NOTES AND SOLUTIONS TO MOHRI'S *FOUNDATIONS OF MACHINE LEARNING*

LUCAS TUCKER

Abstract. The following are a series of notes and solutions to Chapters 2, 3, 4, and 15 from *Foundations of Machine Learning* by Mehryar Mohri.

Contents

Date: September 15, 2023.

Chapter 2 Notes

To show $E[\widehat{R}_S(h)] = R(h)$, or that the expectation of empirical error over m samples drawn from a distribution D is equal to generalization error, we have

$$E_{S\sim D^m}[\widehat{R}_S(h)] = \frac{1}{m}\sum_{i=1}^{m} E_{S\sim D^m}[\chi_{c(x_i)\neq h(x_i)}]$$
$$= E_{S\sim D^m,\, x\in S}[\chi_{c(x)\neq h(x)}] = E_{x\sim D}[\chi_{c(x)\neq h(x)}] = R(h)$$

Definition (PAC-learning): A concept class $\mathcal{C}$ is "PAC-learnable" if there exists an algorithm $\mathcal{A}$ and a polynomial function poly(., ., ., .) such that for any $\epsilon > 0$ and $\delta > 0$, for all distributions $\mathcal{D}$ on $\mathcal{X}$ and for any target concept $c \in \mathcal{C}$,

$$\mathbb{P}_{S\sim D^m}[R(h_S) \leq \epsilon] \geq 1 - \delta$$

where h_S denotes the hypothesis returned by $\mathcal{A}$ after receiving the labeled sample S. If $\mathcal{A}$ further runs in $\text{poly}(1/\epsilon, 1/\delta, n, \text{size}(c))$ then $\mathcal{C}$ is said to be "efficiently PAC-learnable" and $\mathcal{A}$ is deemed a "PAC learning algorithm for $\mathcal{C}$".

Theorem (Learning Bound – finite, $\mathcal{H}$ consistent): Let $\mathcal{H}$ be a finite set of functions from $\mathcal{X}$ to $\mathcal{Y}$. Let $\mathcal{A}$ be an algorithm that for any target concept $c \in \mathcal{H}$ and iid sample S returns a consistent hypothesis h_S such that $\widehat{R}_S(h_S) = 0$. Then for any $\epsilon, \delta > 0$,

$$m \geq \frac{1}{\epsilon}(\log|\mathcal{H}| + \log\frac{1}{\delta})$$
$$\Rightarrow \mathbb{P}_{S\sim D^m}[R(h_S) \leq \epsilon] \geq 1 - \delta$$

Proof: Fix $\epsilon > 0$ and consider $\mathcal{H}_\epsilon := \{h \in \mathcal{H} : R(h) > \epsilon\}$. Then, $\mathbb{P}[\widehat{R}_S(h) = 0] \leq (1-\epsilon)^m$ for $S \sim \mathcal{D}$ of size m. Hence,

$$\mathbb{P}[\exists h \in \mathcal{H}_\epsilon : \widehat{R}_S(h) = 0]$$
$$= \mathbb{P}[\widehat{R}_S(h_1) = 0 \vee \widehat{R}_S(h_2) = 0 \vee ... \vee \widehat{R}_S(|\mathcal{H}|) = 0]$$
$$\leq \sum_{h\in\mathcal{H}_\epsilon} \mathbb{P}[\widehat{R}_S(h) = 0] \leq |\mathcal{H}|(1-\epsilon)^m \leq |\mathcal{H}|e^{-m\epsilon}$$
$$\Rightarrow \mathbb{P}_{S\sim D^m}[R(h_S) \leq \epsilon] = \mathbb{P}[h_S \notin \mathcal{H}_\epsilon | \widehat{R}_S(h_S) = 0] = 1 - \mathbb{P}[h_S \in \mathcal{H}_\epsilon | \widehat{R}_S(h_S) = 0] \geq 1 - \delta$$

Corollary 2.10: Fix $\epsilon > 0$. Then, for any hypothesis $h : \mathcal{X} \to \{0,1\}$, we have

$$\mathbb{P}_{S\sim\mathcal{D}^m}[\widehat{R}_S(h) - R(h) \geq \epsilon] \leq e^{-2m\epsilon^2}$$

and

$$\mathbb{P}_{S\sim\mathcal{D}^m}[\widehat{R}_S(h) - R(h) \leq -\epsilon] \leq e^{-2m\epsilon^2}$$

hence

$$\mathbb{P}_{S\sim\mathcal{D}^m}[|\widehat{R}_S(h) - R(h)| \geq \epsilon] \leq 2e^{-2m\epsilon^2}$$

Proof: Use Hoeffding's Lemma ($E[e^{tX}] \leq e^{\frac{t^2(b-a)^2}{8}}$) and the Chernoff Bounding technique ($\mathbb{P}[X \geq \epsilon] = \mathbb{P}[e^{tX} \geq e^{t\epsilon}] \leq e^{-t\epsilon}E[e^{tX}]$) for Hoeffding's Inequality

($\mathbb{P}[X - E[X] \geq \epsilon] \leq e^{\frac{-2\epsilon^2}{\sum_{i=1}^m (a_i - b_i)^2}}$ for $X = \sum_{i=1}^m X_i$ with $X_i \in (a_i, b_i)$). Note that here $\widehat{R}_S(h) = \frac{1}{m}\sum_{i=1}^m \chi_{h(x)\neq c(x)}$ so that the value $\sum_{i=1}^m (a_i - b_i)^2$ in this case is equal to $\sum_{i=1}^m (\frac{1-0}{m})^2 = m \cdot \frac{1}{m^2} = \frac{1}{m}$.

Corollary 2.11 (Generalization Bound): Set $2\epsilon^{-2m\epsilon^2} = \delta$ in the previous part.

Theorem 2.13 (Learning bound – finite, $\mathcal{H}$ inconsistent case): Let $\mathcal{H}$ be a finite hypothesis set. Then, for any $\delta > 0$ and any $h \in \mathcal{H}$, we have

$$\mathbb{P}\left[R(h) \leq \widehat{R}_S(h) + \sqrt{\frac{\log|\mathcal{H}| + \log\frac{2}{\delta}}{2m}}\right] \geq 1 - \delta$$

.

Proof: We find that

$$\begin{aligned} &\mathbb{P}[\exists h \in \mathcal{H} :\ R(h) - \widehat{R}_S(h) > \epsilon] \\ &= \mathbb{P}[(R(h_1) - \widehat{R}_S(h_1) > \epsilon) \vee ... \vee (R(h_{|\mathcal{H}|}) - \widehat{R}_S(h_{|\mathcal{H}|}) > \epsilon)] \\ &\leq \sum_{i=1}^{|\mathcal{H}|} \mathbb{P}[R(h_i) - \widehat{R}_S(h_i) > \epsilon] \leq 2|\mathcal{H}|e^{-2m\epsilon^2} \end{aligned}$$

so then

$$\delta := 2|\mathcal{H}|e^{-2m\epsilon^2} \Rightarrow -2m\epsilon^2 = \log\frac{\delta}{2|\mathcal{H}|} \Rightarrow \epsilon = \sqrt{\frac{-\log\frac{\delta}{2|\mathcal{H}|}}{2m}} = \sqrt{\frac{\log|\mathcal{H}| + \log\frac{2}{\delta}}{2m}}$$

Definition (Agnostic PAC-learning): Let $\mathcal{H}$ be a hypothesis set. Then, $\mathcal{A}$ is an agnostic PAC-learning algorithm if there exists a polynomial function poly(.,.,.,.) such that for any $\epsilon, \delta > 0$ and any distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$,

$$m \geq \text{poly}(\frac{1}{\epsilon}, \frac{1}{\delta}, n, \text{size}(c)) \Rightarrow \mathbb{P}_{S\sim\mathcal{D}^m}[R(h_S) - \min_{h\in\mathcal{H}} R(h) \leq \epsilon] \geq 1 - \delta$$

Note further that if $\mathcal{A}$ is poly$(\frac{1}{\epsilon}, \frac{1}{\delta}, n, \text{size}(c))$, it is said to be an "efficient agnostic PAC-learning algorithm".

Definition: A scenario is "deterministic" if the label of a point can be uniquely determined by some measurable function $f : \mathcal{X} \to \mathcal{Y}$ with probability 1.

Definition (Bayes Error) Given a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, the Bayes Error

$$R^* := \inf_{\substack{h:\mathcal{X}\to\mathcal{Y} \\ h \text{ measurable}}} R(h)$$

satisfies $R^* = 0$ in the deterministic case, and $R^* \neq 0$ in the stochastic case. A hypothesis h with $R(h) = R^*$ is called a "Bayes classifier".

Ch. 2 Exercises.

2.2. An axis-aligned hyper-rectangle in $\mathbb{R}^n$ is a set of the form $[a_1, b_1] \times ... \times [a_n, b_n]$. Suppose the set of all instances belong in $\mathcal{X} = \mathbb{R}^n$ and $\mathcal{C}$ is the set of all axis-aligned hyper-rectangles in $\mathbb{R}^n$.

Let $R \in \mathcal{C}$ be a target concept and fix $\epsilon > 0$ so that $\mathbb{P}[R] > \epsilon$ (or else the algorithm presented below works immediately). Let $a_1, ..., a_n$ and $b_1, ..., b_n$ be $2n$ real values defining $R = [a_1, b_1] \times ... \times [a_n, b_n]$. We then define rectangles on the perimeter as $R_{i,0} := [a_1, b_1] \times ... \times [r_i, b_i] \times ... \times [a_n, b_n]$ and $R_{i,1} := [a_1, b_1] \times ... \times [a_i, r_i] \times ... \times [a_n, b_n]$ such that $r_i = \inf\{r \in \mathbb{R} : \mathbb{P}[[a_1, b_1] \times ... \times [a_i, r] \times ... \times [a_n, b_n]] \geq \frac{\epsilon}{2n}\}$.

We define our algorithm $\mathcal{A}$ as returning the tightest axis-aligned hyper-rectangle R_S containing the points labeled with 1. If $R(R_S) > \epsilon$, R_S must miss at least one rectangle R_i so that

$$\mathbb{P}_{S\sim\mathcal{D}^m}[R(R_S) > \epsilon] \leq \mathbb{P}_{S\sim\mathcal{D}^m}[\bigcup_{i=1}^{n}\bigcup_{j=0}^{1}\{R_S \cap R_{i,j} = \emptyset\}] \leq \sum_{i=1}^{n}\sum_{j=0}^{1}\mathbb{P}_{S\sim\mathcal{D}^m}[\{R_S \cap R_{i,j} = \emptyset\}]$$

$$\leq \sum_{i=1}^{n} 2(1 - \frac{\epsilon}{2n})^m = 2n(1 - \frac{\epsilon}{2n})^m = 2ne^{m\log(1-\frac{\epsilon}{2n})} \leq 2ne^{-\frac{m\epsilon}{2n}}$$

Hence,

$$\delta \geq 2ne^{-\frac{m\epsilon}{2n}} \iff m \geq \frac{2n}{\epsilon}\log\frac{2n}{\delta}$$

so that $\mathcal{C}$ is PAC-learnable.

2.3. Let $\mathcal{X} = \mathbb{R}^2$ and consider the class $\mathcal{C}$ of concepts of the form $c = \{(x, y) : x^2 + y^2 \leq r^2\}$ for some $r \in \mathbb{R}$. We fix $C \in \mathcal{C}$ as a target concept, along with an $\epsilon > 0$, and we define our algorithm $\mathcal{A}$ as that which returns the infimum of circles containing the points labeled with 1. We denote this infimum as C_S.

We then define the circle C_0 as $C_0 = \text{argmax}_{c\in\mathcal{C}}\{\mathbb{P}[c \backslash C_s] : \mathbb{P}[c \backslash C_s] \leq \epsilon\}$. Therefore, if $R(C_S) > \epsilon$, then $C_S \cap C_0 = \emptyset$, so that

$$\mathbb{P}_{S\sim\mathcal{D}^m}[R(C_S) > \epsilon] \leq \mathbb{P}_{S\sim\mathcal{D}^m}[C_S \cap C_0 = \emptyset] = (1 - \epsilon)^m \leq e^{-m\epsilon}$$

Hence,

$$\delta \geq e^{-m\epsilon} \iff \log\frac{1}{\delta} \leq m\epsilon \iff m \geq (\frac{1}{\epsilon})\log\frac{1}{\delta}$$

as desired.

2.4. Let $\mathcal{X} = \mathbb{R}^2$ and consider the set of concepts of the form $c = \{x \in \mathbb{R}^2 : ||x - x_0|| \leq r\}$ for some $x_0 \in \mathbb{R}^2$ and $r \in \mathbb{R}$. Suppose the target concept $c_0 \in \mathcal{C}$ has $\mathbb{P}[c_0] = k > 0$ and radius r_0 for some $k, r_0 \in \mathbb{R}$. If $p \in r_1 \cap r_2$ and $\ell \in \mathbb{R}^2$ is a line which passes through the intersection $r_1 \cap r_2$, we consider a translation of the circle along ℓ from p toward the center of the circle. In particular, a translation $c' := c_0 + \frac{r_0}{2}$ intersects each of the three regions r_i yet maintains an error of at least $\frac{k}{2}$ so that Gertrude's method does not work.

2.6. Consider now the case where the training points recieved by the learner are subject to the following noise: points labeled positively are randomly flipped to negative with probability less than $\eta' < 1/2$. We again consider the algorithm $\mathcal{A}$ which returns the tightest rectangle containing positive points.

a) For a target concept R we can again assume $\mathbb{P}[R] > \epsilon$. Now suppose that $R(R') > \epsilon$. Then, the probability that R' (due to $\mathcal{A}$) misses a region r_j for $j \in [4]$ is at most $(1 - \frac{\epsilon}{4})^{m\eta'}$ for a sample S of size m.

b) Hence, $\mathbb{P}[R(R') > \epsilon] \leq 4(1 - \frac{\epsilon}{4})^{m\eta'} = 4e^{m\eta' \log(1-\frac{\epsilon}{4})} \leq 4e^{-\frac{m\eta'\epsilon}{4}}$ so that $\delta \geq 4e^{-\frac{m\eta'\epsilon}{4}}$ yields a sample complexity bound of $m \geq \frac{4\log\frac{4}{\delta}}{\epsilon\eta'}$.

2.7. Consider a finite hypothesis set $\mathcal{H}$, assume that the target concept is in $\mathcal{H}$ and that the label of a training point received by the learner is randomly changed with probability $\eta \in (0, \frac{1}{2})$ where $\eta \leq \eta' < \frac{1}{2}$.

a) For any $h \in \mathcal{H}$, let $d(h)$ denote the probability that the label of a training point received by the learner disagrees with the one given by h. Let h^* be the target hypothesis. Since the learner will error with probability η (assuming $R(h) = 0$), we have $d(h^*) = \eta$.

Chapter 3 Notes

Definition: We define $\mathcal{G} := \{g : (x, y) \to L(h(x), y) \mid h \in \mathcal{H}\}$ as a family of loss functions $L : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ and let $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$. Note that many results below hold for arbitrary loss functions $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$.

Definition (Empirical Rademacher Complexity): Let $\mathcal{G}$ be a family of functions mapping from $\mathcal{Z}$ to $[a, b]$ and $S := (z_1, ..., z_m)$ a fixed sample in $\mathcal{Z}$. Then, the Rademacher complexity of $\mathcal{G}$ with respect to sample S is given by

$$\widehat{\mathfrak{R}}_S(\mathcal{G}) = E_\sigma\Big[\sup_{g\in\mathcal{G}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i g(z_i)\Big] = E_\sigma\Big[\sup_{g\in\mathcal{G}} \frac{\sigma \cdot g_S}{m}\Big]$$

where $\sigma := (\sigma_1, ..., \sigma_m)^T$ with independent uniform random variables (Rademacher variables) $\sigma_i \in \{-1, 1\}$, and $g_S := (g(z_1), ..., g(z_m))^T$.

Definition (Rademacher Complexity): Let $\mathcal{D}$ denote the distribution according to which samples are drawn. For $m \in \mathbb{N}$ with $m \geq 1$, we define

$$\mathfrak{R}_m(\mathcal{G}) := E_{S\sim\mathcal{D}^m}[\widehat{\mathfrak{R}}_S(\mathcal{G})]$$

Intuitively, Rademacher Complexity measures how robust a class of loss functions is, as a higher $\widehat{\mathfrak{R}}_S(\mathcal{G})$ for a set S indicates a space of functions more adaptable to arbitrary labelings.

Definition (Martingale Difference Sequence): A sequence of random variables $V_1, V_2, ...$ is a martingale difference sequence with respect to $X_1, X_2, ...$ if for any $i > 0$, V_i is a function of $X_1, ...X_i$ and $E[V_{i+1}|X_1, ..., X_i] = 0$.

Lemma D.6 Let V, Z be random variables such that $E[V|Z] = 0$ and for some function f and constant $c \geq 0$, $f(Z) \leq V \leq f(Z) + c$. Then $t > 0 \Rightarrow E[e^{tV}|Z] \leq e^{\frac{t^2c^2}{8}}$

Proof: Repeat the proof of Hoeffding's Lemma but with conditional expectations.

Theorem D.7 (Azuma's Inequality): Let $V_1, V_2, ...$ be a martingale difference sequence with respect to random variables $X_1, X_2, ...$ and assume that for any $i > 0$ there exists $c_i \geq 0$ and a random variable $Z_i(X_1, ..., X_{i-1})$ such that $Z_i \leq V_i \leq Z_i + c_i$. Then for any $\epsilon > 0$ and $m \in \mathbb{N}$,

$$\mathbb{P}[\sum_{i=1}^{m} V_i \geq \epsilon] \leq e^{\frac{-2\epsilon^2}{\sum_{i=1}^m c_i^2}}$$

and

$$\mathbb{P}[\sum_{i=1}^{m} V_i \leq -\epsilon] \leq e^{\frac{-2\epsilon^2}{\sum_{i=1}^m c_i^2}}$$

Proof: Using Lemma D.6, we find that $S_m := \sum_{i=1}^m V_i$ we have that $\mathbb{P}[S_m \geq \epsilon] = \mathbb{P}[e^{tS_m} \geq e^{t\epsilon}] \leq e^{-t\epsilon} E[e^{tS_m}] = e^{-t\epsilon} E[e^{tS_{m-1}}] E[e^{tV_m}|X_1, ..., X_{m-1}] \leq e^{-t\epsilon} E[e^{tS_{m-1}}] e^{\frac{t^2 c_m^2}{8}} \leq e^{-t\epsilon} e^{\frac{t^2 \sum_{i=1}^m c_i^2}{8}}$. We then choose $t = \frac{4\epsilon}{\sum_{i=1}^m c_i^2}$ and repeat for the other inequality.

Theorem D.8 (McDiarmid's Inequality) Let $X_1, ..., X_m \in \mathcal{X}^m$ be a set of $m \geq 1$ independent random variables and suppose there exists $c_1, ..., c_m > 0$ such that $f : X^m \to \mathbb{R}$ satisfies

$$|f(x_1, ..., x_i, ..., x_m) - f(x_1, ..., x_i', ..., x_m)| \leq c_i$$

for any $i \in [m]$ and $x_1, ..., x_m, x_i' \in \mathcal{X}^m$. Then for $f(S) := f(X_1, ..., X_m)$ and any $\epsilon > 0$ we have

$$\mathbb{P}[f(S) - E[f(S)] \geq \epsilon] \leq e^{\frac{-2\epsilon^2}{\sum_{i=1}^m c_i^2}}$$

and

$$\mathbb{P}[f(S) - E[f(S)] \leq -\epsilon] \leq e^{\frac{-2\epsilon^2}{\sum_{i=1}^m c_i^2}}$$

Proof: We define variables $V = f(S) - E[f(S)]$ and $V_k = E[V|X_1, ..., X_k] - E[V|X_1, ..., X_{k-1}]$. Then, $E[V_k|X_1, ..., X_{k-1}] = E[E[V|X_1, ..., X_k] - E[V|X_1, ..., X_{k-1}]|X_1, ..., X_{k-1}] = 0$ so that the V_k are a martingale difference sequence. Then, we define

$$L_k := \inf_x E[V|X_1, ..., X_{k-1}, x] - E[V|X_1, ..., X_{k-1}]$$

and

$$U_k := \sup_x E[V|X_1, ..., X_{k-1}, x] - E[V|X_1, ..., X_{k-1}]$$

so that $U_k - L_k \leq \sup_{x,x'} E[V|X_1, ..., X_{k-1}, x] - E[V|X_1, ..., X_{k-1}, x'] \leq c_k$ so that $L_k \leq V_k \leq L_k + c_k$ and we may apply Azuma's Inequality.

Theorem 3.3 For $\mathcal{G}$ a family of functions mapping $\mathcal{Z}$ to $[0,1]$, for any $\delta > 0$ and $g \in \mathcal{G}$ we have

$$\mathbb{P}\left[E[g(z)] \leq \frac{1}{m}\sum_{i=1}^{m} g(z_i) + 2\mathfrak{R}_m(\mathcal{G}) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}\right] \geq 1-\delta$$

$$\mathbb{P}\left[E[g(z)] \leq \frac{1}{m}\sum_{i=1}^{m} g(z_i) + 2\widehat{\mathfrak{R}}_S(\mathcal{G}) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2m}}\right] \geq 1-\delta$$

Proof: For any sample $S = (z_1, ..., z_m)$ and $g \in \mathcal{G}$, denote $\widehat{E}_S[g] := \frac{1}{m}\sum_{i=1}^{m} g(z_i)$. We then define

$$\Phi(S) := \sup_{g\in\mathcal{G}}(E[g] - \widehat{E}_S[g])$$

Let S, S' be two different samples (differing by z_m in S and z'_m in S') so

$$\Phi(S') - \Phi(S) \leq \sup_{g\in\mathcal{G}}(E[g] - E[g] - \widehat{E}_S[g] + \widehat{E}_S[g]) \leq \sup_{g\in\mathcal{G}} \frac{g(z_m) - g(z'_m)}{m} \leq \frac{1}{m}$$

Repeating the argument for $\phi(S') - \phi(S)$, we get $|\Phi(S) - \Phi(S')| \leq \frac{1}{m}$. Then, by McDiarmid's Inequality we have

$$\mathbb{P}[\Phi(S) - E[\Phi(S)] \leq \epsilon] \leq e^{\frac{-2\epsilon^2}{\sum_{i=1}^{m}\frac{1}{m^2}}} = e^{-2\epsilon^2 m}$$

. Note further that

$$\frac{\delta}{2} := e^{-2\epsilon^2 m} \Rightarrow \epsilon = \sqrt{\frac{\log\frac{2}{\delta}}{2m}}$$

. Then,

$$E_S[\Phi(S)] = E_S[\sup_{g\in\mathcal{G}}(E[g] - \widehat{E}_S[g])] = E_S[\sup_{g\in\mathcal{G}}(E_{S'}[\widehat{E}_{S'}[g] - \widehat{E}_S[g]])]$$

$$\leq E_{S,S'}[\sup_{g\in\mathcal{G}}(\widehat{E}_{S'}[g] - \widehat{E}_S[g])] = E_{S,S'}[\sup_{g\in\mathcal{G}}(\frac{1}{m}\sum_{i=1}^{m} g(z'_i) - g(z_i))]$$

$$= E_{S,S',\sigma}[\sup_{g\in\mathcal{G}}(\frac{1}{m}\sum_{i=1}^{m}\sigma_i(g(z'_i) - g(z_i)))]$$

$$\leq E_{S',\sigma}[\sup_{g\in\mathcal{G}}(\frac{1}{m}\sum_{i=1}^{m}\sigma_i g(z'_i))] + E_{S,\sigma}[\sup_{g\in\mathcal{G}}(\frac{1}{m}\sum_{i=1}^{m}\sigma_i g(z_i))] = 2\mathfrak{R}_m(\mathcal{G})$$

We then note that, for sets S and S' differing by one point,

$$|\widehat{\mathfrak{R}}_S(\mathcal{G}) - \widehat{\mathfrak{R}}_{S'}(\mathcal{G})| \leq \frac{1}{m}$$

so again by McDiarmid's we have

$$\mathbb{P}[\mathfrak{R}_m(\mathcal{G}) - \widehat{\mathfrak{R}}_{S'}(\mathcal{G}) \geq \epsilon] \leq e^{-2m\epsilon^2}$$

hence

$$\frac{\delta}{2} = e^{-2m\epsilon^2} \Rightarrow \Phi(S) \leq 2\widehat{\mathfrak{R}}_S(\mathcal{G}) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2m}}$$

Lemma 3.4: Let $\mathcal{H}$ be a family of functions taking values in $\{-1, 1\}$, and let $\mathcal{G}$ be a family of loss functions "associated to $\mathcal{H}$ for the zero-one loss", i.e. $\mathcal{G} = \{(x, y) \mapsto \chi_{h(x) \neq y} \mid h \in \mathcal{H}\}$. For any sample $S = ((x_1, y_1), ..., (x_m, y_m))$ of elements in $\mathcal{X} \times \{-1, 1\}$, let $S_{\mathcal{X}} = (x_1, ..., x_m)$. Then, $\widehat{\mathfrak{R}}_S(\mathcal{G}) = \frac{1}{2}\widehat{\mathfrak{R}}_{S_{\mathcal{X}}}(\mathcal{H})$

Proof: We have that

$$\widehat{\mathfrak{R}}_S(\mathcal{G}) = E_\sigma[\sup_{h \in \mathcal{H}}(\frac{1}{m}\sum_{i=1}^{m}\sigma_i \chi_{h(x_i) \neq y_i})]$$

$$= E_\sigma[\frac{1}{m}\sup_{h \in \mathcal{H}}(\sum_{i=1}^{m}\sigma_i \frac{1 - h(x_i)y_i}{2})] = E_\sigma[\frac{1}{2m}\sup_{h \in \mathcal{H}}(\sum_{i=1}^{m}\sigma_i - h(x_i)y_i)]$$

$$= \frac{1}{2}E_\sigma[\sup_{h \in \mathcal{H}}(\frac{1}{m}\sum_{i=1}^{m}\sigma_i h(x_i))] = \frac{1}{2}\widehat{\mathfrak{R}}_{S_{\mathcal{X}}}(\mathcal{H})$$

Theorem 3.5: For a family of functions $\mathcal{H}$ taking values in $\{-1, 1\}$ and $\mathcal{D}$ a distribution over $\mathcal{X}$ (the input space), then for any $\delta > 0$ and any $h \in \mathcal{X}$, over a sample S of size m drawn according to $\mathcal{D}$, we have

$$\mathbb{P}\left[R(h) \leq \widehat{R}_S(h) + \mathfrak{R}_m(\mathcal{H}) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}\right] \geq 1 - \delta$$

$$\mathbb{P}\left[R(h) \leq \widehat{R}_S(h) + \widehat{\mathfrak{R}}_S(\mathcal{H}) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2m}}\right] \geq 1 - \delta$$

Proof: We consider the functions $g : (x, y) \to 1_{h(x) \neq y}$ so that $E[g(z)] = R(h)$ and $\widehat{R}_S(h) = \frac{1}{m}\sum_{i=1}^{m} g(z_i)$. Further, $\widehat{\mathfrak{R}}_s(\mathcal{G}) = \frac{1}{2}\widehat{\mathfrak{R}}_{S_{\mathcal{X}}}(\mathcal{H})$ so that $\mathfrak{R}_m(\mathcal{G}) = \frac{1}{2}\mathfrak{R}_m(\mathcal{H})$. We then combine Theorem 3.3 with Lemma 3.4.

Note:

$$\widehat{\mathfrak{R}}_S(\mathcal{H}) = E_\sigma[\sup_{h \in \mathcal{H}}\frac{1}{m}\sum_{i=1}^{m} -\sigma_i h(x_i)] = -E_\sigma[\inf_{h \in \mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i h(x_i)]$$

which then calculates the negative expectation over sigma of "empirical risk minimization", which is computationally hard for some $\mathcal{H}$.

Definition: The growth function $\Pi_{\mathcal{H}} : \mathbb{N} \to \mathbb{N}$ is defined as

$$\Pi_{\mathcal{H}}(m) = \max_{(x_1, ..., x_m) \subset \mathcal{X}} |\{h(x_1), ..., h(x_m)\} \,:\, h \in \mathcal{H}|$$

where each such distinct classification is referred to as a "dichotomy".

Maximal Inequality: Let $X_1, ..., X_n$ be $n \geq 1$ real-valued random variables such that, for any $j \in [n]$ and $t > 0$, $E[e^{tX_j} \leq e^{\frac{t^2 r^2}{2}}]$ for some $r > 0$. Then, $E[\max_{j \in [n]} X_j] \leq r\sqrt{2\log n}$

Proof: We have that

$$e^{tE[\max_{j\in[n]} X_j]} \leq E[\max_{j\in[n]} e^{tX_j}] \leq \sum_{j=1}^{n} E[e^{tX_j}] \leq ne^{\frac{t^2r^2}{2}}$$

then for $t = \frac{\sqrt{2\log n}}{r}$,

$$E[\max_{j\in[n]} X_j] \leq \frac{\log n + \frac{t^2r^2}{2}}{t} = r\sqrt{2\log n}$$

Corollary D.11: Let $X_1, ..., X_n$ be $n \geq 1$ real-valued random variables such that, for any $j \in [n]$, $X_j = \sum_{i=1}^{m} Y_{ij}$. Suppose that for fixed $j \in [n]$, Y_{ij} are independent, zero mean random variables taking values in $[-r_i, r_i]$ for some $r_i > 0$. Then, $E[\max_{j\in[n]} X_j] \leq \sqrt{2\log(n)\sum_{i=1}^{m} r_i^2}$

Proof: We find that

$$E[e^{tX_j}] = \prod_i E[e^{tY_{ij}}] \leq \prod_i e^{\frac{t^2(2r_i)^2}{8}}$$

hence

$$E[e^{tX_j}] \leq \frac{t\sum_i r_i^2}{2}$$

so that we may apply the Maximal Inequality for $r = \sqrt{\sum_{i=1}^{m} r_i^2}$

Theorem 3.7 (Massart's Lemma): Let $A \subset \mathbb{R}^m$ be a finite set such that $r := \max_{x\in A} ||x||_2$. Then,

$$E_\sigma[\frac{1}{m}\sup_{x\in A}\sum_{i=1}^{m} \sigma_i x_i] \leq \frac{r\sqrt{2\log|A|}}{m}$$

where the $\sigma_i \in \{-1, 1\}$ are independent uniform random variables and $x_1, ..., x_m$ are components of x.

Proof: Apply Corollary D.11 to $X_i = \frac{1}{m}\sum_{j=1}^{m} \sigma_i x_j^i$ for $i \in [|A|]$, noting that each $\sigma_i x_j^i \in \{-|x_j^i|, |x_j^i|\}$ hence $\sum_{i=1}^{m} |x_i|^2 \leq r^2$.

Corollary 3.8: Let $\mathcal{G}$ be a family of functions taking values in $\{-1, 1\}$. Then,

$$\mathfrak{R}_m(\mathcal{G}) \leq \sqrt{\frac{2\log\Pi_{\mathcal{G}}(m)}{m}}$$

Proof: For a fixed sample $S = (z_1, ..., z_m)$, we have

$$\widehat{\mathfrak{R}}_S(\mathcal{G}) = E_\sigma\Big[\sup_{g\in\mathcal{G}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i g(z_i)\Big] \leq \frac{\sqrt{m}\sqrt{2\log\Pi_{\mathcal{G}}(m)}}{m}$$

so the expectation is bounded similarly.

Corollary 3.9: For a family of functions $\mathcal{H}$ valued in $\{-1, 1\}$, for any $\delta > 0$ and any $h \in \mathcal{H}$,

$$\mathbb{P}\left[R(h) \leq \widehat{R}_S(h) + \sqrt{\frac{2 \log \Pi_{\mathcal{H}}(m)}{m}} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}\right] \geq 1 - \delta$$

where we use the Rademacher complexity bound from Corollary 3.8 and Theorem 3.5.

Definition: A set S of $m \geq 1$ points is "shattered" by a hypothesis set $\mathcal{H}$ if $\mathcal{H}$ realizes all possible dichotomies of S, i.e. $\Pi_{\mathcal{H}}(m) = 2^m$.

Definition (VC-dimension): The VC-dimension of a hypothesis set $\mathcal{H}$ is the size of the largest set that can be shattered by $\mathcal{H}$, i.e.

$$\text{VCdim}(\mathcal{H}) = \max\{m \in \mathbb{N} : \Pi_{\mathcal{H}}(m) = 2^m\}$$

Example: Consider the $d+1$ points $x_i := (0, ..., 1, ..., 0)$ for $i \in \{0, 1, ..., d\}$ where the 1 is in the i-th position and x_0 is the origin. Further, let $w = (y_0, y_1, ..., y_d)$ where $y_i \in \{-1, 1\}$. Then, the hyperplane defined as

$$w \cdot x + \frac{y_0}{2} = 0$$

satisfies

$$\text{sgn}(w \cdot x_i + \frac{y_0}{2}) = y_i$$

for $i \in \{1, ..., d\}$ and

$$\text{sgn}(w \cdot x_0 + \frac{y_0}{2}) = y_0$$

hence the VC-dimension of hyperplanes in $\mathbb{R}^d$ is at least $d+1$.

Definition: The convex hull $\text{conv}(\mathcal{X})$ of $\mathcal{X} \subset \mathbb{R}^N$ is defined as

$$\text{conv}(\mathcal{X}) = \left\{ \sum_{i=1}^{|\mathcal{X}|} \alpha_i x_i \mid \sum_{i=1}^{|\mathcal{X}|} \alpha_i = 1,\ x_i \in \mathcal{X},\ \alpha_i \geq 0 \right\}$$

Radon's Theorem: Any set $\mathcal{X}$ of $d+2$ points in $\mathbb{R}^d$ can be partitioned into two subsets $\mathcal{X}_1$ and $\mathcal{X}_2$ such that $\text{conv}(\mathcal{X}_1) \cap \text{conv}(\mathcal{X}_2) \neq \emptyset$

Proof: Let $\mathcal{X} = \{x_1, ..., x_{d+2}\} \subset \mathbb{R}^d$. We find that the system

$$\sum_{i=1}^{d+2} \alpha_i x_i = 0, \ \sum_{i=1}^{d+2} \alpha_i = 0$$

has $d+1$ independent equations and $d+2$ unknowns, so that there exists a non-zero solution $\beta_1, ..., \beta_{d+2}$. Since $\sum_{i=1}^{d+2} \beta_i = 0$, the sets

$$\mathcal{J}_1 := \{i \in [d+2] \mid \beta_i \leq 0\}, \ \ \mathcal{J}_2 := \{i \in [d+2] \mid \beta_i > 0\}$$

are nonempty and they satisfy

$$\sum_{i \in \mathcal{J}_1} \beta_i x_i = -\sum_{i \in \mathcal{J}_2} \beta_i x_i$$

so that

$$\beta := \sum_{i \in \mathcal{J}_1} \beta_i \Rightarrow \frac{1}{\beta} \sum_{i \in \mathcal{J}_1} \beta_i x_i$$

belongs in the convex hulls of both $\mathcal{X}_1$ and $\mathcal{X}_2$.

Theorem 3.17 (Sauer's Lemma): Let $\mathcal{H}$ be a hypothesis set such that $\text{VCdim}(\mathcal{H}) = d$. Then, for any $m \in \mathbb{N}$, $\Pi_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i}$

Proof: We proceed by induction. The statement holds for $m = 1$ and $d = 1$ or $d = 0$. Then, assume the statement holds for $(m-1, d)$ and $(m-1, d-1)$. We then fix a sample S of size m given by $S = (x_1, ..., x_m)$. Let $\mathcal{G}$ denote the space of hypotheses due to S. Identifying each $g \in \mathcal{G}$ with those x_i classified as 1 (rather than -1), let $\mathcal{G}_1$ denote the space of hypotheses due to $(x_1, ..., x_{m-1})$ and let $\mathcal{G}_2$ denote those $g \in \mathcal{G}$ such that if $Z \subset \{0,1\}^{m-1}$ is expressed among the $\{x_1, ..., x_{m-1}\}$, so is $Z \cup x_m$. Hence, $|\mathcal{G}| = |\mathcal{G}_1| + |\mathcal{G}_2|$. Since $\mathcal{G}_1$ has VC dimension at most d while $\mathcal{G}_2$ has VC dimension at most $d-1$ (else $\mathcal{G}$ would also shatter a set of size $d+1$ by adding x_m). Therefore,

$$|\mathcal{G}| \leq \sum_{i=0}^{d-1} \binom{m-1}{i} + \sum_{i=0}^{d} \binom{m-1}{i}$$

$$= \sum_{i=1}^{d} \binom{m-1}{i-1} + \sum_{i=1}^{d} \binom{m-1}{i} = \sum_{i=0}^{d} \binom{m}{i}$$

Corollary 3.18: Let $\mathcal{H}$ be a hypothesis set such that $\text{VCdim}(\mathcal{H}) = d$. Then, for any $m \geq d$, $\Pi_{\mathcal{H}}(m) \leq \left(\frac{em}{d}\right)^d = O(m^d)$

Proof: From Sauer's Lemma, we have that

$$\Pi_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i} \leq \sum_{i=0}^{d} \binom{m}{i} \left(\frac{m}{d}\right)^{d-i} \leq \sum_{i=0}^{m} \binom{m}{i} \left(\frac{m}{d}\right)^{d-i}$$

$$= \left(\frac{m}{d}\right)^d \sum_{i=0}^{m} \binom{m}{i} \left(\frac{d}{m}\right)^i = (\frac{m}{d})^d (1 + \frac{d}{m})^m \leq \left(\frac{em}{d}\right)^d$$

Corollary 3.19: Let $\mathcal{H}$ be a family of functions taking values in $\{-1, 1\}$ with VC-dimension d. Then, for any $\delta > 0$,

$$\mathbb{P}\left[R(h) \leq \widehat{R}_S(h) + \sqrt{\frac{2d \log \frac{em}{d}}{m}} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}\right] \geq 1 - \delta$$

Proof: Combine Corollary 3.18 and Corollary 3.9.

Definition (Relative Entropy): The relative entropy (or Kullback Leibler Divergence) of 2 distributions p and q is denoted $D(p||q)$, and is defined by

$$D(p||q) = E_p\Big[\log \frac{p(x)}{q(x)}\Big] = \sum_{x \in \mathcal{X}} p(x) \log(\frac{p(x)}{q(x)})$$

Sanov's Theorem (D.3): Let $X_1, ..., X_m$ be independent variables drawn according to some distribution $\mathcal{D}$ with mean p and support included in $[0,1]$. Then, for $\widehat{p} := \frac{1}{m}\sum_{i=1}^m X_i$ and any $q \in [0,1]$, we have

$$\mathbb{P}[\widehat{p} \geq q] \leq e^{-mD(p||q)}$$

Proof: We have

$$\mathbb{P}[\widehat{p} \geq q] \leq e^{-tmq} E[e^{tm\widehat{p}}] = e^{-tmq} \prod_{i=1}^{m} E[e^{tX_i}] \leq e^{-tmq}\Big(1 - p + pe^t\Big)^m$$

$$= \Big((1-p)e^{-q\log\frac{q(1-p)}{p(1-q)}} + pe^{(1-q)\log\frac{q(1-p)}{p(1-q)}}\Big)^m = e^{m(-q\log\frac{q}{p} + (q-1)\log\frac{1-q}{1-p})}$$

where $t \geq 0$ is used for the Chernoff bound

Theorem D.4: Let $X_1, ..., X_m$ be independent random variables drawn according to some distribution $\mathcal{D}$ with mean p and support included in $[0,1]$. Then, for any $\gamma \in [0, \frac{1}{p} - 1]$, for $\widehat{p} := \frac{1}{m}\sum_{i=1}^m X_i$, we have

$$\mathbb{P}[\widehat{p} \geq (1+\gamma)p] \leq e^{\frac{-mp\gamma^2}{3}}$$

and

$$\mathbb{P}[\widehat{p} \leq (1-\gamma)p] \leq e^{\frac{-mp\gamma^2}{2}}$$

Proof: For $q = (1+\gamma)p$,

$$D(q||p) = (1+\gamma)p\log\frac{p}{(1+\gamma)p} + (1-(1+\gamma)p)\log\frac{1-p}{1-(1+\gamma)p}$$

$$= -p(1+\gamma)\log(1+\gamma) + (1-(1+\gamma)p)\log(1 + \frac{\gamma p}{1-(1+\gamma)p})$$

$$\leq (1+\gamma)p\frac{-\gamma}{1+\frac{\gamma}{2}} + (1-p-\gamma p)\frac{\gamma p}{1-p-\gamma p} = -\gamma p\Big(1 + \frac{\frac{\gamma}{2}}{1+\frac{\gamma}{2}} - 1\Big) = -\frac{\gamma^2 p}{2+\gamma} \leq -\frac{\gamma^2 p}{3}$$

For $q = (1-\gamma)p$, we have

$$D(q||p) = (1-\gamma)p\log\frac{p}{(1-\gamma)p} + (1-(1-\gamma)p)\log\frac{1-p}{1-(1-\gamma)p}$$

$$= -p(1-\gamma)\log(1-\gamma) + (1-(1-\gamma)p)\log(1 - \frac{\gamma p}{1-(1-\gamma)p})$$

$$\leq (1-\gamma)p\frac{\gamma}{1-\frac{\gamma}{2}} + (1-p+\gamma p)\frac{-\gamma p}{1-p+\gamma p} = \gamma p(\frac{1-\gamma}{1-\frac{\gamma}{2}} - 1) = -\frac{\gamma^2 p}{2-\gamma} \leq -\frac{\gamma^2 p}{2}$$

Theorem 3.20: Let $\mathcal{H}$ be a hypothesis set with VC dimension $d > 1$. Then, for any $m \geq 1$ and any learning algorithm $\mathcal{A}$, there exists a distribution $\mathcal{D}$ over $\mathcal{X}$ and a target function $f \in \mathcal{H}$ such that

$$\mathbb{P}[R_{\mathcal{D}}(h_S, f) > \frac{d-1}{32m}] \geq \frac{1}{100}$$

Proof: Let $\overline{\mathcal{X}} = \{x_0, ..., x_{d-1}\} \subset \mathcal{X}$ be shattered by $\mathcal{H}$. For any $\epsilon > 0$, choose $\mathcal{D}$ such that its support is reduced to $\overline{\mathcal{X}}$ and so that one point (x_0) has probability $1 - 8\epsilon$ with the rest of the mass distributed uniformly, i.e. $\mathbb{P}_{\mathcal{D}}[x_0] = 1 - 8\epsilon$ and for any $i \in [d-1]$, $\mathbb{P}_{\mathcal{D}}[x_i] = \frac{8\epsilon}{d-1}$. Without loss of generality, $\mathcal{A}$ makes no error on x_0. For a sample S, let $\overline{S}$ denote the set of its elements falling in $\{x_1, ..., x_{d-1}\}$ and let $\mathcal{S}$ denote samples S of size m such that $|\overline{S}| \leq \frac{d-1}{2}$. Fix $S \in \mathcal{S}$ and consider the uniform distribution $\mathcal{U}$ over all labelings $f : \overline{\mathcal{X}} \to \{0, 1\}$ (which are all in $\mathcal{H}$ since the set is shattered). Then,

$$E_{f\sim\mathcal{U}}[R_{\mathcal{D}}(h_S, f)] = \sum_f \sum_{x \in \overline{\mathcal{X}}} 1_{h_S(x) \neq f(x)} \mathbb{P}[x]\mathbb{P}[f] \geq \sum_f \sum_{x \notin \overline{S}} 1_{h_S(x) \neq f(x)} \mathbb{P}[x]\mathbb{P}[f]$$

$$= \frac{1}{2} \sum_{x \notin \overline{S}} \mathbb{P}[x] \geq \frac{1}{2} \frac{d-1}{2} \frac{8\epsilon}{d-1} = 2\epsilon \Rightarrow E_{f\sim\mathcal{U}}[E_{S\in\mathcal{S}}[R_{\mathcal{D}}(h_S, f)]] \geq 2\epsilon$$

Hence $E_{S\in\mathcal{S}}[R_{\mathcal{D}}(h_S, f_0)] \geq 2\epsilon$ for at least one labeling $f_0 \in \mathcal{H}$. Since $R_{\mathcal{D}}(h_S, f_0) \leq \mathbb{P}_{\mathcal{D}}[\overline{\mathcal{X}} - \{x_0\}]$, we have that

$$E_{S\in\mathcal{S}}[R_{\mathcal{D}}(h_S, f_0)] = \sum_{S:R_{\mathcal{D}}(h_S, f_0)\geq\epsilon} R_{\mathcal{D}}(h_S, f_0)\mathbb{P}[R_{\mathcal{D}}(h_S, f_0)] + \sum_{S:R_{\mathcal{D}}(h_S, f_0)<\epsilon} R_{\mathcal{D}}(h_S, f_0)\mathbb{P}[R_{\mathcal{D}}(h_S, f_0)]$$

$$\leq \mathbb{P}_{\mathcal{D}}[\overline{\mathcal{X}} - \{x_0\}]\mathbb{P}_{S\in\mathcal{S}}[R_{\mathcal{D}}(h_S, f_0) \geq \epsilon] + \epsilon(1 - \mathbb{P}_{S\in\mathcal{S}}[R_{\mathcal{D}}(h_S, f_0) \geq \epsilon])$$

$$\leq 7\epsilon\mathbb{P}_{S\in\mathcal{S}}[R_{\mathcal{D}}(h_S, f_0) \geq \epsilon] + \epsilon \Rightarrow \frac{\mathbb{P}[\mathcal{S}]}{7} \leq \frac{1}{7} \leq \mathbb{P}_{S\in\mathcal{S}}[R_{\mathcal{D}}(h_S, f_0) \geq \epsilon]$$

Then, for a set $S = (x_1, ..., x_m)$ of size m, define $S_m = \sum_{i=1}^m 1_{x_i \in \overline{\mathcal{X}}}$. Since each $1_{x_i \in \overline{\mathcal{X}}}$ has an expected value of 8ϵ, the mean is $8\epsilon m$ in this case. Then, for any $\gamma > 0$, we use Theorem D.4 as

$$\mathbb{P}[S_m \geq 8\epsilon m(1 + \gamma)] \leq e^{-8\epsilon m \frac{\gamma^2}{3}}$$

hence

$$\epsilon = \frac{(d-1)}{32m},\ \gamma = 1 \Rightarrow 1 - \mathbb{P}[\mathcal{S}] = \mathbb{P}[S_m \geq \frac{d-1}{2}] \leq e^{-\frac{d-1}{12}} \leq e^{-\frac{1}{12}} \leq 1 - 7\delta$$

for $\delta \leq \frac{1}{100} \leq \frac{1-e^{-\frac{1}{12}}}{7}$. Then, $1 - \mathbb{P}[\mathcal{S}] \leq 1 - 7\delta$ so

$$7\delta \leq \mathbb{P}[\mathcal{S}] \Rightarrow \delta \leq \frac{\mathbb{P}[\mathcal{S}]}{7} \leq \mathbb{P}_{S\in\mathcal{S}}[R_{\mathcal{D}}(h_S, f_0) \geq \epsilon]$$

Note: Since there exists a distribution over $\mathcal{X}$ for which the error of the hypothesis returned by $\mathcal{A}$ (with respect to a target function f) is bounded by $C \cdot \frac{d}{m}$, infinite VC-dimension indicates that PAC-learning in the realizable case is not possible.

Slud's Inequality Let X be a random variable following the binomial distribution $B(m, p)$ and let k be an integer such that $p \leq \frac{1}{4}$ and $k \geq mp$ or $p \leq \frac{1}{2}$ and $mp \leq k \leq m(1-p)$. Then,

$$\mathbb{P}[X \geq k] \geq \mathbb{P}\left[N \geq \frac{k - mp}{\sqrt{mp(1-p)}}\right]$$

where N is in standard normal form.

Normal distribution tails: Lower bound: If N is a random variable following the standard normal distribution, then for $u > 0$ we have

$$\mathbb{P}[N \geq u] \geq \frac{1}{2}\Big(1 - \sqrt{1 - e^{-u^2}}\Big)$$

Exercise D.3: Let x_A and x_B be random variables (coins), with $\mathbb{P}[x_A = 0] = \frac{1}{2} - \frac{\epsilon}{2}$ and $\mathbb{P}[x_B = 0] = \frac{1}{2} + \frac{\epsilon}{2}$, where $0 < \epsilon < 1$ is a small positive number, 0 denotes heads and 1 denotes tails. Consider selecting a coin $x \in \{x_A, x_B\}$ uniformly at random, tossing it m times, and predicting which coin was tossed based on the sequence of 0s and 1s obtained.

a) Let S be a sample of size m. Consider playing the above game according to the decision rule $f_o : \{0,1\}^m \to \{x_A, x_B\}$ defined by $f_o(S) = x_A$ if and only if $N(S) < \frac{m}{2}$, where $N(S)$ is the number of 0's in sample S. Suppose m is even. Then, this rule fails in the case that $x = x_A$ yet at least half of the flips were heads. Hence,

$$\begin{aligned}\text{error}(f_0) &= E_x[\mathbb{P}_{\mathcal{D}_x^m}[f_o(S) \neq x]] \\ &= \mathbb{P}[x = x_A]\mathbb{P}_{\mathcal{D}_{x_A}^m}[f_o(S) \neq x_A] + \mathbb{P}[x = x_B]\mathbb{P}_{\mathcal{D}_{x_B}^m}[f_o(S) \neq x_B] \\ &\geq \frac{1}{2}\mathbb{P}\left[N(S) \geq \frac{m}{2} \mid x = x_A\right]\end{aligned}$$

b) Again assuming m is even, we find that $N(S)$ follows the binomial distribution $B(m,p)$ for $p = \frac{1}{2} - \frac{\epsilon}{2}$, where $m(\frac{1}{2} - \frac{\epsilon}{2}) \leq \frac{m}{2} \leq m(\frac{1}{2} + \frac{\epsilon}{2})$. Hence, Slud's Inequality implies

$$\mathbb{P}[N(S) \geq \frac{m}{2}] \geq \mathbb{P}\left[N \geq \frac{\frac{m}{2} - m(\frac{1}{2} - \frac{\epsilon}{2})}{\sqrt{m(\frac{1}{2} - \frac{\epsilon}{2})(\frac{1}{2} + \frac{\epsilon}{2})}}\right] = \mathbb{P}\Big[N \geq \frac{\epsilon\sqrt{m}}{\sqrt{1-\epsilon^2}}\Big]$$

to which we can apply the lower bound for normal distribution tails as

$$\mathbb{P}\Big[N \geq \frac{\epsilon\sqrt{m}}{\sqrt{1-\epsilon^2}}\Big] \geq \frac{1}{2}\Big(1 - \sqrt{1 - e^{-\frac{m\epsilon^2}{1-\epsilon^2}}}\Big)$$

hence

$$\text{error}(f_o) \geq \frac{1}{4}\Big(1 - \sqrt{1 - e^{-\frac{m\epsilon^2}{1-\epsilon^2}}}\Big)$$

c) If m is odd, then note that f_o fails in the case that $N(S) \geq \frac{m}{2} \iff N(S) \geq \lceil \frac{m}{2} \rceil$. Hence, $N(S)$ effectively follows a binomial distribution (by adding an arbitrary element to S) $B(m+1, p)$ for $p = \frac{1}{2} - \frac{\epsilon}{2}$, where $(m+1)(\frac{1}{2} - \frac{\epsilon}{2}) \leq \lceil \frac{m}{2} \rceil \leq (m+1)(\frac{1}{2} + \frac{\epsilon}{2})$. Using Slud's Inequality and the lower bound for normal distribution with $p = \frac{1}{2} - \frac{\epsilon}{2}$, we have

$$\begin{aligned}\frac{1}{2}\mathbb{P}\Big[N(S) \geq \frac{m}{2}\Big] &\geq \frac{1}{2}\mathbb{P}\left[N \geq \frac{\frac{m+1}{2} - (m+1)p}{\sqrt{(m+1)p(1-p)}}\right] = \frac{1}{2}\mathbb{P}\left[N \geq \frac{\epsilon\sqrt{m+1}}{\sqrt{1-\epsilon^2}}\right] \\ &\geq \frac{1}{4}\Big(1 - \sqrt{1 - e^{-\frac{\epsilon^2(m+1)}{1-\epsilon^2}}}\Big) = \frac{1}{4}\Big(1 - \sqrt{1 - e^{-\frac{2\lceil \frac{m}{2} \rceil \epsilon^2}{1-\epsilon^2}}}\Big)\end{aligned}$$

Since the rightmost expression holds as the same bound in the even case, both m odd and even share this bound.

d) If the error of f_o is to be at most δ, where $0 < \delta < \frac{1}{4}$, then

$$\delta \geq \frac{1}{4}\Big(1 - \sqrt{1 - e^{-\frac{2\lceil \frac{m}{2} \rceil \epsilon^2}{1-\epsilon^2}}}\Big) \Rightarrow (1-4\delta)^2 \leq 1 - e^{-\frac{2\lceil \frac{m}{2} \rceil \epsilon^2}{1-\epsilon^2}}$$

$$\Rightarrow -\frac{2\lceil \frac{m}{2} \rceil \epsilon^2}{1-\epsilon^2} \leq \log\Big(1-(1-4\delta)^2\Big) \Rightarrow -\frac{1-\epsilon^2}{2\epsilon^2}\log\Big(1-(1-4\delta)^2\Big) \leq \left\lceil \frac{m}{2} \right\rceil \leq \frac{m+1}{2}$$

$$\Rightarrow m \geq \frac{1-\epsilon^2}{\epsilon^2}\log\Big(\frac{1}{1-(1-4\delta)^2}\Big) - 1$$

Note that $\epsilon \to 0 \Rightarrow m \to \infty$

e) Now consider an arbitrary decision rule $f : \{0,1\}^m \to \{x_A, x_B\}$. Note that, if $f(S') = x_A$ on a particular outcome S' with $N(S) \geq \frac{m}{2}$ then the error of f on S' is at least $\frac{1}{2}\mathbb{P}\Big[N(S) < \frac{m}{2} \mid x = x_A\Big] \geq \frac{1}{2}\mathbb{P}\Big[N(S) \geq \frac{m}{2} \mid x = x_A\Big]$. Similarly, if $f(S') = x_A$ on an outcome S' with $N(S) < \frac{m}{2} - 1$, f errors on S' with at least $\frac{1}{2}\mathbb{P}\Big[N(S) \geq \frac{m}{2} - 1 \mid x = x_A\Big] \geq \frac{1}{2}\mathbb{P}\Big[N(S) \geq \frac{m}{2} \mid x = x_A\Big]$, hence

$$\text{error}(f) \geq \frac{1}{2}\mathbb{P}\Big[N(S) \geq \frac{m}{2} \mid x = x_A\Big]$$

so that the lower bound in part d applies to all decision rules.

Lemma 3.21: Let α be a uniformly distributed random variable taking values in $\{\alpha_-, \alpha_+\}$, where $\alpha_- = \frac{1}{2} - \frac{\epsilon}{2}$ and $\alpha_+ = \frac{1}{2} + \frac{\epsilon}{2}$. Let S be a sample of $m \geq 1$ random variables $X_1, ..., X_m$ taking values in $\{0,1\}$ and drawn i.i.d. according to the distribution $\mathcal{D}_\alpha$ defined by $\mathbb{P}_{\mathcal{D}_\alpha}[X = 1] = \alpha$. Then, if $h : \mathcal{X}^m \to \{\alpha_-, \alpha_+\}$, we have

$$E_\alpha[\mathbb{P}_{\mathcal{D}_\alpha^m}[h(S) \neq \alpha]] \geq \Phi\left(2\left\lceil \frac{m}{2} \right\rceil, \epsilon\right)$$

for $\Phi(m, \epsilon) = \frac{1}{4}\Big(1 - \sqrt{1 - e^{-\frac{m\epsilon^2}{1-\epsilon^2}}}\Big)$ for all m and ϵ.

Proof: This follows from the previous exercise.

Lemma 3.22: Let Z be a random variable taking values in $[0,1]$. Then, for any $\gamma \in [0,1)$, we have

$$\mathbb{P}[Z > \gamma] \geq \frac{E[Z] - \gamma}{1-\gamma} > E[Z] - \gamma$$

Proof: We find that

$$E[Z] \leq (1)(\mathbb{P}[Z > \gamma]) + (\gamma)(\mathbb{P}[Z \leq \gamma])$$

$$= \mathbb{P}[Z > \gamma] + (\gamma)(1 - \mathbb{P}[Z > \gamma]) \Rightarrow E[Z] - \gamma \leq \mathbb{P}[Z > \gamma](1-\gamma)$$

Theorem 3.23 (Lower bound, non-realizable case): let $\mathcal{H}$ be a hypothesis set with VC-dimension $d > 1$. Then, for any $m \geq 1$ and any learning algorithm $\mathcal{A}$, there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0, 1\}$ such that

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left[R_{\mathcal{D}}(h_S) - \inf_{h\in\mathcal{H}} R_{\mathcal{D}}(h) > \sqrt{\frac{d}{320m}}\right] \geq \frac{1}{64}$$

or equivalently, for any learning algorithm, the sample complexity verifies

$$m \geq \frac{d}{320\epsilon^2}$$

Proof: Let $\overline{\mathcal{X}} = \{x_1, ..., x_d\} \subset \mathcal{X}$ be a set shattered by $\mathcal{H}$. For any $\alpha \in [0, 1]$ and any vector $\sigma = (\sigma_1, ..., \sigma_d)^T \in \{-1, 1\}^d$, we define a distribution $\mathcal{D}_\sigma$ with support $\overline{\mathcal{X}} \times \{0, 1\}$ as follows: for any $i \in [d]$,

$$\mathbb{P}_{\mathcal{D}_\sigma}[(x_i, 1)] = \frac{1}{d}\Big(\frac{1}{2} + \frac{\sigma_i\alpha}{2}\Big)$$

For $i \in [d]$, we define the Bayes classifier as

$$h^*_{\mathcal{D}_\sigma}(x_i) = \mathrm{argmax}_{y\in\{0,1\}}\mathbb{P}[y \mid x_i]$$

Note that $h^*_{\mathcal{D}_\sigma}$ is in $\mathcal{H}$ since $\overline{\mathcal{X}}$ is shattered. Further, for all $h \in \mathcal{H}$,

$$R_{\mathcal{D}_\sigma}(h) - R_{\mathcal{D}_\sigma}(h^*_{\mathcal{D}_\sigma}) = E_{\mathcal{D}_\sigma}\Big[\frac{1}{d}\sum_{x\in\overline{\mathcal{X}}} 1_{h(x)\neq y}\Big] - E_{\mathcal{D}_\sigma}\Big[\frac{1}{d}\sum_{x\in\overline{\mathcal{X}}} 1_{h^*_{\mathcal{D}_\sigma}(x)\neq y}\Big]$$

$$= \frac{1}{d}\sum_{x\in\overline{\mathcal{X}}}\left(\Big(\frac{1}{2} + \frac{\alpha}{2}\Big) - \Big(\frac{1}{2} - \frac{\alpha}{2}\Big)\right)1_{h(x)\neq h^*_{\mathcal{D}_\sigma}(x)} = \frac{\alpha}{d}\sum_{x\in\overline{\mathcal{X}}} 1_{h(x)\neq h^*_{\mathcal{D}_\sigma}(x)}$$

Let h_S denote the hypothesis returned by the learning algorithm $\mathcal{A}$ after receiving the labeled sample S drawn according to $\mathcal{D}_\sigma$. Let $|S|_x$ denote the number of occurrences of a point x in S. Let $\mathcal{U}$ denote the uniform distribution over $\{-1, 1\}^d$. Then,

$$E_{\substack{\sigma\sim\mathcal{U}\\ S\sim\mathcal{D}^m_\sigma}}\Big[\frac{1}{\alpha}[R_{\mathcal{D}_\sigma}(h_S) - R_{\mathcal{D}_\sigma}(h^*_{\mathcal{D}_\sigma})]\Big] = \frac{1}{d}\sum_{x\in\overline{\mathcal{X}}} E_{\substack{\sigma\sim\mathcal{U}\\ S\sim\mathcal{D}^m_\sigma}}\Big[1_{h_S(x)\neq h^*_{\mathcal{D}_\sigma}(x)}\Big]$$

$$= \frac{1}{d}\sum_{x\in\overline{\mathcal{X}}} E_{\sigma\sim\mathcal{U}}\Big[\mathbb{P}_{S\sim\mathcal{D}^m_\sigma}[h_S(x) \neq h^*_{\mathcal{D}_\sigma}(x)]\Big]$$

$$= \frac{1}{d}\sum_{x\in\overline{\mathcal{X}}}\sum_{n=0}^{m} E_{\sigma\sim\mathcal{U}}\Big[\mathbb{P}_{S\sim\mathcal{D}^m_\sigma}[h_S(x) \neq h^*_{\mathcal{D}_\sigma}(x) \mid |S|_x = n]\Big]\mathbb{P}[|S|_x = n]$$

$$\geq \frac{1}{d}\sum_{x\in\overline{\mathcal{X}}}\sum_{n=0}^{m}\Phi(n+1, \alpha)\mathbb{P}[|S|_x = n] \geq \frac{1}{d}\sum_{x\in\overline{\mathcal{X}}}\Phi\Big(\frac{m}{d} + 1, \alpha\Big) = \Phi\Big(\frac{m}{d} + 1, \alpha\Big)$$

Hence there exists $\sigma \in \{-1, 1\}^d$ such that

$$E_{S\sim\mathcal{D}^m_\sigma}\Big[\frac{1}{\alpha}[R_{\mathcal{D}_\sigma}(h_S) - R_{\mathcal{D}_\sigma}(h^*_{\mathcal{D}_\sigma})]\Big] > \Phi\Big(\frac{m}{d} + 1, \alpha\Big)$$

By Lemma 3.22, for the same σ and any $\gamma \in [0, 1]$ we have

$$\mathbb{P}_{S\sim\mathcal{D}^m_\sigma}\Big[\frac{1}{\alpha}[R_{\mathcal{D}_\sigma}(h_S) - R_{\mathcal{D}_\sigma}(h^*_{\mathcal{D}_\sigma})] \geq \gamma u\Big] > (1-\gamma)u$$

for $u = \Phi\Big(\frac{m}{d}+1, \alpha\Big)$. If we bound $\delta \le (1-\gamma)u$ and $\epsilon \le \gamma\alpha u$, then

$$\mathbb{P}_{S\sim\mathcal{D}_\sigma^m}\Big[R_{\mathcal{D}_\sigma}(h_S) - R_{\mathcal{D}_\sigma}(h^*_{\mathcal{D}_\sigma}) > \epsilon\Big] > \delta$$

For $\gamma = 1 - 8\delta$, we have

$$\delta \le (1-\gamma)u \iff u \ge \frac{1}{8}$$
$$\iff \frac{1}{4}\Big(1 - \sqrt{1 - e^{-\frac{(\frac{m}{d}+1)\alpha^2}{1-\alpha^2}}}\Big) \ge \frac{1}{8} \iff \frac{1}{4} \ge 1 - e^{-\frac{(\frac{m}{d}+1)\alpha^2}{1-\alpha^2}}$$
$$\iff -\frac{(\frac{m}{d}+1)\alpha^2}{1-\alpha^2} \ge \log\frac{3}{4} \iff \frac{m}{d} \le \frac{1-\alpha^2}{\alpha^2}\log\frac{4}{3} - 1$$

Hence $\alpha = \frac{8\epsilon}{1-8\delta}$ gives $\epsilon = \frac{\gamma\alpha}{8}$ and

$$\frac{m}{d} \le \Big(\frac{(1-8\delta)^2}{64\epsilon^2} - 1\Big)\log\frac{4}{3} - 1 := f\Big(\frac{1}{\epsilon^2}\Big)$$

Then, to obtain a bound of the form $\frac{m}{d} \le \frac{\omega}{\epsilon^2}$, since $\epsilon \le \frac{1}{64}$, it suffices to set $\frac{\omega}{(\frac{1}{64})^2} = f\Big(\frac{1}{(\frac{1}{64})^2}\Big)$. Hence, for $\delta = \frac{1}{64}$, we have $\omega = \frac{1}{(64)^2}((7^2-1)\log\frac{4}{3} - 1) \approx \frac{1}{320}$ so that $\epsilon^2 \le \frac{1}{320(m/d)}$ suffices.

Ch. 3 Exercises.

3.1. Let $\mathcal{H}$ be the set of intervals in $\mathbb{R}$. The VC-dimension of $\mathcal{H}$ is 2, and its growth function satisfies $\Pi_{\mathcal{H}}(m) \le \sum_{i=0}^m (m-i+1) = m^2 + m - \sum_{i=0}^m$.

3.2. Let $\mathcal{H}$ be the family of threshold functions over the real line: $\mathcal{H} = \{x \mapsto 1_{x\le\theta} \mid \theta \in \mathbb{R}\} \cup \{x \mapsto 1_{x\ge\theta} \mid \theta \in \mathbb{R}\}$. In this case, given m points in $\mathbb{R}$, we can exclude or include all, as well as include from opposite sides of the real line. Hence, $\Pi_m(\mathcal{H}) \le 2 + (m-1)(2) = 2m$. Hence,

$$\mathfrak{R}_m(\mathcal{G}) \le \sqrt{\frac{2\log(2m)}{m}}$$

3.3. We define a linearly separable labeling of a set $\mathcal{X}$ of vectors in $\mathbb{R}^d$ as a classification of $\mathcal{X}$ into two sets $\mathcal{X}^+$ and $\mathcal{X}^-$ with $\mathcal{X}^+ = \{x \in \mathcal{X} \mid w \cdot x > 0\}$ and $\mathcal{X}^- = \{x \in \mathcal{X} \mid w \cdot x < 0\}$ for some $w \in \mathbb{R}^d$. Let $\mathcal{X} = \{x_1, ..., x_m\}$ be a subset of $\mathbb{R}^d$.

(a) Let $\{\mathcal{X}^+, \mathcal{X}^-\}$ be a dichotomy of $\mathcal{X}$ and let $x_{m+1} \in \mathbb{R}^d$. Suppose that $\{\mathcal{X}^+, \mathcal{X}^-\}$ is linearly separable by a hyperplane

$$w \cdot x = 0,\ w \in \mathbb{R}^d$$

passing through the origin and $x_{m+1} = (x^1_{m+1}, ..., x^d_{m+1})$. Then, since

$$\sum_{i=1}^d x^i_{m+1} w_i = 0$$

there exist $\epsilon_1, \epsilon_2 \in \mathbb{R}$ and $j, k \in \{1, ..., d\}$ for which $w' := (w_1, ..., w_j \pm \epsilon_1, ..., w_d)$ and $w'' := (w_1, ..., w_k \pm \epsilon_1, ..., w_d)$ satisfy

$$(w_j \pm \epsilon_1)x^j_{m+1} + \sum_{i\ne j} x^i_{m+1} w_i > 0$$

$$(w_k \pm \epsilon_2)x^j_{m+1} + \sum_{i \neq k} x^i_{m+1} w_i < 0$$

and $w \cdot x = 0$ still separates $\{\mathcal{X}^+, \mathcal{X}^-\}$.

Conversely, if $\{\mathcal{X}^+, \mathcal{X}^- \cup \{x_{m+1}\}\}$ and $\{\mathcal{X}^+ \cup \{x_{m+1}\}, \mathcal{X}^-\}$ are linearly separable by hyperplanes, those hyperplanes separate $\{\mathcal{X}^+, \mathcal{X}^-\}$.

b) Let $\mathcal{X} = \{x_1, ..., x_m\}$ be a subset of $\mathbb{R}^d$ such that any k-element subset of $\mathcal{X}$ with $k \leq d$ is linearly independent. Let $C(m, d)$ denote the number of linearly separable labelings of $\mathcal{X}$. Then, we find that $C(m+1, d)$ counts the linearly separable labelings in the m case for $\mathbb{R}^d$, and also double counts those cases in which the hyperplane (given by a vector $w \in \mathbb{R}^d$) can intersect the $m+1$-th vector. In such cases, the $m+1$-th vector may belong to either $\mathcal{X}^+$ or $\mathcal{X}^-$ by part (a), thereby defining two linearly separable labelings. Hence, $C(m+1, d) = C(m, d) + C(m, d-1)$. For $m = 1$, we have $1 = C(2, 1) = C(1, 1) + C(1, 0) = 1 + 0$. We may now inductively assume

$$C(m, d) = 2\sum_{k=0}^{d-1} \binom{m-1}{k}, \quad C(m, d-1) = 2\sum_{k=0}^{d-2} \binom{m-1}{k}$$

Then,

$$\begin{aligned} C(m+1, d) &= 2\sum_{k=0}^{d-1} \binom{m-1}{k} + 2\sum_{k=0}^{d-2} \binom{m-1}{k} \\ &= 2\sum_{k=0}^{d-1} \binom{m-1}{k} + 2\sum_{k=0}^{d-1} \binom{m-1}{k-1} = 2\sum_{k=0}^{d-1} \binom{m}{k} \end{aligned}$$

c) Let $f_1, ..., f_p$ be p functions mapping $\mathbb{R}^d$ to $\mathbb{R}$. Define $\mathcal{F}$ as the family of classifiers based on linear combinations of the functions:

$$\mathcal{F} = \left\{ x \mapsto \operatorname{sgn}\Big(\sum_{k=1}^{p} a_k f_k(x) \Big) : a_1, ..., a_p \in \mathbb{R} \right\}$$

Define Ψ by $\Psi(x) = (f_1(x), ..., f_p(x))$. Assume that there exists $x_1, ..., x_m \in \mathbb{R}^d$ such that every p-subset of $\{\Psi(x_1), ..., \Psi(x_m)\}$ is linearly independent. In this case,

$$\Pi_{\mathcal{F}}(m) = \sup_{\{x_1, ..., x_m\} \subset \mathbb{R}^d} |\{g(x_1), ..., g(x_m) \ : \ g \in \mathcal{F}\}|$$

so since each set $\{g(x_1), ..., g(x_m)\}$ represents a linearly separable labeling of the p-dimensional points $\{\Psi(x_1), ..., \Psi(x_m)\}$,

$$\sup_{\{x_1, ..., x_m\} \subset \mathbb{R}^d} |\{g(x_1), ..., g(x_m) \ : \ g \in \mathcal{F}\}| = 2\sum_{i=0}^{p-1} \binom{m-1}{i}$$

using part (b) and . Therefore,

$$\Pi_{\mathcal{F}}(m) = 2\sum_{i=0}^{p-1} \binom{m-1}{i}$$

3.11. For an input space $\mathcal{X} := \mathbb{R}^{n_1}$, we consider the family of regularized neural networks defined by the following set of functions mapping $\mathcal{X}$ to $\mathbb{R}$:

$$\mathcal{H} = \left\{ x \mapsto \sum_{j=1}^{n_2} w_j \sigma(u_j \cdot x) \ : \ ||w||_1 \leq \Lambda', \ ||u_j||_2 \leq \Lambda, \text{ for any } j \in [n_2] \right\}$$

where σ is an L-Lipschitz function (e.g. σ could be the sigmoid function which is 1-Lipschitz).

a) We find that

$$\widehat{\mathfrak{R}}_S(\mathcal{H}) = E_\sigma\Big[\sup_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^m \sigma_i h(x_i)\Big] = E_\sigma\Big[\sup_{w,u_j} \frac{1}{m}\sum_{i=1}^m \sigma_i \sum_{j=1}^{n_2} w_j \sigma(u_j \cdot x_i)\Big]$$

$$= \frac{1}{m}E_\sigma\Big[\sup_w \sum_{j=1}^{n_2} w_j \sup_{||u||_2\leq\Lambda} \sum_{i=1}^m \sigma_i \sigma(u\cdot x_i)\Big] = \frac{\Lambda'}{m}E_\sigma\Big[\sup_{||u||_2\leq\Lambda} \sum_{i=1}^m \sigma_i\sigma(u\cdot x_i)\Big]$$

b) We now use the following form of Talagrand's lemma valid for all hypothesis sets $\mathcal{H}$ and L-lipschitz functions Φ:

$$\frac{1}{m}E_\sigma\left[\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^m \sigma_i(\Phi\circ h)(x_i)\right|\right] \leq \frac{L}{m}E_\sigma\left[\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^m \sigma_i h(x_i)\right|\right]$$

so that

$$\widehat{\mathfrak{R}}_S(\mathcal{H}) \leq \frac{\Lambda' L}{m}E_\sigma\Big[\sup_{||u||_2\leq\Lambda}\sum_{i=1}^m \sigma_i(u\cdot x_i)\Big] \leq \Lambda' L E_\sigma\left[\sup_{h\in\mathcal{H}'}\frac{1}{m}\sum_{i=1}^m \sigma_i h(x_i)\right]$$

$$= \Lambda' L \widehat{\mathfrak{R}}_S(\mathcal{H}')$$

c) We then find that

$$\widehat{\mathfrak{R}}_S(\mathcal{H}') = E_\sigma\Big[\sup_{s,u}\frac{1}{m}\sum_{i=1}^m \sigma_i s(u\cdot x_i)\Big] = E_\sigma\left[\frac{1}{m}\Big|\Big|u\Big|\Big|_2\left|\left|\sum_{i=1}^m \sigma_i x_i\right|\right|_2\right]$$

$$= \frac{\Lambda}{m}E_\sigma\left[\left|\left|\sum_{i=1}^m \sigma_i x_i\right|\right|_2\right]$$

d) By Jensen's inequality, we have

$$E_v[||v||_2] \leq \sqrt{E_v[||v||_2^2]}$$

hence

$$\widehat{\mathfrak{R}}_S(\mathcal{H}') \leq \frac{\Lambda}{m}\sqrt{E_\sigma\left[\left|\left|\sum_{i=1}^m \sigma_i x_i\right|\right|_2^2\right]}$$

e) If for any $x \in S$ we have $||x||_2 \leq r$ for some $r > 0$, then

$$\widehat{\mathfrak{R}}_S(\mathcal{H}) \leq \Lambda' L\left(\frac{\Lambda}{m}\sqrt{\Big(\sum_{i=1}^{m}||\sigma_i x_i||_2\Big)^2}\right) \leq \Lambda' L\Big(\frac{\Lambda}{m}(mr)\Big) = \Lambda'\Lambda L r$$

3.27. Let $\mathcal{C}$ be a concept class over $\mathbb{R}^r$ with VC-dimension d. A $\mathcal{C}$-neural network with one intermediate layer is a concept defined over $\mathbb{R}^n$ that can be represented by a direct acyclic graph in which the input nodes are those at the bottom and in which each other node is labeled with a concept $c \in \mathcal{C}$.

The output of the neural network for a given input vector $(x_1, ..., x_n)$ is obtained as follows. First, each of the n input nodes is labeled with the corresponding value $x_i \in \mathbb{R}$. Next, the value at a node u in the higher layer (labeled with c) is obtained by applying c to the values of the input nodes admitting an edge ending in u. Since $c \in \{0, 1\}$, $u \in \{0, 1\}$. The value at the top (output) node is obtained similarly by applying the corresponding concept to the values of the nodes admitting an edge to the output node.

a) Let $\mathcal{H}$ denote the set of all neural networks defined with $k \geq 2$ internal nodes. Let $\Pi_{\mathcal{C}}(m) = \max_{z_1,...,z_m \subset \mathbb{R}^r} |\{(c(z_1), ..., c(z_m)) : c \in \mathcal{C}\}|$ denote the growth function of the concept class $\mathcal{C}$. We then have $\Pi_{\mathcal{H}}(m) \leq \Big(\Pi_c(m)\Big)^{k+1}$ if there are k intermediate nodes and 1 final node.

b) Since $\Pi_{\mathcal{H}}(m) \leq \Pi_{\mathcal{C}}(m)^{k+1}$, by Sauer's Lemma we have

$$\Pi_{\mathcal{C}}(m) \leq \Big(\frac{em}{d}\Big)^d \Rightarrow \Pi_{\mathcal{H}}(m) \leq \Big(\frac{em}{d}\Big)^{d(k+1)}$$

so that

$$m := 2(k+1)d\log_2(ek+e) \Rightarrow m > d(k+1)\log_2\Big(\frac{em}{d}\Big)$$

hence

$$2^m > \Big(\frac{em}{d}\Big)^{d(k+1)}$$

so since we must have

$$2^{m^*} \leq \Big(\frac{em^*}{d}\Big)^{d(k+1)}$$

for the VC-dimension m^*, we have that

$$\text{VCdim}(\mathcal{H}) \leq 2(k+1)d\log_2(ek+e)$$

c) Let $\mathcal{C}$ be the family of concept classes defined by threshold functions $\mathcal{C} = \Big\{\text{sgn}\Big(\sum_{j=1}^{r} w_j x_j\Big) : w \in \mathbb{R}^r\Big\}$. In this case, $\text{VCdim}(\mathcal{C}) = r$ since the r-dimensional vectors with 1's in the i-th spot may be shattered but not the origin x_0 (since $\mathcal{C}$ does not involve a term added to the dot product. Hence,

$$\text{VCdim}(\mathcal{H}) \leq 2(k+1)r\log_2(ek+e)$$

3.31. Let $\mathcal{H}$ be a family of functions mapping $\mathcal{X}$ to a subset of real numbers $\mathcal{Y} \subset \mathbb{R}$. For any $\epsilon > 0$, the "covering number" $\mathcal{N}(\mathcal{H}, \epsilon)$ of $\mathcal{H}$ for the L_∞ norm is the minimal $k \in \mathbb{N}$ such that $\mathcal{H}$ can be covered with k balls of radius ϵ, i.e. there exists $\{h_1, ..., h_k\} \subset H$ such that for all $h \in \mathcal{H}$ there exists $i \leq k$ with $||h - h_i||_\infty = \max_{x \in mcX} |h(x) - h_i(x)| \leq \epsilon$. Hence, when $\mathcal{H}$ is compact, the finite subcover due to an ϵ covering of $\mathcal{H}$ indicates that $\mathcal{N}(\mathcal{H}, \epsilon)$ is finite.

Let $\mathcal{D}$ denote a distribution of $\mathcal{X} \times \mathcal{Y}$ according to which labeled examples are drawn. Then, for $h \in \mathcal{H}$, $R(h) = E_{(x,y)\sim\mathcal{D}}[(h(x) - y)^2]$ and $\widehat{R}_S(h) = \frac{1}{m}\sum_{i=1}^m (h(x_i) - y_i)^2$ for a lebeled sample $S = ((x_1, y_1), ..., (x_m, y_m))$. Suppose $\mathcal{H}$ is bounded and that there exists $M > 0$ such that $|h(x) - y| \leq M$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$.

a) Let $L_S(h) = R(h) - \widehat{R}_S(h)$. Then, we find that

$$|L_S(h_1) - L_S(h_2)| = \Big|E[(h_1(x)-y)^2 - (h_2(x)-y)^2] + \frac{1}{m}\sum_{i=1}^m (h_2(x_i)-y_i)^2 - (h_1(x_i)-y_i)^2\Big|$$

$$= \Big|E[h_1(x)^2 - 2h_1(x)y - (h_2(x)^2 - 2h_2(x)y)] + \frac{1}{m}\sum_{i=1}^m h_1(x_i)^2 - 2h_1(x_i)y_i - (h_2(x_i)^2 - 2h_2(x_i)y_i)\Big|$$

$$= \Big|E[(h_1(x) - h_2(x))(h_1(x) - y) - (h_2(x) - h_1(x))(h_2(x) - y)] +$$
$$\frac{1}{m}\sum_{i=1}^m (h_1(x_i) - h_2(x_i))(h_1(x_i) - y_i) - (h_2(x_i) - h_1(x_i))(h_2(x_i) - y_i)\Big|$$

$$\leq |ME[h_1(x) - h_2(x)]| + |ME[h_2(x) - h_1(x)]| + \frac{1}{m}\sum_{i=1}^m 2M \max_i |h_1(x_i) - h_2(x_i)|$$
$$\leq 4M||h_1 - h_2||_\infty$$

b) Assume that $\mathcal{H}$ can be covered by k subsets $\mathcal{B}_1, ..., \mathcal{B}_k$, i.e. $\mathcal{H} = \mathcal{B}_1 \cup ... \cup \mathcal{B}_k$. Fix $\epsilon > 0$. We then have that

$$\mathbb{P}_{S\sim\mathcal{D}^m}\Big[\sup_{h\in\mathcal{H}} |L_S(h)| \geq \epsilon\Big] = \mathbb{P}_{S\sim\mathcal{D}^m}\Big[\sup_{h\in\mathcal{B}_1} |L_S(h)| \geq \epsilon \vee ... \vee \sup_{h\in\mathcal{B}_k} |L_S(h)| \geq \epsilon\Big]$$
$$\leq \sum_{i=1}^k \mathbb{P}_{S\sim\mathcal{D}^m}\Big[\sup_{h\in\mathcal{B}_i} |L_S(h)| \geq \epsilon\Big]$$

by the union bound.

c) We then let $k = \mathcal{N}(\mathcal{H}, \frac{\epsilon}{8M})$ and let $\mathcal{B}_1, ..., \mathcal{B}_k$ be balls of radius $\frac{\epsilon}{8M}$ centered at $h_1, ..., h_k$ covering $\mathcal{H}$. Fix $i \in [k]$. Note that if $h' := \text{argmax}_{h\in\mathcal{B}_i}|L_S(h)|$, then since

$$|L_S(h') - L_S(h_i)| \leq 4M||h' - h_i||_\infty \leq \frac{\epsilon}{2}$$

we have

$$|L_S(h')| \geq \epsilon \Rightarrow |L_S(h_i)| \geq \frac{\epsilon}{2}$$

hence

$$\mathbb{P}_{S\sim\mathcal{D}^m}\Big[\sup_{h\in\mathcal{B}_i} |L_S(h)| \geq \epsilon\Big] \leq \mathbb{P}_{S\sim\mathcal{D}^m}\Big[|L_S(h_i)| \geq \frac{\epsilon}{2}\Big]$$

so by Hoeffding's Inequality and part b),

$$\mathbb{P}_{S\sim\mathcal{D}^m}\Big[\sup_{h\in\mathcal{H}}|L_S(h)|\geq\epsilon\Big]\leq\sum_{i=1}^{k}\mathbb{P}_{S\sim\mathcal{D}^m}\Big[\sup_{h\in\mathcal{B}_i}|L_S(h)|\geq\epsilon\Big]$$

$$\leq\sum_{i=1}^{k}\mathbb{P}_{S\sim\mathcal{D}^m}\Big[|L_S(h_i)|\geq\frac{\epsilon}{2}\Big]=\sum_{i=1}^{k}\mathbb{P}_{S\sim\mathcal{D}^m}\Big[|R(h)-\widehat{R}_S(h)|\geq\frac{\epsilon}{2}\Big]$$

$$\leq 2ke^{-\frac{2(\frac{\epsilon}{2})^2}{\Sigma_{i=1}^m(\frac{M^2}{m})^2}}=2\mathcal{N}\Big(\mathcal{H},\frac{\epsilon}{8M}\Big)e^{-\frac{m\epsilon^2}{2M^2}}$$

Chapter 4 Notes

Definition: A standard algorithm to bound estimation error is Empirical Risk Minimization (ERM):

$$h_S^{\text{ERM}}=\text{argmin}_{h\in\mathcal{H}}\widehat{R}_S(h)$$

Proposition 4.1: For any sample S, the following inequality holds for the hypothesis returned by ERM:

$$\mathbb{P}\Big[R(h_S^{\text{ERM}})-\inf_{h\in\mathcal{H}}R(h)>\epsilon\Big]\leq\mathbb{P}\Big[\sup_{h\in\mathcal{H}}|R(h)-\widehat{R}_S(h)|>\frac{\epsilon}{2}\Big]$$

Proof: We find that

$$\epsilon<R(h_S^{\text{ERM}})-\inf_{h\in\mathcal{H}}R(h)\leq|R(h_S^{\text{ERM}})-\widehat{R}_S(h_S^{\text{ERM}})|+|\inf_{h\in\mathcal{H}}R(h)-\widehat{R}_S(h_S^{\text{ERM}})|$$

so at least one of the terms on the right hand side exceeds $\frac{\epsilon}{2}$, hence

$$\sup_{h\in\mathcal{H}}|R(h)-\widehat{R}_S(h)|>\frac{\epsilon}{2}$$

satisfying

$$\mathbb{P}\Big[R(h_S^{\text{ERM}})-\inf_{h\in\mathcal{H}}R(h)>\epsilon\Big]\leq\mathbb{P}\Big[\sup_{h\in\mathcal{H}}|R(h)-\widehat{R}_S(h)|>\frac{\epsilon}{2}\Big]$$

Definition: Regularization-based algorithms consist of selecting a family $\mathcal{H}$ that is an uncountable union of nested hypothesis sets $\mathcal{H}_\gamma$, i.e. $\mathcal{H}=\bigcup_{\gamma>0}\mathcal{H}_\gamma$, and $\mathcal{H}$ is often chosen to be dense in the space of continuous functions over $\mathcal{X}$. Often there exists $\mathcal{R}:\mathcal{H}\to\mathbb{R}$ such that, for any $\gamma>0$, the constrained optimization problem

$$\text{argmin}_{\gamma>0,h\in\mathcal{H}}\widehat{R}_S(h)+\text{pen}(\gamma,m)$$

where $\text{pen}(\gamma,m)$ refers to a penalty term such as $\mathfrak{R}_m(\mathcal{H}_\gamma)+\sqrt{\frac{\log\gamma}{m}}$, can be written as the unconstrained optimization problem

$$\text{argmin}_{h\in\mathcal{H}}\widehat{R}_S(h)+\lambda\mathcal{R}(h)$$

for some $\lambda>0$. Note that $\mathcal{R}(h)$ is a "regularization term— and λ is treated as a "regularization" hyperparameter (optimal value not known). Larger λ helps penalize more complex hypotheses while $\lambda\approx 0$ coincides with ERM. Cross-validation or n-fold cross-validation help select a value for λ.

Remark: Solving the ERM optimization problem is often NP-hard since the zero-one loss function is not convex, hence using a convex "surrogate" loss function can help upper bound the zero-one loss. In particular, for real-valued $h : \mathcal{X} \to \mathbb{R}$, we denote the binary classifier

$$f_h(x) = \begin{cases} 1 & h(x) \geq 0 \\ -1 & h(x) < 0 \end{cases}$$

and define the expected error $R(h)$ as

$$R(h) = E_{(x,y)\sim\mathcal{D}}[1_{f_h(x)\neq y}]$$

For any $x \in \mathcal{X}$ we write $\eta(x) := \mathbb{P}[y = 1|x]$. For $\mathcal{D}_{\mathcal{X}}$ the marginal distribution over $\mathcal{X}$ and any h, we then have

$$R(h) = E_{(x,y)\sim\mathcal{D}}[1_{f_h(x)\neq y}] = E_{x\sim\mathcal{D}_{\mathcal{X}}}\left[\eta(x)1_{h(x)<0} + (1-\eta(x))1_{h(x)\geq 0}\right]$$

We then define the "Bayes scoring function" $h^* : \mathcal{X} \to \mathbb{R}$ as

$$h^*(x) := \eta(x) - \frac{1}{2}$$

where

$$R^* := R(h^*)$$

denotes the error of the Bayes scoring function.

Lemma 4.5: The "excess error" of any hypothesis $h : \mathcal{X} \to \mathbb{R}$ can be expressed as

$$R(h) - R^* = 2E_{x\sim\mathcal{D}_{\mathcal{X}}}\left[|h^*(x)|1_{h(x)h^*(x)\leq 0}\right]$$

Proof: For any h we have

$$\begin{aligned} R(h) &= E_{x\sim\mathcal{D}_{\mathcal{X}}}[\eta(x)1_{h(x)<0} + (1-\eta(x))1_{h(x)\geq 0}] \\ &= E_{x\sim\mathcal{D}_{\mathcal{X}}}[\eta(x)1_{h(x)<0} + (1-\eta(x))(1-1_{h(x)<0})] \\ &= E_{x\sim\mathcal{D}_{\mathcal{X}}}[2\eta(x)1_{h(x)<0} + 1 - 1_{h(x)<0} - \eta(x)] \\ &= E_{x\sim\mathcal{D}_{\mathcal{X}}}[2h^*(x)1_{h(x)<0} + (1-\eta(x))] \end{aligned}$$

so that

$$\begin{aligned} R(h) - R^* &= 2E_{x\sim\mathcal{D}_{\mathcal{X}}}[h^*(x)1_{h(x)<0} - h^*(x)1_{h^*(x)<0}] \\ &= 2E_{x\sim\mathcal{D}_{\mathcal{X}}}[1_{h(x)h^*(x)\leq 0}|h^*(x)|] \end{aligned}$$

Definition: Let $\Phi : \mathbb{R} \to \mathbb{R}$ be a convex and non-decreasing function so that for any $u \in \mathbb{R}$, $1_{u\leq 0} \leq \Phi(-u)$. The "$\Phi$-loss" of a function $h : \mathcal{X} \to \mathbb{R}$ at a point $(x, y) \in \mathcal{X} \times \{-1, 1\}$ is defined as $\Phi(-yh(x))$ and its expected loss is given by

$$\begin{aligned} \mathcal{L}_\Phi(h) &:= E_{(x,y)\sim\mathcal{D}}[\Phi(-yh(x))] \\ &= E_{x\sim\mathcal{D}_{\mathcal{X}}}[\eta(x)\Phi(-h(x)) + (1-\eta(x))\Phi(h(x))] \end{aligned}$$

Note that $1_{u\leq 0} \leq \Phi(-u) \Rightarrow R(h) \leq \mathcal{L}_\Phi(h)$.

Definition: We further define $u \mapsto L_\Phi(x,u)$ for any $x \in \mathcal{X}$ and $u \in \mathbb{R}$ as

$$L_\Phi(x,u) = \eta(x)\Phi(-u) + (1-\eta(x))\Phi(u)$$

so that $\mathcal{L}_\Phi(h) = E_{x\sim\mathcal{D}_\mathcal{X}}[L_\Phi(x,h(x))]$ Note that since Φ is convex, so is $u \mapsto L_\Phi(x,u)$.

Definition: Let $h^*_\Phi : \mathcal{X} \to [-\infty,\infty]$ denote the "Bayes solution for the loss function L_Φ", i.e. $h^*_\Phi(x)$ solves the convex optimization problem:

$$h^*_\Phi(x) = \text{argmin}_{u\in[-\infty,\infty]} L_\Phi(x,u)$$

Note that this solution may not be unique. We lastly define

$$\mathcal{L}^*_\Phi := E_{(x,y)\sim\mathcal{D}}[\Phi(-yh^*_\Phi(x))]$$

Proposition 4.6: Let Φ be a convex non-decreasing function with $\Phi'(0) > 0$. Then, for any $x \in \mathcal{X}$, $h^*_\Phi(x) > 0 \iff h^*(x) > 0$ and $h^*(x) = 0 \iff h^*_\Phi(x) = 0$, hence $\mathcal{L}^*_\Phi = R^*$

Theorem 4.7: Let Φ be a convex and non-decreasing function. Assume that there exists $s \geq 1$ and $c > 0$ such that the following holds for all $x \in \mathcal{X}$:

$$|h^*(x)|^s = |\eta(x) - \frac{1}{2}|^s \leq c^s[L_\Phi(x,0) - L_\Phi(x,h^*_\Phi(x))]$$

Then, for any hypothesis h, the excess error of h satisfies

$$R(h) - R^* \leq 2c(\mathcal{L}_\Phi(h) - \mathcal{L}^*_\Phi)^{\frac{1}{s}}$$

Proof: First note that, for $\text{sgn}(h) \neq \text{sgn}(h^*)$

$$(*) \quad \eta(x)\Phi(0) + (1-\eta(x))\Phi(0) = \Phi(0) \leq \eta(x)(\Phi(-h(x))) + (1-\eta(x))\Phi(h(x))$$

as $h > 0$ for $\eta(x) < \frac{1}{2}$ and $h < 0$ for $\eta > \frac{1}{2}$, and Φ is non-decreasing with non-decreasing derivative.

We find that

$$\begin{aligned} R(h) - R^* &= 2E_{x\sim\mathcal{D}_\mathcal{X}}[|h^*(x)|1_{h(x)h^*(x)\leq 0}] \\ &\leq 2E_{x\sim\mathcal{D}_\mathcal{X}}[c(L_\Phi(x,0) - L_\Phi(x,h^*_\Phi(x)))^{\frac{1}{s}}1_{h(x)h^*(x)\leq 0}] \\ &= 2cE_{x\sim\mathcal{D}_\mathcal{X}}[((L_\Phi(x,0) - L_\Phi(x,h^*_\Phi(x)))1_{h(x)h^*(x)\leq 0})^{\frac{1}{s}}] \end{aligned}$$

and since $x \mapsto x^{\frac{1}{s}}$ is a concave function for $s \geq 1$,

$$\leq 2c(E_{x\sim\mathcal{D}_\mathcal{X}}[(L_\Phi(x,0) - L_\Phi(x,h^*_\Phi(x)))1_{h(x)h^*(x)\leq 0}])^{\frac{1}{s}}$$

By $(*)$ we then have

$$\leq 2c(E_{x\sim\mathcal{D}_\mathcal{X}}[(L_\Phi(x,h(x)) - L_\Phi(x,h^*_\Phi(x)))1_{h(x)h^*(x)\leq 0}])^{\frac{1}{s}}$$

so since since $L_\Phi(x,h(x)) \geq L_\Phi(x,h^*_\Phi(x))$ for any h,

$$\leq 2c(E_{x\sim\mathcal{D}_\mathcal{X}}[L_\Phi(x,h(x)) - L_\Phi(x,h^*_\Phi(x))])^{\frac{1}{s}} = 2c(\mathcal{L}_\Phi(h) - \mathcal{L}^*_\Phi)^{\frac{1}{s}}$$

Ch. 4 Exercises.

4.1. We find that, for any $h \in \mathcal{H}$, $\widehat{R}_S(h_S^{\mathrm{ERM}}) \leq \widehat{R}_S(h)$, hence $E_{S\sim\mathcal{D}^m}[\widehat{R}_S(h_S^{\mathrm{ERM}})] \leq \inf_{h\in\mathcal{H}} E_{S\sim\mathcal{D}^m}[\widehat{R}_S(h)]$. Further, $R(h_S^{\mathrm{ERM}}) \geq \inf_{h\in\mathcal{H}} R(h)$ for any $S \sim \mathcal{D}^m$, hence $\inf_{h\in\mathcal{H}} E_{S\sim\mathcal{D}^m}[\widehat{R}_S(h)] \leq E_{S\sim\mathcal{D}^m}[R(h_S^{\mathrm{ERM}})]$

4.2. Let $\Phi(u) = (1+u)^2$, so that Φ is non-decreasing on $[-1,\infty]$ and convex with $\Phi''(u) = 2 > 0$. We observe that

$$\eta(x)\Phi(-u) + (1-\eta(x))\Phi(u) = (1+u)^2 - 4\eta(x)u$$

so for $\eta = 0$,

$$|h^*(x)|^2 = \frac{1}{4} = (\frac{1}{2})^2(1 - \inf_u((1+u)^2))$$

For $\eta = \frac{1}{2}$ we have

$$|h^*(x)|^2 = 0 = \frac{1-\inf_u(1+u^2)}{4} = (\frac{1}{2})^2(1-\inf_u((1+u)^2 - 2u))$$

For $\eta = \frac{1}{2} + \epsilon$ with $\epsilon \in (0, \frac{1}{2}]$, since $\inf_u \frac{u^2-4u\epsilon}{4} \leq -\epsilon^2$,

$$|h^*(x)|^2 = \epsilon^2 = -\frac{4\epsilon^2 - 8\epsilon^2}{4} \leq -\inf_u \frac{u^2-4u\epsilon}{4} = \frac{1-\inf_u((1+u)^2 - 4u(\frac{1}{2}+\epsilon))}{4}$$

Similarly, for $\eta = \frac{1}{2} - \epsilon$ with $\epsilon \in (0, \frac{1}{2}]$, since $\inf_u \frac{u^2-4u\epsilon}{4} \leq -\epsilon^2$ (choosing $u = -2\epsilon$),

$$|h^*(x)|^2 = \epsilon^2 = -\frac{4\epsilon^2-8\epsilon^2}{4} \leq -\inf_u \frac{u^2+4u\epsilon}{4}$$

$$= \frac{1-\inf_u((1+u)^2 - 4u(\frac{1}{2}-\epsilon))}{4} = \frac{1}{4}(\Phi(0) - L_\Phi(x, h_\Phi^*(x))) = \frac{1}{4}(L_\Phi(x,0) - L_\Phi(x, h_\Phi^*(x)))$$

Hence, for $s = 2$ and $c = \frac{1}{2}$ we have

$$R(h) - R^* \leq [\mathcal{L}_\Phi(h) - \mathcal{L}_\Phi^*]^{\frac{1}{2}}$$

4.3. We then consider the Hinge loss $\Phi(u) = \max(0, 1+u)^2$. Since this function is the same as that in 4.2 on $[-1,\infty]$, the same bounds hold.

4.4. Define the loss of $h : \mathcal{X} \to \mathbb{R}$ at a point $(x,y) \in \mathcal{X} \times \{-1,1\}$ to be $1_{yh(x)\leq 0}$.

a) The Bayes classifier in this case is

$$h'(x) := \mathrm{argmin}_{y\in\{-1,1\}} \mathbb{P}[y|x]$$

hence a scoring function could be

$$h^*(x) := \begin{cases} \eta(x) - \frac{1}{2} & \eta(x) \neq \frac{1}{2} \\ -1 & \eta(x) = \frac{1}{2} \end{cases}$$

where $\eta(x) = \mathbb{P}[1|x]$.

b) In this case, replacing $1_{h(x)\leq 0}$ with $1_{h(x)<0} + 1_{h(x)=0}$ yields

$$R(h) = E_{x\sim\mathcal{D}_\mathcal{X}}[\eta(x)(1 - 1_{h(x)>0}) + (1-\eta(x))(1_{h(x)>0} + 1_{h(x)=0})]]$$

$$R(h) - R^* = E_{(x,y)\in\mathcal{D}}[1_{yh(x)\leq 0} - 1_{yh^*(x)\leq 0}]$$

$$= E_{x\sim\mathcal{D}_\mathcal{X}}[\eta(x)1_{h(x)\leq 0} + (1-\eta(x))1_{h(x)\geq 0} - (\eta(x)1_{h^*(x)\leq 0} + (1-\eta(x))1_{h^*(x)\geq 0})]$$

where replacing $1_{h(x)\leq 0}$ with $1_{h(x)<0} + 1_{h(x)=0}$ yields

$$= E_{x\sim\mathcal{D}_\mathcal{X}}[2|h^*(x)|1_{h(x)*h^*(x)\le 0} + (-h^*(x) + \frac{1}{2})(1_{h(x)=0} - 1_{h^*(x)=0})]$$

Chapter 15 Notes

Definition: A projection on a vector space V is a linear operator $P : V \to V$ such that $P^2 = P$. A projection on a Hilbert space V is an orthogonal projection if $\langle Px, y\rangle = \langle x, Py\rangle$

Definition: The "Frobenius norm", denoted by $||.||_F$ is a matrix norm defined over $\mathbb{R}^{m\times n}$ as

$$||\mathbf{M}||_F := \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}\mathbf{M}_{ij}^2}$$

Definition: For a sample $S = (x_1, ..., x_m)$ and feature mapping $\mathbf{\Phi} : \mathcal{X} \to \mathbb{R}^N$, we define the data matrix $(\mathbf{\Phi}(x_1), ..., \mathbf{\Phi}(x_m)) =: \mathbf{X} \in \mathbb{R}^{N\times m}$. If $\mathbf{X}$ is a mean-centered data matrix ($\sum_{i=1}^m \mathbf{\Phi}(x_i) = \mathbf{0}$), let $\mathcal{P}_k$ denote the set of N-dimensional rank$-k$ orthogonal projection matrices. PCA (Principal Component Analysis) is defined by the orthogonal projection matrix

$$\mathbf{P}^* := \text{argmin}_{\mathbf{P}\in\mathcal{P}_k}||\mathbf{PX} - \mathbf{X}||_F^2$$

Definition: The "top singular vector" of a matrix $\mathbf{M}$ is the vector $\mathbf{x}$ which maximizes the Rayleigh quotient

$$r(\mathbf{x}, \mathbf{M}) = \frac{\mathbf{x}^T\mathbf{M}\mathbf{x}}{\mathbf{x}^T\mathbf{x}}$$

Theorem 15.1: Let $\mathbf{P}^* \in \mathcal{P}_k$ be the PCA solution for a centered data matrix $\mathbf{X}$. Then, $\mathbf{P}^* = \mathbf{U}_k\mathbf{U}_k^T$, where $\mathbf{U}_k \in \mathbb{R}^{N\times k}$ is the matrix formed by the top k singular vectors of $\mathbf{C} := \frac{1}{m}\mathbf{X}\mathbf{X}^T$, the sample covariance matrix corresponding to $\mathbf{X}$. Note that this is the sample covariance matrix since

$$\frac{1}{m}(\mathbf{X}\mathbf{X}^T)_{ij} = \frac{1}{m}\sum_{\ell=1}^{m}\mathbf{X}_{i\ell}\mathbf{X}_{\ell j}^T = \frac{1}{m}\sum_{\ell=1}^{m}\mathbf{\Phi}(x_\ell)_i\mathbf{\Phi}(x_\ell)_j$$

$$= E[\mathbf{\Phi}(x)_i\mathbf{\Phi}(x)_j] = E[\mathbf{\Phi}(x)_i\mathbf{\Phi}(x)_j] - E[\mathbf{\Phi}(x)_i]E[\mathbf{\Phi}(x)_j] = \text{Cov}(\mathbf{\Phi}(x)_i, \mathbf{\Phi}(x)_j)$$

where the right hand term is the covariance between i-th and j-th coordinates of the feature output based on m samples. Moreover, the associated k-dimensional representation of $\mathbf{X}$ is given by $\mathbf{Y} = \mathbf{U}_k^T\mathbf{X}$.

Proof: For $\mathbf{P} = \mathbf{P}^T$ an orthogonal projection matrix, we seek to minimize

$$||\mathbf{PX} - \mathbf{X}||_F^2 = \sum_{i=1}^{N}\sum_{j=1}^{N}((\mathbf{PX} - \mathbf{X})_{ij})^2 = \text{Tr}[(\mathbf{PX} - \mathbf{X})^T(\mathbf{PX} - \mathbf{X})]$$

$$= \text{Tr}[\mathbf{X}^T\mathbf{P}^2\mathbf{X} - \mathbf{X}^T\mathbf{P}^T\mathbf{X} - \mathbf{X}^T\mathbf{P}\mathbf{X} + \mathbf{X}^T\mathbf{X}] = \text{Tr}[\mathbf{X}^T\mathbf{P}\mathbf{X} - 2\mathbf{X}^T\mathbf{P}\mathbf{X} + \mathbf{X}^T\mathbf{X}]$$

$$= \text{Tr}[\mathbf{X}^2] - \text{Tr}[\mathbf{X}^T\mathbf{P}\mathbf{X}]$$

hence we seek to maximize

$$\mathrm{Tr}[\mathbf{X}^T\mathbf{P}\mathbf{X}] = \mathrm{Tr}[\mathbf{X}^T\mathbf{U}_k\mathbf{U}_k^T\mathbf{X}] = \mathrm{Tr}[\mathbf{U}_k^T\mathbf{X}\mathbf{X}^T\mathbf{U}_k]$$

$$= \sum_{i=1}^{k}\Big(\sum_{j=1}^{N}(\mathbf{U}_k^T\mathbf{X}\mathbf{X}^T)_{ij}(\mathbf{U}_k)_{ji}\Big) = \sum_{i=1}^{k}\Big(\sum_{j=1}^{N}\Big(\sum_{\ell=1}^{N}(\mathbf{U}_k^T)_{i\ell}(\mathbf{X}\mathbf{X}^T)_{\ell j}\Big)(\mathbf{U}_k)_{ji}\Big)$$

so for $\mathbf{u}_i := ((\mathbf{U}_k)_{1i}, ..., (\mathbf{U}_k)_{Ni})$,

$$= \sum_{i=1}^{N}\Big(\mathbf{u}_i^T\mathbf{X}\mathbf{X}^T\mathbf{u}_i\Big)$$

where

$$\mathbf{P}\mathbf{X} = \mathbf{U}_k\mathbf{U}_k^T\mathbf{X}$$

so that $\mathbf{Y} := \mathbf{U}_k^T\mathbf{X}$ is a k-dimensional representation of $\mathbf{X}$.

Note: The top singular vectors of $\mathbf{C}$ are the directions of maximal variance in the data, and the $\mathbf{u}_i$ are the variances, so that PCA may be understood as projection onto the subspace of maximal variance.

b) In the 1-dimensional case, PCA seeks to minimize $||\mathbf{P}\mathbf{X} - \mathbf{X}||_F^2$, which by part a) gives the direction in which projection yields maximal variance.

Remark: In Kernel principle component analysis (KPCA), the feature map Φ send $\mathcal{X}$ to an arbitrary Reproducing Kernel Hilbert Space (RKHS) equipped with its own inner product (kernel function K).

Definition: Isomap extracts the low-dimensional data that best preserves pairwise distances between inputs based on their geodesic distances along a manifold. The algorithm is specified as follows:

1. Using the L_2 norm, find the t closest neighbors for each data point and construct an undirected neighborhod graph $\mathcal{G}$, in which points are nodes and links are edges.

2. Compute approximate geodesic distances Δ_{ij} between all pairs of nodes (i, j) by computing all-pairs shortest distances in $\mathcal{G}$.

3. Calculate the $m \times m$ similarity matrix as $\mathbf{K}_{\mathrm{Iso}} := -\frac{1}{2}(\mathbf{I}_m - \frac{1}{m}\mathbf{1}\mathbf{1}^T)\boldsymbol{\Delta}(\mathbf{I}_m - \frac{1}{m}\mathbf{1}\mathbf{1}^T)$, where $\mathbf{1}$ is a column vector of all ones and Δ is the squared distance matrix.

4. Find the optimal k-dimensional representation $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^n$ where

$$\mathbf{Y} = \mathrm{argmin}_{\mathbf{Y}'}\sum_{i,j}\Big(||\mathbf{y}_i' - \mathbf{y}_j'||_2^2 - \boldsymbol{\Delta}_{ij}^2\Big)$$

given by

$$\mathbf{Y} = (\boldsymbol{\Sigma}_{\mathrm{Iso,\ j}})^{\frac{1}{2}}\mathbf{U}_{\mathrm{Iso,k}}^T$$

Note that $\boldsymbol{\Sigma}_{\mathrm{Iso,\ j}}$ is the diagonal matrix of the top k singular values of $\mathbf{K}_{\mathrm{Iso}}$ and $\mathbf{u}_{\mathrm{Iso,\ k}}$ are the corresponding singular vectors. Further, $\mathbf{K}_{\mathrm{Iso}}$ serves as a kernel matrix (similarity matrix for data points in feature space) if it is positive semidefinite.

Definition The Laplacian Eigenmaps algorithm aims to find a k-dimensional representation of the data matrix $\mathbf{X}$ which best preserves the weighted neighborhood relations specified by a matrix $\mathbf{W}$:

1. Find the t nearest neighbors of each point

2. Define $\mathbf{W} \in \mathbb{R}^{m\times m}$ as $\mathbf{W}_{ij} := e^{\frac{||\mathbf{x}_i - \mathbf{x}_j||_2^2}{\sigma^2}}$ if $\mathbf{x}_i$ and $\mathbf{x}_j$ are neighbors, or as 0 otherwise, where σ is a scaling parameter.

3. Construct a diagonal matrix $\mathbf{D} \in \mathbb{R}^{m\times m}$ as $\mathbf{D}_{ii} = \sum_{j=1}^{m} \mathbf{W}_{ij}$.

4. Find $\mathbf{Y} \in \mathbb{R}^{k\times m}$ satisfying

$$\mathrm{argmin}_{\mathbf{Y}'}\Big\{\sum_{i,j} \mathbf{W}_{ij}||\mathbf{y}'_i - \mathbf{y}'_j||_2^2\Big\}$$

Intuitively, the above minimization penalizes k-dimensional representations of neighbors that differ largely under the L_2 norm.

Proposition (LE definition): The solution to the Laplacian eigenmap minimization is $\mathbf{U}_{\mathbf{L},k}^T$, where $\mathbf{L} = \mathbf{D} - \mathbf{W}$ is the "graph Laplacian" and $\mathbf{U}_{\mathbf{L},k}^T$ are the bottom k singular vectors of $\mathbf{L}$ (excluding 0 if the underlying neighborhood graph has connections).

Proof: We find that, for $\mathbf{x} \in \mathbb{R}^N$ and $\mathbf{Y} \in \mathbb{R}^{k\times m}$ we have

$$(\mathbf{Y}\mathbf{L}\mathbf{Y}^T)_{ij} = \sum_{\ell=1}^{m} \mathbf{Y}_{i\ell}^T(\mathbf{L}\mathbf{Y})_{\ell j} = \sum_{\ell=1}^{m} \mathbf{Y}_{\ell i}\Big(\sum_{t=1}^{m} \mathbf{L}_{\ell t}\mathbf{Y}_{tj}\Big)$$

$$(*) \quad = \sum_{\ell=1} \mathbf{Y}_{\ell i} \sum_{t\neq \ell} \mathbf{W}_{\ell t}(\mathbf{Y}_{\ell j} - \mathbf{Y}_{tj})$$

while

$$\sum_{i,\ell} \mathbf{W}_{i\ell}||\mathbf{y}'_i - \mathbf{y}'_\ell||_2^2 = \sum_{i=1}^{m}\sum_{\ell=1}^{m} \mathbf{W}_{i\ell}(\mathbf{y}'_i - \mathbf{y}'_\ell)^T(\mathbf{y}'_i - \mathbf{y}'_\ell)$$

$$= \sum_{i=1}^{m}\sum_{\ell=1}^{m} \mathbf{W}_{i\ell}((\mathbf{y}'_i)^2 - 2(\mathbf{y}'^T_\ell \mathbf{y}'_i) + (\mathbf{y}'_\ell)^2)$$

$$= \sum_{i=1}^{m}\sum_{\ell=1}^{m} \mathbf{W}_{i\ell}\Big(\sum_{j=1}^{m} (\mathbf{y}'_i)_j^2 - 2(\mathbf{y}'_\ell)_j(\mathbf{y}'_i)_j + (\mathbf{y}'_\ell)_j^2\Big)$$

$$= \sum_{i=1}^{m}\sum_{\ell=1}^{m} \mathbf{W}_{i\ell}\Big(\sum_{j=1}^{m} \mathbf{Y}'_{ji}{}^2 - 2\mathbf{Y}'_{j\ell}\mathbf{Y}'_{ji} + \mathbf{Y}'_{j\ell}{}^2\Big)$$

hence by $(*)$

$$= \sum_{i=1}^{k} (\mathbf{Y}'\mathbf{L}\mathbf{Y}'^T)_{ii}$$

so for $\mathbf{Y} := \mathbf{Y}'^T$, by the final simplication used in Theorem 15.1,

$$= \sum_{i=1}^{k} \mathbf{y}_i^T \mathbf{L} \mathbf{y}_i$$

Remark (PCA Gradient Descent): From Theorem 15.1, we have that

$$\frac{\partial}{\partial (U_k)_{ab}} ||PX - X||_F^2 = -\frac{\partial}{\partial (U_k)_{ab}} \sum_{i=1}^{k} \sum_{j=1}^{N} \sum_{\ell=1}^{N} (U_k^T)_{i\ell} (XX^T)_{\ell j} (U_k)_{ji}$$
$$= -\left(2(U_k)_{ab}(XX^T)_{aa} + \sum_{\ell \neq a}^{N} (U_k^T)_{b\ell}(XX^T)_{\ell a} + \sum_{j \neq a}^{N} (XX^T)_{aj}(U_k)_{jb} \right)$$
$$= -2 \sum_{\ell=1}^{N} (U_k)_{\ell b} (XX^T)_{a\ell}$$

since

$$XX_{ij}^T = \sum_{s=1}^{m} X_{is} X_{sj}^T = \sum_{s=1}^{m} X_{js} X_{si}^T = XX_{ji}^T$$

so for $F(U_k) = ||U_k U_k^T X - X||_F^2$ and $DF(U_k)_{ji} = \frac{\partial}{\partial (U_k)_{ji}} ||PX - X||_F^2$, we perform gradient descent steps as

$$U_k - \lambda DF(U_k)$$

for step size λ.

Ch. 15 Exercises.

15.1. Let $\mathbf{X}$ be an uncentered data matrix and let $\overline{\mathbf{x}} := \frac{1}{m} \sum_{i=1}^{N} \mathbf{x}_i$ be the sample mean of the columns of $\mathbf{X}$.

a) We require

$$\mathbf{C}_{ij} = \text{Cov}(\mathbf{\Phi}(x)_i, \mathbf{\Phi}(x)_j) = E[\mathbf{\Phi}(x)_i \mathbf{\Phi}(x)_j] - E[\mathbf{\Phi}(x)_i] E[\mathbf{\Phi}(x)_j]$$
$$= \frac{1}{m} \sum_{\ell=1}^{m} \mathbf{\Phi}(x_\ell)_i \mathbf{\Phi}(x_\ell)_j - \overline{\mathbf{x}}_i \overline{\mathbf{x}}_j = \frac{1}{m} \sum_{\ell=1}^{m} (\mathbf{x}_\ell)_i (\mathbf{x}_\ell)_j - \overline{\mathbf{x}}_i \overline{\mathbf{x}}_j$$
$$= \frac{1}{m} \Big(\sum_{\ell=1}^{m} (\mathbf{x}_\ell)_i (\mathbf{x}_\ell)_j - (\mathbf{x}_\ell)_i (\overline{\mathbf{x}}_j) - (\mathbf{x}_\ell)_j (\overline{\mathbf{x}}_i) + (\overline{\mathbf{x}}_i)(\overline{\mathbf{x}}_j) \Big)$$

hence

$$\mathbf{C} = \frac{1}{m} \sum_{\ell=1}^{m} (\mathbf{x}_\ell \mathbf{x}_\ell^T - \mathbf{x}_\ell \overline{\mathbf{x}}^T - \overline{\mathbf{x}}^T \mathbf{x}_\ell + \overline{\mathbf{x}}^T \overline{\mathbf{x}}) = \frac{1}{m} \sum_{i=1}^{m} (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T$$

Then, for a vector $\mathbf{u} \in \mathbb{R}^N$, we have

$$\text{Var}(\mathbf{u}^T \mathbf{x}_i) = E[(\mathbf{u}^T \mathbf{x}_i)^2] - E[\mathbf{u}^T \mathbf{x}_i]^2$$
$$= \frac{1}{m} \Big(\sum_{i=1}^{m} (\mathbf{u}^T \mathbf{x}_i)^2 \Big) - (\mathbf{u}^T \overline{\mathbf{x}})^2 = \frac{1}{m} \Big(\sum_{i=1}^{m} (\mathbf{u}^T \mathbf{x}_i)^2 - (\mathbf{u}^T \overline{\mathbf{x}})^2 \Big)$$
$$= \frac{1}{m} \sum_{i=1}^{m} \mathbf{u}^T (\mathbf{x}_i \mathbf{x}_i^T - \mathbf{x}_i \overline{\mathbf{x}}^T - \overline{\mathbf{x}}^T \mathbf{x}_i + \overline{\mathbf{x}}^T \overline{\mathbf{x}}) \mathbf{u} = \mathbf{u} \mathbf{C} \mathbf{u}^T$$

15.2. In this problem we prove the correctness of double centering (computing $\mathbf{K}_{\text{Iso}}$) using Euclidean distance. Define $\mathbf{X}$ as in 15.1, and define $\mathbf{X}^*$ to have $\mathbf{x}_i^* := \mathbf{x}_i - \overline{\mathbf{x}}$ as its i-th column. Let $\mathbf{K} := \mathbf{X}\mathbf{X}^T$ and let $\mathbf{D}$ denote the Euclidean distance matrix with $\mathbf{D}_{ij} = ||\mathbf{x}_i - \mathbf{x}_j||$. Further, let $\mathbf{\Delta}$ denote the squared distance matrix with $\mathbf{\Delta}_{ij} = \mathbf{D}_{ij}^2$.

a) We find that

$$\mathbf{K}_{ij} = \sum_{\ell=1}^{m} \mathbf{X}_{i\ell}^T \mathbf{X}_{\ell j} = \frac{1}{2}\Big(\sum_{\ell=1}^{m} \mathbf{X}_{\ell i}^2 - \mathbf{X}_{\ell i}^2 + \mathbf{X}_{\ell j}^2 - \mathbf{X}_{\ell j}^2 + 2\mathbf{X}_{\ell i}\mathbf{X}_{\ell j}\Big)$$
$$= \frac{1}{2}\Big(\sum_{\ell=1}^{m} \mathbf{X}_{\ell i}^2 + \mathbf{X}_{\ell j}^2 - (\mathbf{X}_{\ell j} - \mathbf{X}_{\ell i})^2\Big) = \frac{1}{2}(\mathbf{K}_{ii} + \mathbf{K}_{jj} - ||\mathbf{x}_i - \mathbf{x}_j||^2)$$
$$= \frac{1}{2}(\mathbf{K}_{ii} + \mathbf{K}_{jj} - \mathbf{D}_{ij}^2)$$

b) Let $\mathbf{K}^* := \mathbf{X}^{*T}\mathbf{X}^*$. We first find that

$$\frac{1}{m}(\mathbf{K}\mathbf{1}\mathbf{1}^T)_{ij} = \frac{1}{m}\sum_{t=1}^{m} \mathbf{K}_{it} = \frac{1}{m}\sum_{t=1}^{m}\sum_{\ell=1}^{m} \mathbf{X}_{\ell i}\mathbf{X}_{\ell t} = \sum_{\ell=1}^{m} (\overline{\mathbf{x}})_\ell (\mathbf{x}_i)_\ell$$
$$\frac{1}{m}(\mathbf{1}\mathbf{1}^T\mathbf{K})_{ij} = \frac{1}{m}\sum_{t=1}^{m} \mathbf{K}_{tj} = \frac{1}{m}\sum_{t=1}^{m}\sum_{\ell=1}^{m} \mathbf{X}_{\ell t}\mathbf{X}_{\ell j} = \sum_{\ell=1}^{m} (\overline{\mathbf{x}})_\ell (\mathbf{x}_j)_\ell$$

and

$$\frac{1}{m^2}(\mathbf{1}\mathbf{1}^T\mathbf{K}\mathbf{1}\mathbf{1}^T)_{ij} = \frac{1}{m^2}\sum_{t=1}^{m} (\mathbf{1}\mathbf{1}^T)_{it}(\mathbf{K}\mathbf{1}\mathbf{1}^T)_{tj} = \frac{1}{m}\sum_{t=1}^{m}\sum_{\ell=1}^{m} (\overline{\mathbf{x}})_\ell (\mathbf{x}_t)_\ell = \sum_{\ell=1}^{m} (\overline{\mathbf{x}}_\ell)^2$$

Then,

$$\mathbf{K}_{ij}^* = \sum_{\ell=1}^{N} \mathbf{X}_{i\ell}^{*\,T}\mathbf{X}_{\ell j}^* = \sum_{\ell=1}^{N} (\mathbf{x}_i - \overline{\mathbf{x}})_\ell (\mathbf{x}_j - \overline{\mathbf{x}})_\ell$$
$$= \sum_{\ell=1}^{N} (\mathbf{x}_i)_\ell (\mathbf{x}_j)_\ell - (\mathbf{x}_i)_\ell (\overline{\mathbf{x}})_\ell - (\mathbf{x}_j)_\ell (\overline{\mathbf{x}})_\ell + (\overline{\mathbf{x}})_\ell^2$$
$$= \mathbf{K}_{ij} - \frac{1}{m}(\mathbf{K}\mathbf{1}\mathbf{1}^T)_{ij} - \frac{1}{m}(\mathbf{1}\mathbf{1}^T\mathbf{K})_{ij} + \frac{1}{m^2}(\mathbf{1}\mathbf{1}^T\mathbf{K}\mathbf{1}\mathbf{1}^T)_{ij}$$

so that

$$\mathbf{K}^* = \mathbf{K} - \frac{1}{m}\mathbf{K}\mathbf{1}\mathbf{1}^T - \frac{1}{m}\mathbf{1}\mathbf{1}^T\mathbf{K} + \frac{1}{m^2}\mathbf{1}\mathbf{1}^T\mathbf{K}\mathbf{1}\mathbf{1}^T$$

c) We find that

$$\mathbf{K}_{ij}^* = \mathbf{K}_{ij} - \frac{1}{m}(\mathbf{K}\mathbf{1}\mathbf{1}^T)_{ij} - \frac{1}{m}(\mathbf{1}\mathbf{1}^T\mathbf{K})_{ij} + \frac{1}{m^2}(\mathbf{1}\mathbf{1}^T\mathbf{K}\mathbf{1}\mathbf{1}^T)_{ij}$$
$$= \frac{1}{2}(\mathbf{K}_{ii} + \mathbf{K}_{jj} - \mathbf{D}_{ij}^2) - \frac{1}{m}(\mathbf{K}\mathbf{1}\mathbf{1}^T)_{ij} - \frac{1}{m}(\mathbf{1}\mathbf{1}^T\mathbf{K})_{ij} + \frac{1}{m^2}(\mathbf{1}\mathbf{1}^T\mathbf{K}\mathbf{1}\mathbf{1}^T)_{ij}$$
$$= \frac{1}{2}(\mathbf{K}_{ii} + \mathbf{K}_{jj} - \mathbf{D}_{ij}^2) - \frac{1}{m}\sum_{t=1}^{m} \mathbf{K}_{it} - \frac{1}{m}\sum_{t=1}^{m} \mathbf{K}_{tj} + \frac{1}{m^2}\sum_{t=1}^{m}\sum_{\ell=1}^{m} \mathbf{K}_{t\ell}$$

$$= \frac{1}{2}(\mathbf{K}_{ii}+\mathbf{K}_{jj}-\mathbf{D}_{ij}^2)-\frac{1}{2m}\sum_{t=1}^{m}\Big((\mathbf{K}_{ii}+\mathbf{K}_{tt}-\mathbf{D}_{it}^2)+(\mathbf{K}_{tt}+\mathbf{K}_{jj}-\mathbf{D}_{tj}^2)-\frac{1}{m}\sum_{\ell=1}^{m}(\mathbf{K}_{tt}+\mathbf{K}_{\ell\ell}-\mathbf{D}_{t\ell}^2)\Big)$$

$$= \frac{1}{2}(-\mathbf{D}_{ij}^2) - \frac{1}{2m}\sum_{t=1}^{m}\Big((\mathbf{K}_{tt} - \mathbf{D}_{it}^2) - \mathbf{D}_{tj}^2 - \frac{1}{m}\sum_{\ell=1}^{m}(\mathbf{K}_{\ell\ell} - \mathbf{D}_{t\ell}^2)\Big)$$

$$= \frac{1}{2}\Big(- \mathbf{D}_{ij}^2 - \frac{1}{m}\sum_{t=1}^{m}(\mathbf{K}_{tt} - \mathbf{D}_{it}^2 - \mathbf{D}_{tj}^2) + \frac{1}{m^2}\sum_{t=1}^{m}\sum_{\ell=1}^{m}(\mathbf{K}_{\ell\ell} - \mathbf{D}_{t\ell}^2)\Big)$$

$$= -\frac{1}{2}\Big(\mathbf{D}_{ij}^2 - \frac{1}{m}\sum_{t=1}^{m}(\mathbf{D}_{it}^2 + \mathbf{D}_{tj}^2) + \frac{1}{m^2}\sum_{t=1}^{m}\sum_{\ell=1}^{m}\mathbf{D}_{t\ell}^2\Big)$$

d) We then find that

$$(\boldsymbol{\Delta}(\mathbf{I}_m - \frac{1}{m}\mathbf{1}\mathbf{1}^T))_{\ell j} = \boldsymbol{\Delta}_{\ell j} - \frac{1}{m}\sum_{t=1}^{m}\boldsymbol{\Delta}_{\ell t}$$

hence we may solve for $(\mathbf{H}\Delta\mathbf{H})_{ij}$ as

$$((\mathbf{I}_m - \frac{1}{m}\mathbf{1}\mathbf{1}^T)\boldsymbol{\Delta}(\mathbf{I}_m - \frac{1}{m}\mathbf{1}\mathbf{1}^T))_{ij} = \boldsymbol{\Delta}_{ij} - \frac{1}{m}\sum_{t=1}^{m}\boldsymbol{\Delta}_{it} - \frac{1}{m}\sum_{\ell=1}^{m}(\boldsymbol{\Delta}_{\ell j} - \frac{1}{m}\sum_{t=1}^{m}\boldsymbol{\Delta}_{\ell t})$$

$$= -2\mathbf{K}_{ij}^* \Rightarrow \mathbf{K}^* = -\frac{1}{2}\mathbf{H}\boldsymbol{\Delta}\mathbf{H}$$

15.3. Assume $k = 1$ and we seek a one-dimensional representation $\mathbf{y}$. By Proposition (LE Definition), the Laplacian eigenmap optimization problem is equivalent to $\mathbf{y} = \text{argmin}_{\mathbf{y}'}\mathbf{y}'^T\mathbf{L}\mathbf{y}'$

Remark: We now seek to understand such algorithms in the context of the Fenchel game no-regret dynamics framework (FGNRD) introduced by Wang-Abernethy-Levy.

Definition (Conjugate function): For a function $f : D \to \mathbb{R}\cup\infty$ where $D \subset \mathbb{R}^d$, we define its conjugate $f^* : \mathbb{R}^d \to \mathbb{R} \cup \infty$ as

$$f^*(y) := \sup_{x\in D}\{\langle y, x\rangle - f(x)\}$$

Proposition (Conjugate convex): Conjugate functions of convex functions are convex.

Proof: For $f : D \to \mathbb{R}$ convex where $D \subset \mathbb{R}^d$, we find that

$$f^*(\lambda x + (1-\lambda)y) = \sup_{x'\in D}\{\langle x', \lambda x + (1-\lambda)y\rangle - f(x')\}$$

$$= \sup_{x'\in D}\{\langle x', \lambda x + (1-\lambda)y\rangle - f(x')\}$$

$$= \sup_{x'\in D}\{\langle x', \lambda x\rangle + \langle x', y\rangle - \lambda\langle x', y\rangle - f(x')\}$$

$$= \sup_{x'\in D}\{\lambda\langle x, x'\rangle - \lambda f(x') + \langle y, x'\rangle - f(x') - \lambda\langle y, x'\rangle + \lambda f(x')\}$$

$$\begin{aligned}&= \sup_{x'\in D}\{\lambda(\langle x,x'\rangle - f(x')) + (1-\lambda)(\langle y,x'\rangle - f(x'))\}\\ &\le \lambda \sup_{x'\in D}\{\langle x,x'\rangle - f(x')\} + (1-\lambda)\sup_{x''\in D}\{\langle y,x''\rangle - f(x'')\}\\ &= \lambda f^*(x) + (1-\lambda) f^*(y)\end{aligned}$$

Definition (subdifferential): The subdifferential $\partial f(x)$ is the set of all subgradients of f at x, i.e.

$$\partial f(x) = \{f_x : f(z) \ge \langle f_x, z-x\rangle + f(x),\ \forall z\}$$

Proposition (Equivalence) : For a closed convex function $f : \mathbb{R}^d \to \mathbb{R}$, the following are equivalent:

I. $y \in \partial f(x)$

II. $x \in \partial f^*(y)$

III. $\langle x, y\rangle = f(x) + f^*(y)$

Proof: We first note that $f^*(x)$ is convex as the supremum over

First suppose $y \in \partial f(x)$, i.e. $f(z) - f(x) \ge \langle y, z-x\rangle$ for all $z \in \mathbb{R}^d$. NOT YET DONE!!

Definition (Payoff function) We define our two-input "payoff" function $g : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ as

$$g(x,y) := \langle x,y\rangle - f^*(y)$$

We will understand this function as a zero-sum game in which, if player 1 selects action x and player 2 selects action y, $g(x,y)$ is the "cost" for player 1 and the "gain" for player 2.

Definition (Min-max problems, Nash equilibrium): Given a zero-sum game with a payoff function $g(x,y)$ which is convex in x and concave in y, we define

$$V^* := \inf_{x\in\mathcal{X}}\sup_{y\in\mathcal{Y}} g(x,y)$$

We further define an "ϵ-equilibrium" of $g(.,.)$ as a pair $\widehat{x}, \widehat{y}$ for which

$$V^* - \epsilon \le \inf_{x\in\mathcal{X}} g(x,\widehat{y}) \le V^* \le \sup_{y\in\mathcal{Y}} g(\widehat{x},y) \le V^* + \epsilon$$

where $\mathcal{X}$ and $\mathcal{Y}$ are convex decision spaces of the x-player and y-player respectively.

Definition (Fenchel Game): To solve for $\inf_{x\in D} f(x)$, we define $g : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}$ as

$$g(x,y) := \langle x,y\rangle - f^*(y) = \langle x,y\rangle - \sup_{x'\in D}\{\langle x',y\rangle - f(x')\}$$

and attempt to find an ϵ-equilibrium for $g(x,y)$.

Proposition: An equilibrium for the Fenchel game function solves the minimization problem $\inf_{x\in D} f(x)$.

Proof: For an ϵ-equilibrium $\widehat{x}, \widehat{y}$ of g defined as above, we have

$$\inf_{x \in D} f(x) = -\sup_{x \in D}\{-f(x)\} = -\sup_{x' \in D}\{\langle x', y\rangle - \langle x', y\rangle - f(x')\} =: h(y)$$

so that

$$\inf_{x \in \mathcal{X}}\left\{\langle x, \widehat{y}\rangle - \sup_{x' \in D}\{\langle x', \widehat{y}\rangle - f(x')\}\right\} \leq h(\widehat{y}) \leq \sup_{y \in \mathcal{Y}}\left\{\langle \widehat{x}, y\rangle - \sup_{x' \in D}\{\langle x', y\rangle - f(x')\}\right\}$$

hence

$$(*) \quad |V^* - h(y)| \leq 2\epsilon$$

where

$$V^* = \inf_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}}\left\{\langle x, y\rangle - \sup_{x' \in D}\{\langle x', y\rangle - f(x')\}\right\}$$

and as $\epsilon \to 0$ we have

$$\begin{aligned} V^* &= \sup_{y \in \mathcal{Y}}\left\{\langle \widehat{x}, y\rangle - \sup_{x' \in D}\{\langle x', y\rangle - f(x')\}\right\} \\ &= \sup_{y \in \mathcal{Y}}\{\langle \widehat{x}, y\rangle - f^*(y)\} = f(\widehat{x}) \end{aligned}$$

which follows from Proposition (Equivalence)

Corollary (mine): If $(\widehat{x}, \widehat{y})$ is an ϵ-equilibrium of the Fenchel Game as defined above, then

$$|f(\widehat{x}) - \inf_x f(x)| \leq \epsilon$$

Proof: Follows from $(*)$ above for $\epsilon' := \frac{\epsilon}{2}$.

Definition (Online Convex Optimization): Online convex optimization works as follows. At each round t (of T many), the learner selects a point $z_t \in \mathcal{Z}$ and suffers a loss $\alpha_t \ell_t(z_t)$ for this selection, where $\boldsymbol{\alpha}$ is the weight vector and $\mathcal{Z} \subset \mathbb{R}^d$ is a convex decision set of actions.

In general it is assumed that, upon selecting z_t during round t, the learner has observed all loss functions $\alpha_1 \ell_1(.), ..., \alpha_{t-1}\ell_{t-1}(.)$ up to but not including time t. An exception to this are the "prescient" learners (whose algorithms, marked with a "+" superscript, have access to the loss ℓ_t prior to selecting z_t) maintain knowledge of the t-th loss function.

Algorithm 1 Protocol for weighted online convex optimization

Require: convex decision set $\mathcal{Z} \subset \mathbb{R}^d$
Require: number of rounds T
Require: weights $\alpha_1, \alpha_2, ..., \alpha_T > 0$
Require: algorithm OAlg
  for $t = 1, 2, \ldots, T$ **do**
    Return: $z_t \leftarrow$ OAlg
    Receive: $\alpha_t, \ell_t(\cdot) \rightarrow$ OAlg
    Evaluate: Loss $\leftarrow$ Loss + $\alpha_t \ell_t(z_t)$
  end for

Remark: The "OAlg" referenced above refers to an algorithm performed within the current algorithm, and "OAlg^X" will refer to the algorithm updating the x coordinate in the Fenchel Game No Regret Dynamics.

Definition (regret): We define a learner's "regret" as

$$\boldsymbol{\alpha}\text{-REG}^z(z^*) := \sum_{t=1}^{T} \alpha_t \ell_t(z_t) - \sum_{t=1}^{T} \alpha_t \ell_t(z^*)$$

where $z^* \in \mathcal{Z}$ is the "comparator" to which the online learner is compared. We further define "average regret" as that normalized by the time weight $A_T : \sum_{t=1}^{T} \alpha_t$ and denote it by

$$\overline{\boldsymbol{\alpha}\text{-REG}}^z(z^*) := \frac{\boldsymbol{\alpha}\text{-REG}^z(z^*)}{A_T}$$

Finally, "no-regret algorithms" guarantee $\overline{\boldsymbol{\alpha}\text{-REG}}^z(z^*) \to 0$ as $A_T \to \infty$

Definition (online learning strategies): The following batch-style online-learning strategies modify the central algorithm Follow The Leader (FTL):

Algorithm 2 Online Learning Strategies

Require: convex set $\mathcal{Z}$, initial point $z_{\text{init}} \in \mathcal{Z}$
Require: $\alpha_1, ..., \alpha_T > 0$, $\ell_1, ..., \ell_T : \mathcal{Z} \to \mathbb{R}$
FTL$[z_{\text{init}}]$: $z_t \leftarrow z_{\text{init}}$ **if** $t = 1$, **else**
$z_t \leftarrow \text{argmin}_{z \in \mathcal{Z}} \left(\sum_{s=1}^{t-1} \alpha_s \ell_s(z) \right)$
FTL$^+$ $z_t \leftarrow \text{argmin}_{z \in \mathcal{Z}} \left(\sum_{s=1}^{t} \alpha_s \ell_s(z) \right)$
FTRL$[R(.), \eta]$: $z_t \leftarrow \text{argmin}_{z \in \mathcal{Z}} \left(\sum_{s=1}^{t} \alpha_s \ell_s(z) + \frac{1}{\eta} R(z) \right)$

Vishnoi Problems (work in progress):

1. Let $f^0, f^1, ... : K \to \mathbb{R}$ be a sequence of convex and differentiable functions, and $x^0, x^1, ... \in K$ a sequence of points where $x^0 := \text{argmin}_x R(x)$ and $R : K \to \mathbb{R}$ is a convex regularizer. In this case, we define regret up to time T as

$$\text{Regret}_T := \sum_{t=0}^{T-1} f^t(x^t) - \min_{x \in K} \sum_{t=0}^{T-1} f^t(x)$$

and x^t is defined as follows (as in FTRL)

$$x^t := \text{argmin}_x \left(\sum_{i=0}^{t-1} f^i(x) + R(x) \right)$$

We further assume that the gradient of each f^i is bounded everywhere by G and the diameter of K is bounded by D.

(a) We wish to show

$$\text{Regret}_T \le \sum_{t=0}^{T-1} (f^t(x^t) - f^t(x^{t+1})) - R(x^0) + R(x^*)$$

for all $T \in \mathbb{N}_0$ where

$$x^* := \text{argmin}_{x \in K} \sum_{t=0}^{T-1} f^t(x)$$

Proof: We first use induction to show that

$$(*) \qquad \sum_{t=0}^{T-1} f^t(x^{t+1}) \leq \sum_{t=0}^{T-1} f^t(x^T)$$

As a base case, for $T = 1$ we find

$$f^0(x^1) \leq f^0(x^T)$$

as equality holds. We then assume the $T-1$ case $(*)$ and observe

$$\sum_{t=0}^{T} f^t(x^{t+1}) \leq f^T(x^{T+1}) + \sum_{t=0}^{T-1} f^t(x^T) \leq \sum_{t=0}^{T} f^t(x^{T+1})$$

Then, since we have

$$\sum_{t=0}^{T-1} f^t(x^T) + R(x^T) \leq \sum_{t=0}^{T-1} f^t(x^{T+1}) + R(x^{T+1})$$

To show

$$\sum_{t=0}^{T-1} f^t(x^{t+1}) - \min_x \sum_{t=0}^{T-1} f^t(x) \leq R(x^*) - R(x^0)$$

we first prove

As a base case, observe that

$$f^0(x^*) + R(x^*) \geq f^0(x^1) + R(x^1) \geq f^0(x^1) + R(x^0)$$

Then, as an inductive hypothesis suppose

$$\sum_{t=0}^{T-1} f^t(x_T^*) + R(x_T^*) \geq \sum_{t-0}^{T-1} f^t(x^{t+1}) + R(x^0)$$

where $x_T^* = \text{argmin}_x \sum_{t=0}^{T-1} f^t(x)$. In this case, we have that

$$\sum_{t=0}^{T} f^t(x^*) + R(x^*) \geq \sum_{t=0}^{T} f^t(x^{T+1}) + R(x^{T+1})$$

$$\geq f^T(x^{T+1}) + \sum_{t=0}^{T-1} f^t(x_T^*) + R(x^0)$$

(b) Given an $\epsilon > 0$, we now use this method for

$$R(x) := \frac{1}{\eta}||x||_2^2$$

such that

$$\frac{1}{T}\text{Regret}_T \leq \epsilon$$

Proof: We wish to find T and η for which

$$\text{Regret}_T \leq |f^0(x^0) - f^{T-1}(x^T)| + R(x^*) - R(x^0) \leq \epsilon T$$

$$\Rightarrow \frac{1}{T}\Big(|f^0(x^0) - f^{T-1}(x^T)| + \frac{1}{\eta}(||x^*||_2^2 - ||x^0||_2^2)\Big) \leq \epsilon$$

Hence, we choose $\eta = \frac{D}{G}$ and $T = \frac{2GD}{\epsilon}$ so that

$$\begin{aligned}\frac{1}{T}\Big(|f^0(x^0) - f^{T-1}(x^T)| &+ \frac{1}{\eta}(||x^*||_2^2 - ||x^0||_2^2)\Big) \\ &\leq \frac{\epsilon}{2GD}\Big(G||x^0 - x^T||_2 + \frac{G}{D}||x^* - x^0||_2^2\Big) \\ &\leq \frac{\epsilon}{2GD}\Big(GD + GD\Big) = \epsilon\end{aligned}$$

Lemma (Legendre): For convex and differentiable f, we have

$$y^* = \text{argmax}_y(\langle x, y\rangle - f(y)) \iff x = \nabla f(y^*)$$

Definition (First-order oracle): A first-order oracle for a function $f : \mathbb{R}^n \to \mathbb{R}$ is a primitive that, given $x \in \mathbb{Q}^n$, outputs the value $f(x) \in \mathbb{Q}$ and a vector $h(x) \in \mathbb{Q}^n$ such that, for any $z \in \mathbb{R}^n$,

$$f(z) \geq f(x) + \langle h(x), z - x\rangle$$

so $h(x) = \nabla f(x)$ for f differentiable, else it is a subgradient of f at x.

Definition (BESTRESP$^+[\ell]$**):** This strategy, for prescient learners, is simply given by

$$\text{argmin}_{z\in\mathcal{Z}}\{\ell_t(z)\}$$

Definition (FW): The Frank-Wolfe method accesses a linear optimization oracle and remains within the domain D:

Algorithm 3 Frank-Wolfe Method and its FGNRD equivalent

Require: L-smooth (Lipschitz constant L) function $f(\cdot)$
Require: convex domain $D \subset \mathbb{R}^d$
Require: arbitrary w_0, iterations T
 FW (iterative)
 $\gamma_t \leftarrow \frac{2}{t+1}$
 $v_t \leftarrow \text{argmin}_{v\in D}\langle v, \nabla f(w_{t-1})\rangle$
 $w_t \leftarrow w_{t-1} + \gamma_t(v_t - w_{t-1})$
 FGNRD Equivalent
 $g(x, y) := \langle x, y\rangle - f^*(y)$
 $\alpha_t \leftarrow t$
 $\text{OAlg}^Y := \text{FTL}[\nabla f(w_0)]$
 $\text{OAlg}^X := \text{BESTRESP}^+[\text{g}]$

Note that the FTL loss function at time t is $-g(x_t, \cdot)$ in this case, while the loss function for BESTRESP$^+$ is $g(\cdot, y_t)$.

Proof of equivalence: To show the equivalence of the above FGNRD and Frank-Wolfe algorithms, we prove that the following three equalities hold at every time step t:

$$\text{I. } \nabla f(w_{t-1}) = y_t,$$
$$\text{II. } v_t = x_t,$$
$$\text{III. } w_t = \overline{x}_t$$

where $\overline{x}_t := \frac{\sum_{s=1}^t \alpha_s x_s}{\sum_{s=1}^t \alpha_s}$ is the weighted-average point produced by the dynamic.

We proceed by induction. As a base case, for $t = 1$ we have $\nabla f(w_0) = y_0$. Then, we show I $\Rightarrow$ II $\Rightarrow$ III $\Rightarrow$ I (for $t + 1$). For I $\Rightarrow$ II we have

$$\begin{aligned}\nabla f(w_{t-1}) = y_t \Rightarrow x_t &= \operatorname{argmin}_{x\in D}(\langle x, y_t\rangle - f^*(y_t))\\ &= \operatorname{argmin}_{x\in D}(\langle x, \nabla f(w_{t-1})) = v_t\end{aligned}$$

For II $\Rightarrow$ III, we note that

$$\overline{x}_t = \overline{x}_{t-1} + \gamma_t(x_t - \overline{x}_{t-1}) \Rightarrow \frac{\sum_{s=1}^t \alpha_s x_s}{\sum_{s=1}^t \alpha_s} = \frac{\sum_{s=1}^t \alpha_s x_s}{\sum_{s=1}^t \alpha_s} + \gamma_t\left(\frac{\sum_{s=1}^{t-1} \alpha_s(x_t - x_s)}{\sum_{s=1}^{t-1} \alpha_s}\right)$$

$$\Rightarrow \gamma_t = \frac{\alpha_t \sum_{s=1}^{t-1} \alpha_s(x_t - x_s)}{(\sum_{s=1}^t \alpha_s)(\sum_{s=1}^{t-1} \alpha_s(x_t - x_s))} = \frac{\alpha_t}{\sum_{s=1}^t \alpha_s} = \frac{t}{\sum_{s=1}^t s} = \frac{2t}{t(t+1)} = \frac{2}{t+1}$$

so that $\overline{x}_t = w_t$ for $w_0 = \overline{x}_0$. Finally, for III $\Rightarrow$ I,

$$\begin{aligned}y_t = \operatorname{argmin}_y \sum_{s=1}^{t-1} \alpha_s(-g(x_s, y)) &= \operatorname{argmin}_y\left(-\frac{1}{\sum_{s=1}^{t-1} s}\sum_{s=1}^{t-1} sg(x_s, y)\right)\\ &= \operatorname{argmin}_y\left(\frac{1}{\sum_{s=1}^{t-1} s}\sum_{s=1}^{t-1} s(f^*(y) - \langle x_s, y\rangle)\right)\\ = \operatorname{argmin}_y\left(f^*(y) - \left\langle \frac{\sum_{s=1}^{t-1} sx_s}{\sum_{s=1}^{t-1} s}, y\right\rangle\right) &= \operatorname{argmax}_y \langle \overline{x}_{t-1}, y\rangle - f^*(y) = \nabla f(\overline{x}_{t-1})\\ &= \nabla f(w_{t-1})\end{aligned}$$

from III, so we are done. Note that the penultimate equality is due to Lemma (Legendre).