt{\frac{V(\mathcal{F})}{n}}$$

We define $\mathcal{F}$ to be set of indicators on half-interval. It is easy to show that, the VC-dimension for a class of functions consisting of half-intervals is one. Also, we can treat the CDF function, as empirical risk for a set of half-interval functions. By this assumption, we know that, using the symmetrization trick we can find the following bound (proof in Lemma 1.3):

$$\mathbb{E}\left[\sup_{x \in \mathbb{R}} \left| \hat{F}_n(x) - F_n(x) \right| \right] \leq 2\mathbb{E}R_n(\mathcal{F}(Z^n))$$

Using these facts, we can find the following bound,

$$\mathbb{E}\left[\sup_{x \in \mathbb{R}} \left| \hat{F}_n(x) - F_n(x) \right| \right] \leq 2\mathbb{E}R_n(\mathcal{F}(Z^n)) \leq \frac{C}{\sqrt{n}}, \quad \text{for some } C > 0$$

Now we can use Lemma 1.3 which will give us the desired result.

2. If $C = 1$, prove *Massart*'s inequality,

$$\mathbb{P}\left(\sup_{x \in \mathbb{R}} \left| \hat{F}_n(x) - F_n(x) \right| > t \right) \leq 2e^{-2nt^2}, \quad \forall t > 0.$$

Note: it can be shown $C = 1$ is optimal.
Answer: We use the Lemma 1.15. Defining $Z = \sup_{x\in\mathbb{R}} \left|\hat{F}_n(x) - F_n(x)\right|$, we know that,

$$\mathbb{E}[Z] \leq \sqrt{\frac{\ln(e \times 2)}{2n}} \leq \sqrt{\frac{2\ln(e)}{2n}} = \frac{1}{\sqrt{n}}.$$

■

Theorem 1.6 Let F be class of function defined on space X mapped to Y, from an unknown distribution $\mathbb{P}_{XY}$. Then given i.i.d. samples $\{(X_i, Y_i)\}_{i=1}^n$, then with probability at least $1-\delta$:

$$\begin{aligned}\mathbb{P}(h(X) \neq Y) &\leq \hat{\mathbb{P}}(h(x_i) \neq y_i) + 4R_n(\mathcal{F}) + \sqrt{\frac{1/\delta}{2n}} \\ &\leq \hat{\mathbb{P}}(h(x_i) \neq y_i) + 4\sqrt{\frac{2d\log(en) - 2d\log d}{n}} + \sqrt{\frac{1/\delta}{2n}}\end{aligned}$$

This could be genelized to all of the samples,

$$\mathbb{P}(h(X) \neq Y) \leq \hat{\mathbb{P}}(h(X) \neq Y) + O\left(\sqrt{\frac{d\log n + \log 1/\delta}{n}}\right)$$

1.6 Union bound for risk

Let's assume we have proven the following bound for any $f \in \mathcal{F}$,

$$p(L(f) - L_n(f) \geq a(\delta)) \leq \delta, \quad \text{for any } f \in \mathcal{F}$$

which is equivalent to,

$$L_n(f) \geq L(f) + b(\delta) \quad \text{with probability at least } 1-\delta \tag{1.20}$$

for some values a, b (functions of parameters). Then,

$$p(\exists f \in \mathcal{F} \wedge L_n(f) = 0 \wedge L(f) \geq a) \leq |\mathcal{F}|\delta$$

or, equivalently,

$$L_n(f) \geq L(f) + b(\delta/|\mathcal{F}|) \quad \text{with probability at least } 1-\delta$$

Proof.

$$\begin{aligned}p(\exists f \in \mathcal{F} \wedge L_n(f) = 0 \wedge L(f) \geq a) &\leq p\left(\cup_{f\in\mathcal{F}}\left(L_n(f) = 0 \wedge L(f) \geq a\right)\right) \\ &\leq \sum_{f\in\mathcal{F}} p\left(\left(L_n(f) = 0 \wedge L(f) \geq a\right)\right) \\ &\leq |\mathcal{F}|\delta\end{aligned}$$

Now define $\delta' = \frac{\delta}{|\mathcal{F}|}$, and then using 1.20 we have

$$L_n(f) \geq L(f) + b(\delta') = L(f) + b(\delta/|\mathcal{F}|) \quad \text{with probability at least } 1-\delta$$

which proves our desired statement. ■

1.7 Kernels and Hilbert spaces

Theorem 1.7 — Mercer's theorem. Suppose K is a continuous symmetric non-negative definite kernel. Then there is a set of orthonormal basis $\{\varphi_i \in L^2(\mathsf{X}, P)\}$ consisting of eigenfunctions of T_k, i.e. $T_K\varphi_j = \lambda_j\varphi_j$, such that the corresponding sequence of eigenvalues $\{\lambda_i\}$ is nonnegative. The eigenfunctions corresponding to non-zero eigenvalues are continuous on X and K have the representation

$$K(s,t) = \sum_{j=1}^{\infty} \lambda_j \, \varphi_j(s) \, \varphi_j(t), \quad \forall s,t \in \mathsf{X}$$

where the convergence is absolute and uniform.

1.7.1 Reproducing Kernel Hilbert Spaces(RKHS)

The RKHS property says that, projecting any function in $\mathcal{L}_K(\mathsf{X})$ will produce exact the same function:

$$\begin{aligned}\langle f, K(x,.)\rangle_K &= \langle \sum_j c_j K(x_j,.), K(x,.)\rangle_K \\ &= \sum_j c_j \langle K(x_j,.), K(x,.)\rangle_K \\ &= \sum_j c_j K(x_j,x) = f(x)\end{aligned}$$

Another representation of RKHS is based on the eigen functions spanning the space of the kernels. Any function $f \in \mathcal{L}_K(\mathsf{X})$ can be represented as,

$$f(x) = \sum_i c_i K(x_i,x) = \sum_i c_i \sum_j \lambda_j \, \varphi_j(x_i) \, \varphi_j(x) = \sum_j \sum_i c_i \lambda_j \, \varphi_j(x_i) \, \varphi_j(x) = \sum_j d_j \varphi_j(x)$$

■ **Example 1.9** Let X be a compact (i.e. closed and bounded) subset of $\mathbb{R}^d$, and let $K : \mathsf{X} \times \mathsf{X} \to \mathbb{R}$ be Mercer kernel defined over X. With a fixed probability distribution P on X, consider the Hilbert space $L^2(\mathsf{X}, P)$ of functions $g : \mathsf{X} \to \mathbb{R}$, where,

$$\int_{\mathsf{X}} g^2(x) P(dx) < \infty$$

with the norm defined as,

$$\langle g, g' \rangle = \int_{\mathsf{X}} g(x) g'(x) P(dx) = \mathbb{E}\left[g(X) g'(X)\right]$$

Also consider the operator T_K

$$[T_K\phi](x) = \int_{\mathsf{X}} K(x,t)\varphi(t) P(dt), \quad \forall x \in \mathsf{X}$$

which maps a function?? .

For a given kernel K, define $\mathcal{L}_K(\mathsf{X})$ to be the set of all functions such that,

$$f(x) = \sum_j c_j K(x_j, x)$$

Using Mercer's reproducing kernel theorem, prove that,

1. Let $J \triangleq \{j \in \mathbb{N} : \lambda_j > 0\}$, and for each $j \in J$ define the function $\psi \triangleq \varphi_j \sqrt{\lambda_j}$. Then $\{\psi_j\}_{j \in J}$ is an orthonormal system in the RKHS $\mathcal{H}_K$, i.e. $\langle \psi_j, \psi_k \rangle_K = \delta_{jk}$, for all $j, k \in J$. **Answer:** The answer is inspired from the formulation in Cucker and Zhou (2007). Based on the definitions we have

$$\begin{aligned}
\langle \psi_j(x), \psi_k(x) \rangle_K &= \left\langle \sqrt{\lambda_j}\varphi_j(x), \sqrt{\lambda_k}\varphi_k(x) \right\rangle_K \\
&= \left\langle \frac{1}{\sqrt{\lambda_j}} \int_{\mathsf{X}} K(x,t)\varphi_j(t)P(dt), \sqrt{\lambda_k}\varphi_k(x) \right\rangle_K && \text{projections} \\
&= \frac{\sqrt{\lambda_k}}{\sqrt{\lambda_j}} \int_{\mathsf{X}} \varphi_j(t) \langle K(x,t), \varphi_k(x) \rangle_K P(dt) \\
&= \frac{\sqrt{\lambda_k}}{\sqrt{\lambda_j}} \int_{\mathsf{X}} \varphi_j(t)\varphi_k(t)P(dt) && \text{RKHS property} \\
&= \frac{\sqrt{\lambda_k}}{\sqrt{\lambda_j}} \langle \varphi_j(t), \varphi_k(t) \rangle_{L^2(\mathsf{X},P)} \\
&= \frac{\sqrt{\lambda_k}}{\sqrt{\lambda_j}} \delta_{k,j} && \varphi_j\text{:orthonormal} \\
&= \delta_{k,j}
\end{aligned}$$

2. Let $\mathcal{F}$ be the unit ball of $\mathcal{H}_K$, and let $X_1, X_2, \ldots, X_n$ be drawn i.i.d. from P. Then

$$\mathbb{E}R_n\left(\mathcal{F}(X^n)\right) \leq \sqrt{\frac{1}{n}\sum_{j=1}^{+\infty} \lambda_j}$$

Answer: We consider the ball of K:

$$\mathcal{F}_\lambda = \{f \in \mathcal{H}_K : ||f||_K \leq 1\}$$

Here I am just reviewing the procedure introduced for finding the risk for

this,

$$
\begin{aligned}
R_n(\mathcal{F}_\lambda(X^n)) &= \sup_{f:\|f\|_K\leq\lambda} \frac{1}{n}\mathbb{E}_{\sigma^n}\left|\sum_{i=1}^n \sigma_i f(X_i)\right| && (1.21)\\
&= \sup_{f:\|f\|_K\leq\lambda} \frac{1}{n}\mathbb{E}_{\sigma^n}\left|\sum_{i=1}^n \sigma_i \langle f, K_{X_i}\rangle_K\right| && (1.22)\\
&= \sup_{f:\|f\|_K\leq\lambda} \frac{1}{n}\mathbb{E}_{\sigma^n}\left|\left\langle f, \sum_{i=1}^n \sigma_i K_{X_i}\right\rangle_K\right| && (1.23)\\
&= \frac{\lambda}{n}\mathbb{E}_{\sigma^n}\left\|\sum_{i=1}^n \sigma_i K_{X_i}\right\|_K && (1.24)\\
&= \frac{\lambda}{n}\sqrt{\sum_{i=1}^n \|K_{X_i}\|_K^2} && (1.25)\\
&= \frac{\lambda}{n}\sqrt{\sum_{i=1}^n \langle K_{X_i}, K_{X_i}\rangle_K} && (1.26)\\
& && (1.27)
\end{aligned}
$$

Now we first simplify $\langle K_{X_i}, K_{X_i}\rangle$ and plug in the results in the above bound. But before that, we use the result we found in the previous part. Previously we proved that, $\langle \psi_i, \psi_j\rangle_K = \delta_{i,j}$, we can use this result:

$$\langle \psi_i, \psi_j\rangle_K = \sqrt{\lambda_i\lambda_j}\langle \varphi_i, \varphi_j\rangle_K = \delta_{i,j} \Rightarrow \langle \varphi_i, \varphi_j\rangle_K = \frac{1}{\sqrt{\lambda_i\lambda_j}}\delta_{i,j} \qquad (1.28)$$

Using this result, we simplify $\langle K_{X_i}, K_{X_i}\rangle$ in Equation 1.21.

$$
\begin{aligned}
\sum_{i=1}^n \langle K_{X_i}, K_{X_i}\rangle &= \sum_{i=1}^n \left\langle \sum_{j=1}^{+\infty} \lambda_j\varphi_j(X_i)\varphi(X), \sum_{k=1}^{+\infty} \lambda_k\varphi_k(X_i)\varphi(X)\right\rangle_K\\
&= \sum_{i=1}^n \sum_{j=1}^{+\infty}\sum_{k=1}^{+\infty} \lambda_j\lambda_k\varphi_j(X_i)\varphi_k(X_i)\langle \varphi(X), \varphi(X)\rangle_K\\
&= \sum_{i=1}^n \sum_{j=1}^{+\infty}\sum_{k=1}^{+\infty} \frac{\lambda_j\lambda_k}{\sqrt{\lambda_j\lambda_k}}\varphi_j(X_i)\varphi_k(X_i)\delta_{j,k}\\
&= \sum_{i=1}^n \sum_{j=1}^{+\infty} \lambda_j\varphi_j^2(X_i)
\end{aligned}
$$

Now we plug this result in the bound we found in Equation 1.21, with $\lambda = 1$ (the unit ball).

$$R_n(\mathcal{F}_\lambda(X^n)) \leq \frac{1}{n}\sqrt{\sum_{i=1}^n \sum_{j=1}^{+\infty} \lambda_j\varphi_j^2(X_i)}$$

Now we take expectation with respect to samples,

$$\begin{aligned}
\mathbb{E}R_n(\mathcal{F}_\lambda(X^n)) &\leq \frac{1}{n}\mathbb{E}\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{+\infty}\lambda_j\varphi_j^2(X_i)} \\
&\leq \frac{1}{n}\sqrt{\mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{+\infty}\lambda_j\varphi_j^2(X_i)\right]} \\
&= \frac{1}{n}\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{+\infty}\lambda_j\mathbb{E}\left[\varphi_j^2(X_i)\right]} \\
&= \frac{1}{n}\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{+\infty}\lambda_j} \\
&= \frac{1}{n}\sqrt{n\sum_{j=1}^{+\infty}\lambda_j} \\
&= \sqrt{\frac{1}{n}\sum_{j=1}^{+\infty}\lambda_j}
\end{aligned}$$

Which gives the desired result.

■

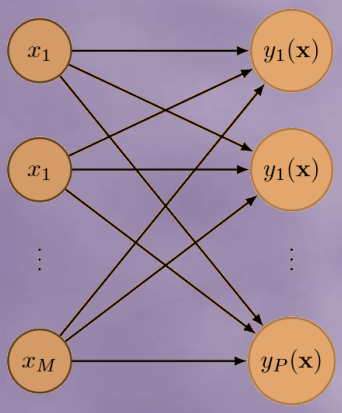
1.8 Perceptron algorithm

This algorithm was discovered around 60s, but at that time no one really appreciate it! Discovering this algorithm was one of the fundamental steps for the research in neural networks. For around two decades the research on this algorithm was dormant, until around 80s people started using this simple, but powerful method. [TBW]

1.9 Bibliographical notes

This is mostly written during Statistical Learning Theory course at UIUC, by Maxim Raginsky Raginsky (2011). In particular some sample questions from Maxim's homework assignments. I have also used notes from Sham Kakade's course and John Duchi's "for fun!" notes.

1.10 Problems



Bibliography

Felipe Cucker and Ding Xuan Zhou. *Learning theory: an approximation theory viewpoint*. Number 24. Cambridge University Press, 2007.

Maxim Raginsky. Lecture notes: Ece 299: Statistical learning theory. *Tutorial*, 2011. URL `http://maxim.ece.illinois.edu/teaching/spring11/schedule.html`.