Note on Machine Learning Theory

Zepeng CHEN
The HK PolyU

Date: April 29, 2023

1 Introduction

In this note, we introduce the learning theory. Our goal is to learn a function $f : \mathcal{X} \to \mathcal{Y}$ from some input space $\mathcal{X}$ to some output or label space $\mathcal{Y}$. There is an underlying distribution $P_{X,Y}$ over $\mathcal{X} \times \mathcal{Y}$. However, it is normally unknown. On the basis of n i.i.d. samples $(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)$, we aim to provide a prediction $\hat{f}$ of f.

Definition 1.1 (Supervised Learning)

Definition 1.2 (Unsupervised Learning)

Definition 1.3 (Online Learning)

Definition 1.4 (Mathematical Machine Learning Model)

When we consider machine learning theory rather than classical statistics theory, it is because

- Statistics mostly focuses on asymptotic scenarios, and does not provide non-asymptotic guarantees.
- Statistics requires a well-behaved statistical model, e.g., distribution. However, real data, e.g., text and image, could be more complex.
- Statistics aims to find the entire prob distribution, which could be too costly to compute and analyze.

2 Probability Inequalities

Lemma 2.1 (The Union Bound)

Consider events $A_1, \cdots, A_t$, we have

$$\mathbb{P}(A_1 \cup \cdots \cup A_t) \leq \mathbb{P}(A_1) + \cdots + \mathbb{P}(A_t).$$

Note on *This bound is very useful in machine learning theory.*

2.1 Chernoff Bounds

Theorem 2.1 (Chernoff Bounds)

Let X be a random variable with moment generating function $M(t) = \mathbb{E}[e^{tX}]$. Then for any $\epsilon > 0$

$$P\{X \geq \epsilon\} \leq e^{-t\epsilon}M(t) \leq \inf_{t>0} e^{-t\epsilon}M(t) \quad \forall t > 0$$
$$P\{X \leq \epsilon\} \leq e^{-t\epsilon}M(t) \quad \forall t < 0$$

Proof For $t > 0$, based on Markov's inequality, we have

$$P\{X \geq \epsilon\} = P\left\{e^{tX} \geq e^{t\epsilon}\right\} \leq E\left[e^{tX}\right] e^{-t\epsilon}$$

And similarly, we can get another bound. Since the Chernoff bounds hold for all t in either the positive or negative quadrant, we obtain the best bound by using the t that minimizes $e^{-t\epsilon}M(t)$. ■

Corollary 2.1 (Chernoff Bounds for i.i.d. Samples)

Suppose X_i are i.i.d., then for any $t > 0$,

$$\mathbb{P}(\hat{\mu}_n - \mu \geq \epsilon) \leq (M_{X-\mu}(t)e^{-t\epsilon})^n \leq (\inf_{t>0} M_{X-\mu}(t)e^{-t\epsilon})^n.$$

Proof The important result is that Chernoff bounds "play nicely" with summations Ng, 2022, which is a consequence of the moment generating function, i.e., if X_i are independent, then

$$M_{X_1+\cdots+X_n}(t) = \prod_{i=1}^{n} M_{X_i}(t).$$

Thus,

$$\begin{aligned}
\mathbb{P}(\hat{\mu}_n - \mu \geq \epsilon) &= \mathbb{P}(\sum_{i=1}^{n} X_i - n\mu \geq n\epsilon) \\
&\leq e^{-tn\epsilon} M_{X_1+\cdots+X_n-n\mu}(t) \quad \text{(Chernoff Bounds)} \\
&= e^{-tn\epsilon} \prod_{i=1}^{n} M_{X_i-\mu}(t) \\
&= (M_{X-\mu}(t)e^{-t\epsilon})^n
\end{aligned}$$

■

Note on *The exponetial decay shown by Chernoff's inequality will be much faster than the $O(\frac{1}{n})$ decay suggested by Chebyshev's inequality.*

Corollary 2.2 (Chernoff Tail Bound for Gaussians)

The optimized Chernoff tail bound for $X \sim \mathcal{N}(0, \sigma^2)$ will be

$$\mathbb{P}(X \geq \epsilon) \leq \exp(-\frac{\epsilon^2}{2\sigma^2}).$$

Proof

$$\begin{aligned}\mathbb{P}(X \geq \epsilon) &\leq e^{-t\epsilon} M(t) \quad \text{(Chernoff's Bounds)} \\ &\leq \inf_{t>0} \exp(\frac{t^2\sigma^2}{2} - t\epsilon) \\ &= \exp(\inf_{t>0} \frac{t^2\sigma^2}{2} - t\epsilon) \quad (\exp \text{ is increasing}) \\ &= \exp(-\frac{\epsilon^2}{2\sigma^2}) \quad (t^* = \frac{\epsilon}{\sigma^2})\end{aligned}$$

■

Corollary 2.3 (Chernoff Tail Bound for Gaussian Samples)

Given IID samples $x_1, \cdots, x_n \sim \mathcal{N}(\mu, \sigma^2)$ we have the following error bound for empirical mean $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n x_i$:

$$\mathbb{P}(\hat{\mu}_n - \mu \geq \epsilon) \leq \exp(-\frac{n\epsilon^2}{2\sigma^2}).$$

Proof A direct result of Corollary 2.1 and Corollary 2.1. ■

2.2 Hoffeding's Inequality for Bounded Random Variables

Lemma 2.2 (Hoffeding's Lemma)

Suppose that r.v. X is bounded and satisfies $a \leq X \leq b$ for scalars $a, b \in \mathbb{R}$. Then, X is sub-Gaussian with parameter $\frac{(b-a)^2}{4}$, i.e., we have

$$\mathbb{E}[e^{t(X-\mathbb{E}[X])}] \leq \exp(\frac{(b-a)^2t^2}{8}).$$

Proof WLOG, by replacing Z by $X - \mathbb{E}[X]$, we can assume $\mathbb{E}[Z] = 0$, so that $a \leq 0 \leq b$. Since e^{tz} is convex, we have that for all $z \in [a, b]$,

$$e^{tz} \leq \frac{b-z}{b-a}e^{ta} + \frac{z-a}{b-a}e^{tb}.$$

Thus,

$$\begin{aligned}\mathbb{E}[e^{tZ}] &\leq \frac{b-\mathbb{E}[Z]}{b-a}e^{ta} + \frac{\mathbb{E}[Z]-a}{b-a}e^{tb} \\ &= \frac{b}{b-a}e^{ta} + \frac{-a}{b-a}e^{tb} \quad (\mathbb{E}[Z] = 0) \\ &= \exp(-\gamma u + \log(\gamma e^u + (1-\gamma))) = \exp(g(u)),\end{aligned}$$

where $u = t(b-a)$ and $\gamma = -\frac{a}{b-a}$ and the last equality can be established by solving $e^{g(u)} = \frac{b}{b-a}e^{ta} + \frac{-a}{b-a}e^{tb}$. Note that $g(0) = g'(0) = 0$ and $g''(u) \leq \frac{1}{4}$. By Taylor's theorem, we have

$$\exp(g(u)) = \exp(g(0) + ug'(0) + \frac{u^2}{2}g''(\varepsilon)) = \exp(\frac{u^2}{2}g''(\varepsilon)) \leq \exp(\frac{u^2}{8}) = \exp(\frac{(b-a)^2t^2}{8}).$$

■

Theorem 2.2 (Hoffeding's Inequality (Liao, 2020))

Suppose that r.v. $X_1, \cdots, X_n$ *are independent and bounded as* $a_i \leq X_i \leq b_i$. *Then, defining the empirical mean* $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n X_i$ *and underlying mean* $\mu = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i]$ *results in the following concentration inequality:*

$$\mathbb{P}(\mu - \hat{\mu}_n \geq \epsilon) \leq \exp(-\frac{2n^2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}),$$
$$\mathbb{P}(\hat{\mu}_n - \mu \geq \epsilon) \leq \exp(-\frac{2n^2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}).$$

Particularly, if $a_i = a$ *and* $b_i = b$, *we have*

$$\mathbb{P}(\hat{\mu}_n - \mu \geq \epsilon) \leq \exp(-\frac{2n\epsilon^2}{(b-a)^2}).$$

Proof

■

Theorem 2.3 (McDiarmid's Inequality (Liao, 2020))

Let $f : \mathcal{X}^n \to \mathbb{R}$ *be a function such that for every* $x_1, \cdots, x_n, x_1', \cdots, x_n' \in \mathcal{X}$ *the following bounded differences condition holds:*

$$\forall 1 \leq i \leq n : |f(x_1, \cdots, x_i, \cdots, x_n) - f(x_1, \cdots, x_i', \cdots, x_n)| \leq c_i.$$

Then, assuming $X_1, \cdots, X_n \in \mathcal{X}$ *are independent r.v., we have*

$$\mathbb{P}(f(X_1, \cdots, X_n) - \mathbb{E}[f(X_1, \cdots, X_n)] \geq \epsilon) \leq \exp(-\frac{2\epsilon^2}{\sum_{i=1}^n c_i^2}),$$
$$\mathbb{P}(\mathbb{E}[f(X_1, \cdots, X_n)] - f(X_1, \cdots, X_n) \geq \epsilon) \leq \exp(-\frac{2\epsilon^2}{\sum_{i=1}^n c_i^2}).$$

Note on *Hoeffding's inequality is a special case of McDiarmid's inequality for*

$$f(x_1, \cdots, x_n) = \frac{1}{n}\sum_{i=1}^n x_i.$$

3 Small Sample Theory

We summarize some useful definitions and theorems for limited sample theory which are not covered in large sample theory. For definitions and theorems in large sample theory, please refer to my notes on statistics.

3.1 Sub-Gaussian Random Variables

Definition 3.1 (Sub-Gaussian Random Variables)

X with mean μ is called as a sub-Gaussian r.v. with parameter σ^2 if the MGF of $X-\mu$ ($M_{X-\mu}$) satisfies the following inequality at every $t \in \mathbb{R}$:

$$M_{X-\mu}(t) := \mathbb{E}[e^{t(X-\mu)}] \leq \exp(\frac{\sigma^2 t^2}{2}).$$

Corollary 3.1 (Sum of Independent Sub-Gaussians)

If $X_1, \cdots, X_n$ are independent sub-Gaussian r.v. with parameters $\sigma_1^2, \cdots, \sigma_n^2$, then $\sum_{i=1}^n X_i$ will be sub-Gaussian with parameter $\sum_{i=1}^n \sigma_i^2$.

Proof

$$\begin{aligned} M_{\sum_{i=1}^n X_i - \mathbb{E}[\sum_{i=1}^n X_i]}(t) &= \mathbb{E}[e^{t(\sum_{i=1}^n X_i - \sum_{i=1}^n \mu_i)}] \\ &= \prod_{i=1}^n \mathbb{E}[e^{t(Z_i-\mu_i}] \quad \text{(Independence)} \\ &\leq \prod_{i=1}^n \exp(\frac{\sigma_i^2 t^2}{2}) = \exp(\frac{\sum_{i=1}^n \sigma_i^2 t^2}{2}) \end{aligned}$$

■

Corollary 3.2 (Scalar Product of Sub-Gaussians)

If X is sub-Gaussian r.v. with parameter σ^2, then cX for scalar $c \in \mathbb{R}$ will be sub-Gaussian with parameter $c^2\sigma^2$.

Corollary 3.3 (Chernoff Tail Bound for Sub-Gaussians)

The optimized Chernoff tail bound for a sub-Gaussian X with parameter σ^2 and mean μ will be

$$\mathbb{P}(X - \mu \geq \epsilon) \leq \exp(-\frac{\epsilon^2}{2\sigma^2}).$$

Proof

$$\begin{aligned} \mathbb{P}(X-\mu \geq \epsilon) &\leq e^{-t\epsilon} M_{X-\mu}(t) \quad \text{(Chernoff's Bounds)} \\ &\leq \exp(\frac{t^2\sigma^2}{2} - t\epsilon) \quad \text{(Definition of Sub-Gaussians)} \\ &\leq \exp(\inf_{t>0} \frac{t^2\sigma^2}{2} - t\epsilon) \quad (\exp \text{ is increasing}) \\ &= \exp(-\frac{\epsilon^2}{2\sigma^2}) \quad (t^* = \frac{\epsilon}{\sigma^2}) \end{aligned}$$

■

Corollary 3.4 (Chernoff-based Concentration Inequality for Sub-Gaussians)

If $x_1, \cdots, x_n$ are i.i.d. samples for sub-Gaussian X with parameter σ^2 and mean μ, we have the following error bound on their empirical mean $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n x_i$:

$$\mathbb{P}(\hat{\mu}_n - \mu \geq \epsilon) \leq \exp(-\frac{n\epsilon^2}{2\sigma^2}).$$

4 Model-based Statistical Learning

This section covers exponential families, maximum likelihood, method of moments and maximum entropy principle, which overlaps statistics. For more detail, please refer to my note on Advanced Statistics directly.

4.1 Information Theory (Braverman, 2011; Pillow, 2018)

Definition 4.1 (Entropy)

Given a probability vector $\mathbf{q} = [q_1, \cdots, q_k]$ *for a discrete random variable* X*, the (Shannon) entropy of* X *is defined as*

$$H_{\mathbf{q}}(X) = \sum_{i=1}^{k} q_i log \frac{1}{q_i} = -\mathbb{E}_q log q.$$

Note on *Entropy is simply the negative expected loglikelihood.*

Note on *The entropy value is always non-negative and concave.*

Note on *The entropy measures the uncertainty in a given distribution. Moreover, the entropy is upper-bounded by* $logk$ *(Jensen's Inequality). Particularly, the upper-bound is achieved by the discrete uniform distribution, i.e.,* $q_1 = \cdots = q_k = \frac{1}{k}$*. This can be proved by solving the entropy maximization problem:*

$$\begin{aligned} \max_{\mathbf{q} \in \mathbb{R}^k} \quad & \sum_{i=1}^{k} q_i log \frac{1}{q_i} \\ s.t. \quad & \sum_{i=1}^{k} q_i = 1, \\ & q_i \geq 0, i = 1, \cdots, k. \end{aligned}$$

Definition 4.2 (Conditional Entropy)

Definition 4.3 (Joint Entropy)

Note on *Cross-entropy is also the negative expected loglikelihood, and is not calculated under its truth, but under some other distribution.*

4.2 Maximum Entropy Principle

Definition 4.4 (Empirical Distribution)

Suppose we have n *observations such as* $x_1, \cdots, x_n$ *from an unknown distribution* p*. The empirical distribution is defined as* $\tilde{p} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(x = x_i)$.

Definition 4.5 (Maximum Entropy Principle)

Given samples $\mathbf{x}_1, \cdots, \mathbf{x}_n$*, we want to find a model in a set of probability distributions*

$$M_\phi := \left\{ q \in \mathcal{P}_\mathcal{X} : \mathbb{E}_{\hat{\boldsymbol{\theta}}}[\phi(\mathbf{x})] = \frac{1}{n}\sum_{i=1}^{n} \phi(\mathbf{x}_i) \right\},$$

conduct the inference and base the decision on the distribution maximizing the Entropy function:

$$\underset{q \in M_\phi}{\operatorname{argmax}} H_q(\mathbf{X}) := \sum_{\mathbf{x}\in\mathcal{X}} q(\mathbf{x}) \log \frac{1}{q(\mathbf{x})}.$$

Note on *This principle chooses the most uncertain model based on the given set* M*.*

Theorem 4.1

The distribution that maximizes the entropy is an exponential family model with feature function ϕ*.*

Proof Consider the maximum entropy problem

$$\begin{aligned} \max_{\mathbf{q}\in\mathbb{R}^{|\mathcal{X}|}} \quad & \sum_{\mathbf{x}\in\mathcal{X}} q_\mathbf{x} log \frac{1}{q_\mathbf{x}} = -\sum_{\mathbf{x}\in\mathcal{X}} q_\mathbf{x} log q_\mathbf{x} \\ \text{s.t.} \quad & \sum_{\mathbf{x}\in\mathcal{X}} q_\mathbf{x}\phi(\mathbf{x}) = \hat{\boldsymbol{\mu}}, \\ & \sum_{\mathbf{x}\in\mathcal{X}} q_\mathbf{x} = 1, \\ & q_\mathbf{x} \geq 0, \mathbf{x} \in \mathcal{X}, \end{aligned}$$

as a problem without inequality constraints, i.e.,

$$\begin{aligned} \max_{\mathbf{q}\in\mathbb{R}^{|\mathcal{X}|}} \quad & -\sum_{\mathbf{x}\in\mathcal{X}} q_\mathbf{x} log q_\mathbf{x} \\ \text{s.t.} \quad & \sum_{\mathbf{x}\in\mathcal{X}} q_\mathbf{x} \begin{bmatrix} \phi(\mathbf{x}) \\ 1 \end{bmatrix} = \begin{bmatrix} \hat{\boldsymbol{\mu}} \\ 1 \end{bmatrix}. \end{aligned}$$

Next we consider its Lagrangian problem

$$\mathcal{L}(\mathbf{q}, \boldsymbol{\gamma}) = \sum_{\mathbf{x}\in\mathcal{X}} q_\mathbf{x} \left(-log q_\mathbf{x} - \phi(\mathbf{x})^\top \boldsymbol{\gamma}_{1:k} - \gamma_{k+1} \right) + \hat{\boldsymbol{\mu}}^\top \boldsymbol{\gamma}_{1:k} + \gamma_{k+1},$$

the stationary KKT condition

$$\nabla_{q_\mathbf{x}} \mathcal{L}(\mathbf{q}, \boldsymbol{\gamma}) = -log q^*_\mathbf{x} - \phi(\mathbf{x})^\top \boldsymbol{\gamma}_{1:k} - \gamma_{k+1} + 1 = 0$$

leads to

$$q^*_\mathbf{x} = exp\left(-\phi(\mathbf{x})^\top \boldsymbol{\gamma}_{1:k} - \gamma_{k+1} + 1 \right) \geq 0.$$

Thus, $q^*_\mathbf{x}$ is also the optimal solution to the original problem. Moreover,

$$q^*_\mathbf{x} \propto exp\left(-\phi(\mathbf{x})^\top \boldsymbol{\gamma}_{1:k} \right)$$

leads to

$$q^*_\mathbf{x} = \frac{exp\left(-\phi(\mathbf{x})^\top \boldsymbol{\gamma}_{1:k} \right)}{\sum_{x\in\mathcal{X}} exp\left(-\phi(\mathbf{x})^\top \boldsymbol{\gamma}_{1:k} \right)}$$

due to the constraint that probability $q_{\mathbf{x}}$'s add up to 1. ■

Note on *Suppose $\Omega = \{0, 1\}$ and $\phi(x) = x$, then the maximum entropy principle leads to a binomial distribution. Suppose $\Omega = (-\infty, \infty)$ and $\phi(x) = [x, x^2]^\top$, then the maximum entropy principle leads to a Gaussian distribution.*

4.3 Maximum Relative Entropy Principle (Meila, 2012, Lec. 8)

Definition 4.6

Based on the logic of maximum entropy principle, suppose we also have a prior distribution q_0.

4.4 Minimum KL-Divergence

Definition 4.7 (Kullback-Leibler Divergence)

Let X be a random variable with possible outcomes $\mathcal{X}$ and let P and Q be two probability distributions on X. The KL-Divergence of P from Q is defined as:

$$KL[P||Q] = \sum_{x \in \mathcal{X}} p(x) log_b \frac{p(x)}{q(x)} = \mathbb{E}_p \log \frac{p}{q},$$

$$KL[P||Q] = \int_{\mathcal{X}} p(x) log_b \frac{p(x)}{q(x)} \, \mathrm{d}x.$$

Note on *KL-Divergence captures how much a model distribution function differs from the true distribution of the data. However, since KL-Divergence is asymmetric in (p, q). We should not call it as the 'distance' between two distributions (Nowak, 2009).*

Note on Perspective from Statistics (halvorsen, 2016) *If we have two hypothesis regarding which distribution is generating the data X, e.g., P and Q. Then $\frac{p(x)}{q(x)}$ is the likelihood ratio for testing H_0: Q against H_1: P. Since KL-Divergence $KL[P||Q]$ is the expected value of the loglikelihood ratio under the alternative hypothesis, so it is a measure of the difficulty of this test. The asymmetry of KL-Divergence actually reflects the asymmetry between null and laternative hypothesis.*

For example, let P be the t_1-distribution and Q be the standard normal distribution. Then

$$KL(P||Q) \approx \infty,$$

$$KL(Q||P) \approx 0.26.$$

That is, if the null model is normal but the data is generated from t-distribution, then it is quite easy to reject the null! The logic here is that data from t-distribution do not look like normal. However, if the null is t and data is normal. Normal distributed data could look like t data.

Corollary 4.1 (KL-Divergence and Entorpy (Mao, 2019))

Lemma 4.1

KL-Divergence is always non-negative. Moreover, $KL(p||q) = 0$ *iff* $p = q$.

Proof By Gibbs' Inequality. ■

Lemma 4.2 (KL-Divergence's Convexity (Soch, 2020))

KL-Divergence is convex in the pair of probability distributions (p, q)*, i.e.,*

$$KL[\lambda p_1 + (1-\lambda)p_2||\lambda q_1 + (1-\lambda)q_2] \leq \lambda KL[p_1||q_1] + (1-\lambda)KL[p_2||q_2],$$

where (p_1, q_1) *and* (p_2, q_2) *are two pairs of probability distributions and* $0 \leq \lambda \leq 1$*. Moreover, KL-divergence attains the minimum* 0 *if* $p = q$.

Proof Firstly, we show that KL-Divergence is bi-convex of p, q, i.e., it is a convex function of p for a fixed q and vice versa,

$$KL[\lambda p_1 + (1-\lambda)p_2||q_1] \leq \lambda KL[p_1||q_1] + (1-\lambda)KL[p_2||q_1],$$
$$KL[p_1||\lambda q_1 + (1-\lambda)q_2] \leq \lambda KL[p_1||q_1] + (1-\lambda)KL[p_1||q_2]$$

on the basis of $xlogx$'s convexity. Next we can prove the convexity from bi-convexity. ■

4.5 Connections

4.5.1 Maximum Entropy and MLE (Lacoste-Julien, 2022)

Theorem 4.2 (Maximum Entropy v.s. MLE)

The maximum entropy problem over M_ϕ *is the dual problem to the MLE for the exponential family with feature function* ϕ.

Proof Suppose we replace the Lagrangian problem with $q^*_{\mathbf{x}}$, the dual of the maximum entropy problem can be written as

$$\min_{\boldsymbol{\gamma}} log\left(\sum_{x\in\mathcal{X}} exp\left(-\phi(\mathbf{x})^\top \boldsymbol{\gamma}_{1:k}\right)\right) + \hat{\boldsymbol{\mu}}^\top \boldsymbol{\gamma}_{1:k}.$$

■

4.5.2 Minimum KL-Divergence and MLE (Lacoste-Julien, 2022)

Theorem 4.3

Let $\{q_\theta\}_{\theta\in\Theta}$ *be a parametric family of distributions, and suppose* $p_n(x)$ *is the empirical pdf from* n *samples, MLE is minimizing KL-Divergence, i.e.,*

$$\arg\min_{\theta\in\Theta} KL(p_n||q_\theta) = \arg\max_{\theta} p(x|\theta).$$

Proof

$$\begin{aligned}\arg\min_{\theta} KL(p||q) &= \arg\min_{\theta} \mathbb{E}_{x\sim p}\left[log\frac{p(x)}{q(x)}\right] \\ &\iff \arg\min_{\theta} \mathbb{E}_{x\sim p}\left[-logq(x)\right] \\ &\iff \arg\max_{\theta} \mathbb{E}_{x\sim p}\left[logq(x)\right]\end{aligned}$$

■

4.5.3 Minimum KL-Divergence and Maximum Entropy

Theorem 4.4 (Minimum KL-Divergence and Maximum Entropy)

The model with maximum entropy is equivalent to the minimum KL divergence to the uniform distribution.

5 Model-free Machine Learning

Compared to model-based learning, here the underlying distribution is unknown. We introduce the theory of supervised learning in this section. In addition, we focus on non-asymptotic rather than asymptotic analysis. Here we provide convergence guarantees without having the number of observations n go to infinity. A key tool for proving such guarantees is uniform convergence (Schramm, 2022, Ch. 4), e.g.,

$$\mathbb{P}[\hat{L}(f) - L(f) \leq \epsilon] \geq 1 - \delta.$$

In other words, the probability that the difference between empirical loss and population loss is larger than ϵ is at most δ.

5.1 Empirical Risk Minimization

Definition 5.1 (Loss Function)

A standard goal in supervised learning is to minimize the averaged prediction loss $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^{\geq 0}$ *where* $\ell(\hat{y}, y)$ *measures the loss suffered under prediction* $\hat{y}$ *for actual label* y.

Note on Example of Loss Function

- *Squared-error loss:* $\ell_2(\hat{y}, y) = (\hat{y} - y)^2$,
- *0-1 loss:* $\ell_{0/1}(\hat{y}, y) = I(\hat{y} \neq y)$.
- *Absolute loss:* $\ell = |f(x) - y|$.

Definition 5.2 (Population Risk Minimization)

Given ℓ, the supervised learning goal is to find a prediction function $f \in \mathcal{F}$ to minimize the expected loss under distribution $P_{X,Y}$, i.e., population risk:

$$\min_{f \in \mathcal{F}} L(f) = \mathbb{E}_{P_{X,Y}}[\ell(f(X), Y)].$$

Let f^* denotes the function with minimum population risk, i.e., $f^* = \arg\min_{f \in \mathcal{F}} L(f)$.

Note on *This problem cannot be solved, since we do not know $P_{X,Y}$.*

Definition 5.3 (Empirical Risk Minimization)

Given loss function ℓ and training data $\{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)\}$, the empirical risk minimization (ERM) approach finds the prediction rule $\hat{f} \in \mathcal{F}$ minimizing the empirical expected loss or empirical risk:

$$\min_{f \in \mathcal{F}} \hat{L}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(\mathbf{x}_i), y_i).$$

Let $\hat{f}_n$ denotes the function with minimum empirical risk, i.e., $\hat{f}_n = \arg\min_{f \in \mathcal{F}} \hat{L}(f)$. Moreover, the expectation of empirical risk is exactly the population risk (Danica, 2018), i.e.,

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

Proof

$$\begin{aligned}
\mathbb{E}[\hat{L}(f)] &= \mathbb{E}[\frac{1}{n} \sum_{i=1}^{n} \ell(f(\mathbf{x}_i), y_i)] \\
&= \frac{1}{n} \sum_{i=1}^{n} \int \ell(f(\mathbf{x}_i), y_i) \, \mathrm{d}P_{\mathbf{x}_i), y_i} \\
&= \frac{1}{n} \sum_{i=1}^{n} L(f) = L(f).
\end{aligned}$$

■

Note on *The least squares problem is exactly the adoption of ERM with squared-error loss and linear model.*

Definition 5.4 (Generalization Risk)

The generalization risk of a prediction function $\hat{f} \in \mathcal{F}$ is defined as the difference between its empirical and population risks:

$$\epsilon_{gen}(\hat{f}) = L(\hat{f}) - \hat{L}(\hat{f}).$$

Note on *If the loss function is bounded, the generalization risk of f^* can be bounded by $\mathcal{O}(\frac{1}{\sqrt{n}})$ via Hoeffding's inequality (Schramm, 2022, Ch. 4).*

Definition 5.5 (Excess Risk)

The excess risk of a prediction function $\hat{f}_n \in \mathcal{F}$ is defined as the difference between its population risk and the population risk of optimal f^*:

$$\begin{aligned}\epsilon_{excess}(\hat{f}_n) &= L(\hat{f}_n) - L(f^*) \\ &= \underbrace{L(\hat{f}_n) - \hat{L}(\hat{f}_n)}_{\text{Gen. Risk for } \hat{f}_n} + \underbrace{\hat{L}(\hat{f}_n) - \hat{L}(f^*)}_{\leq 0} + \underbrace{\hat{L}(f^*) - L(f^*)}_{\text{Gen. Risk for } f^*}.\end{aligned}$$

Note on *The excess risk of every function in $\mathcal{F}$ is non-negative, but the generalization risk may be negative.*

Note on *While f^* is a deterministic function, $\hat{f}_n$ is a random function affected by the randomness of the training samples.*

Note on *A central goal of learning theory is to bound the excess risk.*

5.2 Finite Hypothesis Sets & Uniform Convergence Bounds

Theorem 5.1 (Excess Risk Bound for Finite Function Sets with Realizability Assumption)

Given

1. *0-1 loss: $\ell_{0/1}(\hat{y}, y) = I(\hat{y} \neq y)$,*
2. *realizability assumption: there exists a realizable scenario where $L(f^*) = 0$,*
3. *and a finite function set $\mathcal{F} = \{f_1, \cdots, f_t\}$ with t functions,*

the population risk bound holds for the ERM solution $\hat{f}$ with prob at least $1 - \delta$:

$$\epsilon_{excess}(\hat{f}) = L(\hat{f}) \leq \frac{\log t + \log \frac{1}{\delta}}{n}.$$

Proof Firstly, the realizability assumption implies that $\hat{L}(\hat{f}) = 0$. Since $\hat{L}(\hat{f}) \leq \hat{L}(f^*) = 0$ (Cuong, 2019), and the equality holds since S is a sample from $P_{X,Y}$.

For $\epsilon \geq 0$, define $F_\epsilon = \{f \in \mathcal{F} : L(f) \geq \epsilon\}$. Then our goal is to bound the probability $\mathbb{P}(\hat{f} \in F_\epsilon)$, i.e.,

$$\mathbb{P}(\hat{f} \in F_\epsilon) = \mathbb{P}(\exists f \in F_\epsilon : \hat{L}(f) = 0).$$

If we assume $f \in \mathcal{F}_\epsilon$, then given the 0/1 loss we have Why?

$$\mathbb{P}(\hat{L}(f) = 0) = (1 - L(f))^n \leq (1 - \epsilon)^n \leq e^{-n\epsilon}.$$

Then we use the union bound to show

$$\mathbb{P}(\exists f \in F_\epsilon : \hat{L}(f) = 0) \leq \sum_{f \in F_\epsilon} \mathbb{P}(\hat{L}(f) = 0) \leq |F_\epsilon|(1 - \epsilon)^n \leq te^{-n\epsilon}.$$

That is,

$$\mathbb{P}(\hat{f} \in F_\epsilon) \leq te^{-n\epsilon} \rightarrow \mathbb{P}(L(\hat{f}) \geq \frac{\log t + \log \frac{1}{\delta}}{n}) \leq \delta.$$

■

Note on *The risk bound is based on two restrictive assumptions: the realizability condition and*

the finiteness of the hypothesis set $\mathcal{F}$.

Theorem 5.2 (Excess Risk Bound for Finite Function Sets without Realizability Assumption)

Given

1. *0-1 loss: $\ell_{0/1}(\hat{y}, y) = I(\hat{y} \neq y)$,*
2. *and a finite function set $\mathcal{F} = \{f_1, \cdots, f_t\}$ with t functions,*

the excess risk bound holds for the ERM solution $\hat{f}$ with prob at least $1 - \delta$:

$$\epsilon_{excess}(\hat{f}) \leq \sqrt{\frac{2\log t + 2\log\frac{2}{\delta}}{n}} = \mathcal{O}(\sqrt{\frac{\log(t/\delta)}{n}}).$$

Proof The second part of excess risk is negative since $\hat{f}_n$ is a minimizer of $\hat{L}$. This allows us to write

$$\begin{aligned}\epsilon_{excess}(\hat{f}) &\leq \underbrace{|L(\hat{f}) - \hat{L}(\hat{f})|}_{\text{Gen. Risk for } \hat{f}} + \underbrace{|\hat{L}(f^*) - L(f^*)|}_{\text{Gen. Risk for } f^*} \\ &\leq 2\sup_{f\in\mathcal{F}} |L(f) - \hat{L}(f)|.\end{aligned}$$

Thus, if $\sup_{f\in\mathcal{F}} |L(f) - \hat{L}(f)|$ is small, i.e., $\sup_{f\in\mathcal{F}} |L(f) - \hat{L}(f)| \leq \frac{\epsilon}{2}$, then excess risk is less than ϵ, i.e., $\epsilon_{excess}(\hat{f}) \leq \epsilon$.

In other words, given the best population and empirical risk functions $f^*, \hat{f} \in \mathcal{F}$, the probability of an ϵ-large excess risk is bounded as:

$$\mathbb{P}(L(\hat{f}) - L(f^*) \geq \epsilon) \leq \mathbb{P}(\sup_{f\in\mathcal{F}} |L(f) - \hat{L}(f)| \geq \frac{\epsilon}{2}),$$

i.e.,

$$\mathbb{P}(\epsilon_{excess} \geq \epsilon) \leq \mathbb{P}(\sup_{f\in\mathcal{F}} |\epsilon_{gen}| \geq \frac{\epsilon}{2}).$$

$$\begin{aligned}\mathbb{P}(\sup_{f\in\mathcal{F}} |L(f) - \hat{L}(f)| \geq \frac{\epsilon}{2}) &= \mathbb{P}(\sup_{1\leq i\leq t} |L(f_i) - \hat{L}(f_i)| \geq \frac{\epsilon}{2}) \quad \text{(Finite function set)} \\ &\leq \sum_{i=1}^{t} \mathbb{P}(|L(f_i) - \hat{L}(f_i)| \geq \frac{\epsilon}{2}) \quad \text{(Union bound)} \\ &\leq \sum_{i=1}^{t} 2\exp\left(-\frac{2n(\epsilon/2)^2}{(1-0)^2}\right) = 2t\exp\left(-\frac{n\epsilon^2}{2}\right) \quad \text{(Hoffeding's Inequality)}\end{aligned}$$

Thus, if we define $\delta = 2t\exp\left(-\frac{n\epsilon^2}{2}\right)$, it turns out that $\epsilon = \sqrt{\frac{2\log(2t/\delta)}{n}}$, which completes the proof:

$$\mathbb{P}\left(\epsilon_{excess}(\hat{f}) \geq \sqrt{\frac{2\log(2t/\delta)}{n}}\right) \leq \delta.$$

■

Note on Difference between bounds with and without realizability assumption

- *The $O(\frac{1}{n})$ risk bound in the realizable case (noiseless setting) is vanishing faster than the $O(\frac{1}{\sqrt{n}})$ bound in the non-realizable case (noisy setting).*
- *In learning theory, the risk bounds that decay with $O(\frac{1}{n})$ are called fast rates bounds,*

which require extra assumptions on the learning setting, e.g. realizability or norm-based regularization.

5.3 Rademacher Complexity

5.3.1 Definitions of Rademacher Complexity

Definition 5.6 (Rademacher Random Variable)

A Rademacher r.v. σ is defined as uniformly distributing over $\{-1, +1\}$, i.e.,

$$\mathbb{P}(\sigma = 1) = \mathbb{R}(\sigma = -1) = \frac{1}{2}$$

.

Note on Application *To get rid of the virtual dataset, we use independent Rademacher r.v. $\sigma_1, \cdots, \sigma_n$. Due to the symmetry, the following equations hold:*

$$\begin{aligned} & X_i - X_i' \overset{dist.}{=} \sigma_i(X_i - X_i') \\ \Rightarrow & \frac{1}{n}\sum_{i=1}^{n}(X_i - X_i') \overset{dist.}{=} \frac{1}{n}\sum_{i=1}^{n}(\sigma_i X_i - \sigma_i X_i') \\ \Rightarrow & L_S(f) - L_{S'}(f) \overset{dist.}{=} \frac{1}{n}\sum_{i=1}^{n}(\sigma_i X_i - \sigma_i X_i'). \end{aligned}$$

Definition 5.7 (Rademacher Complexity)

For a function set $\mathcal{H}$ and Rademacher variables in $\boldsymbol{\sigma} = [\sigma_1, \cdots, \sigma_n]$, we define $\mathcal{H}$'s Rademacher complexity as

$$R_n(\mathcal{H}) := \mathbb{E}_{S,\boldsymbol{\sigma}}[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i h(X_i)].$$

Note on Motivation from binary classification *(Chen, 2013, Lec. 9) Suppose f is a classification function which maps data x_i to its label $\sigma_i \in \{-1, 1\}$. Since f is dependent on σ_i, to measure how well $\mathcal{H}$ can correlate with random noise, we take the expectation of the correlation over σ_i, i.e., Rademacher complexity. This intuitively measures the expressiveness of $\mathcal{H}$. For example, $|\mathcal{H}| = 1$ where we only have one choice for a hypothesis, our expectation equals 0 since the max term disappears; and $|\mathcal{H}| = 2^n$ where $\mathcal{H}$ shatters S, our expectation equals 1 since there always exists a hypothesis matching any set of σ_i's. That is, this measure must fall between 0 and 1.*

Corollary 5.1 (Basic Properties of Rademacher Complexity)

1. ***Monotonicity.*** *If $\mathcal{H}_1 \subseteq \mathcal{H}_2$, then $R_n(\mathcal{H}_1) \leq R_n(\mathcal{H}_2)$.*
2. ***Singleton Set.*** *If $\mathcal{H} = \{h\}$ contains only one function, then $R_n(\mathcal{H}) = 0$.*
3. ***Scalar Product.*** *If $c\mathcal{H} = \{ch : h \in \mathcal{H}\}$, then $R_n(\mathcal{H} = |c|R_n(\mathcal{H})$.*
4. ***Lipschitz Composition.*** *If $g : \mathbb{R} \to \mathbb{R}$ is a ρ-Lipschitz function, i.e.,*

$$\forall z, z' \in \mathbb{R} : |g(z) - g(z')| \leq \rho|z - z'|,$$

then $R_n(g \circ \mathcal{H} \leq \rho R_n(\mathcal{H})$.

5. ***Convex Hull.*** *For a function set* $\mathcal{H} = \{h_1, \cdots, h_2\}$*, we define its convex hull:*

$$\textit{convex-hull}(\mathcal{H}) := \{\sum_{i=1}^{t} \alpha_i h_i : \alpha_1, \cdots, \alpha_t \geq 0, \sum_{i=1}^{t} = 1\}.$$

Then, $R_n(\textit{convex-hull}(\mathcal{H})) = R_n(\mathcal{H})$.

Proof

■

Definition 5.8 (Empirical Rademacher Complexity)

For a function set $\mathcal{H}$*, Rademacher variables in* $\boldsymbol{\sigma} = [\sigma_1, \cdots, \sigma_n]$ *and a fixed dataset* $S = \{x_1, \cdots, x_n\}$*, we define* $\mathcal{H}$*'s empirical Rademacher complexity as*

$$\hat{R}_n(\mathcal{H}) := \mathbb{E}_{\boldsymbol{\sigma}}[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i h(x_i)].$$

In other words,

$$R_n(\mathcal{H}) = \mathbb{E}_S[\hat{R}_n(\mathcal{H})].$$

5.3.2 Bounds of Rademacher Complexity

Lemma 5.1 (Difference of Rademacher Complexity(Liao, 2020))

Suppose the loss function is bounded as $0 \leq \ell(y, \hat{y}) \leq c$*. For any* ϵ*,*

$$\mathbb{P}(\hat{R}_n(\mathcal{H}) - R_n(\mathcal{H})) \leq \exp(-2n\epsilon^2/c^2),$$

$$\mathbb{P}(R_n(\mathcal{H}) - \hat{R}_n(\mathcal{H})) \leq \exp(-2n\epsilon^2/c^2).$$

Proof According to the definition of $\hat{R}_n(\mathcal{H})$, any change of one of the samples, e.g., x_i, would change $\hat{R}_n(\mathcal{H})$ by at most c/n. Therefore, we could apply the McDiarmid's inequality to obtain both two inequations. ■

Lemma 5.2 (Massart Lemma)

Suppose that $\mathcal{H} = \{h_1, \cdots, h_t\}$ *is a finite set of* t *functions. Also, suppose that for every* $h \in \mathcal{H}$ *and dataset* $S = \{x_1, \cdots, x_n\}$ *the following holds:*

$$\frac{1}{n} \sum_{i=1}^{n} h(x_i)^2 \leq M.$$

Then, the following bound on the empirical Rademacher complexity holds:

$$\hat{R}_n(\mathcal{H}) \leq \sqrt{\frac{2M \log t}{n}}.$$

Proof Slide 14 ■

Note on *Massart lemma shows that the Rademacher complexity of a finite function set of size* t *is bounded by* $\mathcal{O}(\sqrt{\frac{\log t}{n}})$.

Corollary 5.2 (Empirical Rademacher Complexity of ℓ_2-Norm-bounded Linear Functions)

Consider the following set of ℓ_2-norm-bounded linear functions:

$$\mathcal{H} = \{h_{\boldsymbol{w}}(\mathbf{x}) = \boldsymbol{w}^\top \mathbf{x} : ||\boldsymbol{w}||_2 \leq M\}.$$

Then for a dataset $S = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$, we have the following bound on the empirical Rademacher complexity:

$$\hat{R}_n(\mathcal{H}) \leq \frac{M \max_i ||\mathbf{x}_i||_2}{\sqrt{n}}.$$

Corollary 5.3 (Empirical Rademacher Complexity of ℓ_1-Norm-bounded Linear Functions)

Consider the following set of ℓ_1-norm-bounded linear functions on a d-dimensional $\mathbf{x} \in \mathbb{R}^d$:

$$\mathcal{H} = \{h_{\boldsymbol{w}}(\mathbf{x}) = \boldsymbol{w}^\top \mathbf{x} : ||\boldsymbol{w}||_1 \leq M\}.$$

Then for a dataset $S = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$, we have the following bound on the empirical Rademacher complexity:

$$\hat{R}_n(\mathcal{H}) \leq M \max_i ||\mathbf{x}_i||_\infty \sqrt{\frac{2\log(2d)}{n}}.$$

Proof Slide 14 ■

5.3.3 Rademacher Complexity of ReLU-based Neural Nets

Definition 5.9 (Frobenius Norm)

Given a matrix $W \in \mathbb{R}^{d \times t}$, we define its Frobenius norm as

$$||W||_F = \sqrt{\sum_{i=1}^{d}\sum_{j=1}^{t} w_{ij}^2}.$$

Definition 5.10 (ReLU Function)

The ReLU function is defined as $\psi_{ReLU}(x) = \max\{0, x\}$.

Corollary 5.4 (Rademacher Complexity of ReLU-based Nueral Nets)

Consider the following set of L-layer neural nets with ReLU activation function:

$$\mathcal{H} = \{h_{\boldsymbol{w}}(\mathbf{x}) = W\}$$

TBD

5.4 Infinite Hypothesis Sets

Note that we still have

$$\mathbb{P}(L(\hat{f}) - L(f^*) \geq \epsilon) \leq \mathbb{P}(\sup_{f \in \mathcal{F}} |L(f) - \hat{L}(f)| \geq \frac{\epsilon}{2}),$$

i.e.,

$$\mathbb{P}(\epsilon_{excess} \geq \epsilon) \leq \mathbb{P}(\sup_{f \in \mathcal{F}} |\epsilon_{gen}| \geq \frac{\epsilon}{2}),$$

when the hypothesis set infinite. However, we cannot use Hoffeding's inequality directly since the hypothesis set infinite. We firstly bound the worst-case generalization risk G_n and use this bound to bound the excess risk.

Definition 5.11 (Worst-case Generalization Risk)

Given a random dataset of size n, the worst-case generalization risk is defined as

$$G_n := \sup_{f \in \mathcal{F}} L(f) - \hat{L}(f).$$

Lemma 5.3 (Concentration Bound for Worst-case Generalization Risk)

Suppose the loss function is bounded as $0 \leq \ell(y, \hat{y}) \leq c$. Consider the worst-case generalization risk G_n as a function of independent empirical samples $X_1, \cdots, X_n$. Then

$$\mathbb{P}(G_n \geq \mathbb{E}[G_n] + \epsilon) \leq \exp(-\frac{2n\epsilon^2}{c^2}).$$

Particularly, given the 0-1 loss function, i.e., $c = 1$, we have

$$\mathbb{P}(G_n \geq \mathbb{E}[G_n] + \epsilon) \leq \exp(-2n\epsilon^2).$$

Proof According to the definition of $\hat{L}$, any change of one of the samples, e.g., $\ell(f(\mathbf{x}_i), y_i)$, would change $\hat{L}$ by at most c/n. And L does not depend on samples. Therefore, we could apply the McDiarmid's inequality to obtain the inequation. ■

Lemma 5.4 (Symmetrization Bound for Worst-case Generalization Risk)

Introducing a virtual dataset $S' = \{X'_1, \cdots, X'_n\}$ including n new samples independent from dataset S and denote $L(f) = \mathbb{E}_{S'}[\hat{L}_{S'}(f)]$. The expected worst-case generalization risk can be bounded as

$$\mathbb{E}[G_n] \leq \mathbb{E}_{S,S'}\left[\sup_{f \in \mathcal{F}} \hat{L}_{S'}(f) - \hat{L}_S(f)\right] \quad \textit{(Symmetrization Bound)}$$

$$\leq 2\mathbb{E}_{S,\boldsymbol{\sigma}}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \ell(f(\mathbf{x}_i, y_i)\right] = 2R_n(\mathcal{H}).$$

Proof

$$\begin{aligned}
\mathbb{E}[G_n] &= \mathbb{E}_S\left[\sup_{f\in\mathcal{F}} L(f) - \hat{L}_S(f)\right] \\
&= \mathbb{E}_S\left[\sup_{f\in\mathcal{F}} \mathbb{E}_{S'}[\hat{L}_{S'}(f)] - \hat{L}_S(f)\right] \\
&= \mathbb{E}_S\left[\sup_{f\in\mathcal{F}} \mathbb{E}_{S'}\left[\hat{L}_{S'}(f) - \hat{L}_S(f)\right]\right] \\
&\leq \mathbb{E}_S\left[\mathbb{E}_{S'}\left[\sup_{f\in\mathcal{F}} \hat{L}_{S'}(f) - \hat{L}_S(f)\right]\right] \\
&= \mathbb{E}_{S,S'}\left[\sup_{f\in\mathcal{F}} \hat{L}_{S'}(f) - \hat{L}_S(f)\right],
\end{aligned}$$

where the inequality holds due to Jensen's inequality.

$$\begin{aligned}
\mathbb{E}_{S,S'}\left[\sup_{f\in\mathcal{F}} \hat{L}_{S'}(f) - \hat{L}_S(f)\right] &= \mathbb{E}_{S,S'}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}[X_i' - X_i]\right] \\
&= \mathbb{E}_{S,S',\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}[\sigma_i X_i' - \sigma_i X_i]\right] \quad \text{(Property of Rademacher r.v.)} \\
&\leq \mathbb{E}_{S',\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i X_i'\right] + \mathbb{E}_{S,\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} -\sigma_i X_i\right] \quad \text{(Why?)} \\
&= 2\mathbb{E}_{S,\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i X_i\right]
\end{aligned}$$

■

Lemma 5.5

Let P be a probability distribution over a domain space X. The Rademacher complexity of the function class $\mathcal{F}$ w.r.t. P for i.i.d. sample $S = (x_1, \cdots, x_n)$ with size n is $R_n(\mathcal{F})$. We have

$$\mathbb{E}_{S\sim P^n}[\sup_{f\in\mathcal{F}}(\mathbb{E}_{x\sim P}[f(x)] - \frac{1}{n}\sum_{i=1}^{n} f(x_i))] \leq 2R_n(\mathcal{F}).$$

Proof Construct another independent sample $S' = \{x_1', \cdots, x_n'\}$, we have Why?

$$\mathbb{E}_{x\sim P}[f(x)] = \mathbb{E}_{S'\sim P^n}[\frac{1}{n}\sum_{i=1}^{n} f(x_i')].$$

TBD. ■

Note on *This theorem shows that one can bound the maximum error in estimating the mean of any function f using the Rademacher complexity of the set of functions $\mathcal{F}$.*

Theorem 5.3 (Excess Risk Bound via Rademacher Complexity(Chen, 2013; Liao, 2020))

For a hypothesis set $\mathcal{F}$, define $\mathcal{H} = \{[x, y] \to \ell(f(x), y) : f \in \mathcal{F}\}$ to be the composition of loss function ℓ with the hypotheses in $\mathcal{F}$. If $|f(x) - f(y)| \leq c$,

$$\mathbb{P}(\mathbb{E}[f(x)] - \frac{1}{n}\sum_{i=1}^{n} f(x_i) \geq 2R_n(\mathcal{H}) + \epsilon) \leq \exp(-2n\epsilon^2/c^2),$$

$$\mathbb{P}(\mathbb{E}[f(x)] - \frac{1}{n}\sum_{i=1}^{n} f(x_i) \geq 2\hat{R}_n(\mathcal{H}) + 3\epsilon) \leq 2\exp(-2n\epsilon^2/c^2).$$

Proof We denote event C as

$$\hat{R}_n(\mathcal{H}) \geq R_n(\mathcal{H}) - \epsilon.$$

And from Lemma 5.3.2, we know that $\mathbb{P} \geq 1 - \exp(-2n\epsilon^2/c^2)$. ■

Corollary 5.5 (Excess Risk Bound via Rademacher Complexity given 0-1 loss(Farnia, 2023))

Particularly, under 0-1 loss functions, with probability at least $1 - \delta$,

$$L(\hat{f}) - L(f^*) \leq 4R_n(\mathcal{H}) + \sqrt{\frac{2\log(2/\delta)}{n}},$$

$$L(\hat{f}) - L(f^*) \leq 4\hat{R}_n(\mathcal{H}) + \sqrt{\frac{50\log(4/\delta)}{n}}.$$

Proof

$$\begin{aligned}
\mathbb{P}(L(\hat{f}) - L(f^*) \geq \epsilon) &\leq \mathbb{P}(\sup_{f\in\mathcal{F}} |L(f) - \hat{L}(f)| \geq \frac{\epsilon}{2}) \\
&\leq \mathbb{P}(G_n \geq \frac{\epsilon}{2}) + \mathbb{P}(G'_n \geq \frac{\epsilon}{2}) \\
&= \mathbb{P}(G_n - \mathbb{E}[G_n] \geq \frac{\epsilon}{2} - \mathbb{E}[G_n]) + \mathbb{P}(G'_n - \mathbb{E}[G'_n] \geq \frac{\epsilon}{2} - \mathbb{E}[G'_n]) \\
&\leq \exp(-2n(\frac{\epsilon}{2} - \mathbb{E}[G_n])^2) + \exp(-2n(\frac{\epsilon}{2} - \mathbb{E}[G'_n])^2) \quad \text{(McDiarmid's Inequality)} \\
&\leq 2\exp(-2n(\frac{\epsilon}{2} - 2R_n(\mathcal{H}))^2) \quad \text{(Lemma 5.3.1 + Lemma 5.4)}
\end{aligned}$$

where G'_n is defined for the negative loss. If we define $\delta = 2\exp(-2n(\frac{\epsilon}{2} - 2R_n(\mathcal{H}))^2)$, then we have $\epsilon = 4R_n(\mathcal{H}) + \sqrt{\frac{2\log(2/\delta)}{n}}$, i.e.,

$$\mathbb{P}(L(\hat{f}) - L(f^*) \leq 4R_n(\mathcal{H}) + \sqrt{\frac{2\log(2/\delta)}{n}}) \leq \delta.$$

■

5.5 VC Dimension

Definition 5.12 (Shattering Coefficient)

Given a function set $\mathcal{F}$ whose members map a feature vector $\mathbf{x} \in \mathcal{X}$ to $\mathcal{Y} = \{0, 1\}$, we define shattering coefficient $s(\mathcal{F}, n)$ as the maximum number of different label assignment

over datasets of size n, i.e., $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\} \in \mathcal{X}^n$,

$$s(\mathcal{F}, n) := \max_{\mathbf{x}_1, \cdots, \mathbf{x}_n \in \mathcal{X}} card(\{[f(\mathbf{x}_1), \cdots, f(\mathbf{x}_n)] : f \in \mathcal{F}\}),$$

where $card$ denotes cardinality of the set, i.e., the number of elements in the set.

Property 5.1 (Properties of Shattering Coefficient ("Shattered Set" 2021))

- *$s(\mathcal{F}, n) \leq 2^n$ for all n, note that 2^n is the largest cardinality for a set with n elements in $\{0, 1\}$.*
- *If $s(\mathcal{F}, n) = 2^n$, that means there is a set of cardinality n, which can be shattered by $\mathcal{F}$*
- *If $s(\mathcal{F}, N) < 2^N$ for some $N > 1$, then $s(\mathcal{F}, n) < 2^n$ for all $n \geq N$*

Corollary 5.6 (Massart Lemma applied to Shattering Coefficient)

Consider a hypothesis set $\mathcal{F}$ with boolean output in $\mathcal{Y} = \{0, 1\}$. Then for every dataset S of size n, we have

$$\hat{R}_n(\mathcal{F}) \leq \sqrt{\frac{2 \log s(\mathcal{F}, n)}{n}}.$$

Proof The proof is based on Lemma 5.3.2. Since $\mathcal{Y} = \{0, 1\}$, $M = 1$. The logic is a little weird here, this may comes from different expressions of theorems. ■

Definition 5.13 (VC Dimension)

Consider a hypothesis set $\mathcal{F}$ with boolean output in $\mathcal{Y} = \{0, 1\}$. Its VC dimension $VC(\mathcal{F})$ is defined as the size n of the largest dataset S that can be shattered by $\mathcal{F}$:

$$VC(\mathcal{F}) = \sup\{n : s(\mathcal{F}, n) = 2^n\}.$$

Note on Examples

- *The class of one dimensional half spaces $\mathcal{A}_1 = \{(-\infty, a] | a \in \mathbb{R}\}$ has $s(\mathcal{F}, n) = n + 1$, and so $VC(\mathcal{F}) = 1$ (Martin Wainwright, 2009).*
- *The class of half open intervals $\mathcal{F} = \{x \to \mathbf{1}(x \in (b, a]) : b < a \in \mathbb{R}\}$ has $s(\mathcal{F}, n) = \frac{n(n+1)}{2} + 1$. To see this, WLOG, suppose $x_1 \leq \cdots \leq x_n$. When $n = 1$, the outcomes can be $\{1, 0\}$ since we can find an interval based on $b < a \in \mathbb{R}$ including and excluding x_1. When $n = 2$, the outcomes can be $\{11, 10, 01, 00\}$. When $n = 3$, the outcomes does not include 101. Since if $(b, a]$ include x_1 and x_3, it must include x_2 too. So $VC(\mathcal{F}) = 2$ (Martin Wainwright, 2009).*

Lemma 5.6 (VC Dimension of Finite-dimensional Vector Space (Martin Wainwright, 2009))

Let $\mathcal{G}$ be a finite-dimensional vector space of functions on $\mathbb{R}^d$. Then the class of sets

$$\mathcal{A}_\mathcal{G} = \{x : g(x) \geq 0, g \in \mathcal{G}\}$$

has VC dimension at most $dim\mathcal{G}$.

Property 5.2 (VC Dimension of d-dimension Hyperplanes)

Consider the set of d-dimension linear classification rules:

$$\mathcal{F} = \{\mathbf{1}(\mathbf{w}^T\mathbf{x} \geq 0) : \mathbf{w} \in \mathbb{R}^d\}.$$

Then, the VC dimension will be $VC(\mathcal{F}) = d$.

Lemma 5.7 (Sauer's Lemma)

Consider a function set $\mathcal{F}$ with VC dimension $VC(\mathcal{F}) = d$. Then, for every integer $n \in \mathbb{N}$, we have

$$s(\mathcal{F}, n) \leq \sum_{i=1}^{d} C_n^i \leq (\frac{ne}{d})^d.$$

Note on

- *Case 1 $d \geq n$: we have $\sum_{i=1}^{d} C_n^i = 2^n$*
- *Case 2 $d < n$: we have $\sum_{i=1}^{d} C_n^i \leq (\frac{ne}{d})^d$, which is derived by*

$$\begin{aligned}
(\frac{d}{n})^d s(\mathcal{F}, n) &\leq \sum_{i=1}^{d} C_n^i (\frac{d}{n})^d \\
&\leq \sum_{i=1}^{d} C_n^i (\frac{d}{n})^i \\
&= \sum_{i=1}^{d} C_n^i (\frac{d}{n})^i 1^{n-i} \\
&\leq \sum_{i=1}^{n} C_n^i (\frac{d}{n})^i 1^{n-i} \\
&= ((1 + \frac{1}{d/n})^{n/d})^d \leq e^d
\end{aligned}$$

Theorem 5.4 (Rademacher Complexity Bound with VC-dimension)

Consider a hypothesis set $\mathcal{F}$ with boolean output in $\mathcal{Y} = \{0, 1\}$, whose VC dimension is $VC(\mathcal{F}) = d$. Then for every dataset S of size n, we have

$$\hat{R}_n(\mathcal{F}) \leq \sqrt{\frac{2d(\log(n/d) + 1)}{n}}.$$

Proof A direct result of Sauer's lemma. ■

Corollary 5.7 (Excess Risk Bound for 0-1 loss via VC-dimension)

Consider a hypothesis set $\mathcal{F}$ with boolean output which has $VC(\mathcal{F}) = d$. Suppose the loss function is the 0-1 loss. Then, with probability at least $1 - \delta$,

$$L(\hat{f}) - L(f^*) \leq \sqrt{\frac{32d(\log(n/d) + 1)}{n}} + \sqrt{\frac{2\log(2/\delta)}{n}}.$$

Proof A direct result of Theorem 5.5 and Corollary 5.4. ■

6 Kernel Methods (Mairal and Vert, 2020; Francis Bach, 2021)

In addition to minimize the excess risk, we also want to minimize the approximation error:

$$L(\hat{f}) = \underbrace{L(\hat{f}_n) - L(f^*)}_{\text{Excess Risk}} + \underbrace{L(f^*)}_{\text{Approximation Error}} .$$

Moreover, we also want to extend the learning theory from linear models to non-linear models, which relies on the kernel method enabling us to transfer the input to a potentially high-dimension space $m \gg d$.

6.1 Kernels

Definition 6.1 (Feature map)

The feature map $\phi : \mathbb{R}^d \to \mathbb{R}^m$ used in kernel methods maps $\mathbf{X} \in \mathbb{R}^{\mathbf{d}}$ to an often high-dimensional space $m \gg d$.

Note on Examples

- *Degree-m scalar polynomial:* $\phi(x) = [x, x^2, \cdots, x^m]$.
- *Counting features for a string:* $\phi(x) = [\#(\text{substring } s(i) \text{ in } x)]$

Note on Additional costs *Suppose we have a linear model, e.g.,*

$$\min_{\boldsymbol{w} \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \left(\boldsymbol{w}^\top \boldsymbol{x}_i - y_i \right)^2 .$$

Then the updated model will be

$$\min_{\boldsymbol{w} \in \mathbb{R}^m} \frac{1}{n} \sum_{i=1}^{n} \left(\boldsymbol{w}^\top \phi(\boldsymbol{x}_i) - y_i \right)^2 .$$

Obviously, including m decision variables brings more computational costs.

Definition 6.2 (Positive Semi-definite Kernels)

A function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is defined as a kernel function, if for every integer $t \in \mathbb{N}$ and vectors $\mathbf{x}_1, \cdots, \mathbf{x}_t$, the matrix $K \in \mathbb{R}^{t \times t}$ with the (i, j)-entry $k(\mathbf{x}_i, \mathbf{x}_j)$ will be positive semi-definite:

$$K := \begin{bmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & \cdots & k(\mathbf{x}_1, \mathbf{x}_n) \\ . & \cdots & . \\ k(\mathbf{x}_n, \mathbf{x}_1) & \cdots & k(\mathbf{x}_n, \mathbf{x}_n) \end{bmatrix} \succeq 0.$$

Lemma 6.1

$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel function iff there exists a feature map $\phi : \mathbb{R}^d \to \mathbb{R}^m$ such that for every $\boldsymbol{x}, \boldsymbol{x}'$:

$$k(\boldsymbol{x}, \boldsymbol{x}') = \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{x}') \rangle.$$

Note on Example

1. *Linear kernel:* $k(\boldsymbol{x}, \boldsymbol{x}') = \langle \boldsymbol{x}, \boldsymbol{x}' \rangle = \boldsymbol{x}^\top \boldsymbol{x}'$

2. *Degree-r polynomial kernel:* $k(\boldsymbol{x}, \boldsymbol{x}') = (1 + \langle \boldsymbol{x}, \boldsymbol{x}' \rangle)^r$. *Since* $1 + \langle \boldsymbol{x}, \boldsymbol{x}' \rangle$ *is a valid kernel functions (positive semi-definite), and so does the degree-r polynomial kernel via Corollary 6.1.*
3. *Gaussian kernel with bandwidth parameter* σ*:* $k(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\frac{||\boldsymbol{x}-\boldsymbol{x}'||_2^2}{2\sigma^2})$. *This actually can be rewritten as*

$$k(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\frac{||\boldsymbol{x}||_2^2}{2\sigma^2}) \exp(-\frac{||\boldsymbol{x}'||_2^2}{2\sigma^2}) \exp(-\frac{||\langle \boldsymbol{x}, \boldsymbol{x}' \rangle||_2^2}{2\sigma^2}).$$

And $\exp(-\frac{||\boldsymbol{x}||_2^2}{2\sigma^2}) \exp(-\frac{||\boldsymbol{x}'||_2^2}{2\sigma^2})$ *satisfies the properties of a kernel function via Corollary 6.1. Moreover,* $\exp(-\frac{||\langle \boldsymbol{x}, \boldsymbol{x}' \rangle||_2^2}{2\sigma^2}) = \sum_{t=0}^{\infty} \frac{\langle \boldsymbol{x}, \boldsymbol{x}' \rangle^t}{t! \sigma^{2t}}$ *is the sum of the powers of linear kernel and a kernel function too.*

Property 6.1 (Sum and Product of Kernels)

- *If k_1 and k_2 are two valid kernel functions, then their sum, i.e., $k = k_1 + k_2$, will also be a kernel function.*
- *If k_1 and k_2 are two valid kernel functions, then their product, i.e., $k = k_1 \times k_2$, will also be a kernel function.*

6.2 Reproducing Kernel Hilbert Space

Definition 6.3 (Reproducing Kernel Hilbert Space (RKHS))

For kernel function k, we define the reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ as the following set of functions

$$\mathcal{H} = \left\{ f(\boldsymbol{x}) = \sum_{i=1}^{t} \alpha_i k(\boldsymbol{x}_i, \boldsymbol{x}) : t \in \mathbb{N}, \alpha_1, \cdots, \alpha_t \in \mathbb{R}, \boldsymbol{x}_1, \cdots, \boldsymbol{x}_t \in \mathcal{X} \right\}.$$

Definition 6.4 (Inner Product in an RKHS)

Given two functions $f(\boldsymbol{x}) = \sum_{i=1}^{t} \alpha_i k(\boldsymbol{x}_i, \boldsymbol{x})$, $g(\boldsymbol{x}) = \sum_{j=1}^{t'} \beta_j k(\boldsymbol{x}'_j, \boldsymbol{x})$ *in the RKHS $\mathcal{H}$ for kernel-k, we define their inner product as*

$$\langle f, g \rangle = \sum_{i=1}^{t} \sum_{j=1}^{t'} \alpha_i \beta_j k(\boldsymbol{x}_i, \boldsymbol{x}'_j).$$

Property 6.2 (Inner Product's Property in an RKHS)

The RKHS defined for a kernel k coupled with the above inner product will result in a Hilbert Space, where the followings hold for every $f, f_1, f_2, g \in \mathcal{H}$:

1. ***Symmetry.*** $\langle f, g \rangle = \langle g, f \rangle$
2. ***Linearity.*** *For all* $\gamma \in \mathbb{R}$: $\langle f_1 + \gamma f_2, g \rangle = \langle f_1, g \rangle + \gamma \langle f_2, g \rangle$
3. ***Positive definiteness.*** $\langle f, f \rangle \geq 0$ *where the equality holds only for* $f = 0$.

Definition 6.5 (Norm in an RKHS)

For a function $f(\boldsymbol{x}) = \sum_{i=1}^{t} \alpha_i k(\boldsymbol{x}_i, \boldsymbol{x})$ in the RKHS $\mathcal{H}$, we define its norm as

$$||f||_{\mathcal{H}} = \langle f, f \rangle = \sum_{i=1}^{t} \sum_{j=1}^{t} \alpha_i \alpha_j k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{\alpha}^\top K \boldsymbol{\alpha},$$

where $\boldsymbol{K} \in \mathbb{R}^{t \times t}$ has $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ as (i, j)-entry.

6.3 Shift-invariant Kernels and Bochner's Theorem

Definition 6.6 (Shift-invariant Kernel)

A kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a shift-invariant kernel if there exists a function $\kappa : \mathcal{X} \to \mathbb{R}$ such that for every $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$:

$$k(\boldsymbol{x}, \boldsymbol{x}') = \kappa(\boldsymbol{x} - \boldsymbol{x}').$$

Note on Examples *Linear and polynomial kernels are not shift-invariant Why?, but Gaussian kernel is shift-invariant.*

Lemma 6.2 (Euler's formula)

$$e^{i\theta} = \cos\theta + i \sin\theta$$

Lemma 6.3 (Inner product and Norm of Complex Numbers (mb-, 2016))

The standard inner product of two complex numbers $z_1, z_2 \in \mathbb{C}$ is $z_1 \overline{z}_2$. And the norm induced for $z = a + bi$ is

$$||z|| = \sqrt{z_1 \overline{z}_1} = \sqrt{a^2 + b^2}.$$

Definition 6.7 (Fourier Transform)

Given a function $f : \mathbb{R}^d \to \mathbb{R}$, we define its Fourier transform $\hat{f} : \mathbb{R}^d \to \mathbb{R}$ as

$$\hat{f}(\boldsymbol{w}) = \int f(\boldsymbol{x}) \exp(-i \boldsymbol{w}^\top \boldsymbol{x}) d\boldsymbol{x}.$$

Note on Examples

1. ***Double-sided exponential (Pauly, 2001).*** *$f(t) = e^{-a|t|}$ with $a > 0$*

$$\begin{aligned} F(w) &= \int_{-\infty}^{0} \exp(at) \exp(-iwt) dt + \int_{0}^{\infty} \exp(-at) \exp(-iwt) dt \\ &= \frac{1}{a - iw} + \frac{1}{a + iw} = \frac{2a}{a^2 + w^2} \end{aligned}$$

2. ***Gaussian-shape function (Bagla, 2020).*** $\kappa(\boldsymbol{x}) = \exp(-\frac{||\boldsymbol{x}||_2^2}{2\sigma^2})$

$$
\begin{aligned}
F(w) &= \int \exp(-\frac{\sum_{j=1}^d x_j^2}{2\sigma^2}) \exp(-i\sum_{j=1}^d w_j x_j) dx \\
&= \prod_{j=1}^d \int \exp(-\frac{x_j^2}{2\sigma^2}) \exp(-iw_j x_j) dx_j \\
&= \prod_{j=1}^d \exp(-\frac{w_j^2\sigma^2}{2}) \int \exp(-(\frac{1}{\sqrt{2}\sigma}x_j + \frac{iw_j\sigma}{\sqrt{2}})^2) dx_j \\
&= \prod_{j=1}^d \exp(-\frac{w_j^2\sigma^2}{2}) \sqrt{2}\sigma \int \exp(-(\frac{1}{\sqrt{2}\sigma}x_j + \frac{iw_j\sigma}{\sqrt{2}})^2) d(\frac{1}{\sqrt{2}\sigma}x_j + \frac{iw_j\sigma}{\sqrt{2}}) \\
&= \prod_{j=1}^d \exp(-\frac{w_j^2\sigma^2}{2}) \sqrt{2}\sigma\sqrt{\pi} \quad (\int \exp(-x^2)dx = \sqrt{\pi}) \\
&= (\sqrt{2\pi}\sigma)^d \exp(-\frac{\sigma^2||w||_2^2}{2})
\end{aligned}
$$

Property 6.3 (Properties of Fourier Transform)

1. ***Symthesis (Inverse Fourier transform).***
$$f(x) = \frac{1}{(2\pi)^d} \int \hat{f}(w) \exp(iw^\top x) dw$$
2. ***Linearity.*** *For any f_1, f_2, α, β, we have*
$$\alpha f_1 \hat{+} \beta f_2 = \alpha \hat{f}_1 + \beta \hat{f}_2$$
3. ***Convolution property.*** *Denote the convolution as $f * g(z) = \int f(x)g(z-x)dx$, then we have*
$$f \hat{*} g = \hat{f}\hat{g}$$

Note on Examples

- *Find $F^{-1}(\frac{1}{(9+\lambda^2)(4+\lambda^2)})$ (Bagla, 2020).*

$$
\begin{aligned}
F^{-1}(\frac{1}{(9+\lambda^2)(4+\lambda^2)}) &= \frac{1}{5}F^{-1}(\frac{1}{2^2+\lambda^2} - \frac{1}{3^2+\lambda^2}) \\
&= \frac{1}{20}F^{-1}(\frac{4}{4+\lambda^2}) - \frac{1}{30}F^{-1}(\frac{6}{9+\lambda^2}) \quad (\textit{Linearity}) \\
&= \frac{1}{20}\exp(-2|x|) - \frac{1}{30}\exp(-3|x|) \quad (F(\exp(-a|x|)) = \frac{2a}{a^2+\lambda^2})
\end{aligned}
$$

- *Using convolution property, find $F^{-1}(\frac{1}{12-7i\lambda-\lambda^2})$ (Bagla, 2020).*

$$F^{-1}(\frac{1}{12-7i\lambda-\lambda^2}) = F^{-1}(\frac{1}{4-i\lambda}\frac{1}{3-i\lambda}) = F^{-1}(\frac{1}{4-i\lambda})F^{-1}(\frac{1}{3-i\lambda})$$

Theorem 6.1 (Bochner's Theorem)

A function $\kappa : \mathbb{R}^d \to \mathbb{R}$ results in a valid shift-invariant kernel $k(\boldsymbol{x}, \boldsymbol{x}') = \kappa(\boldsymbol{x} - \boldsymbol{x}')$ iff the Fourier transform $\hat{\kappa}$ is non-negative everywhere:

$$\forall \boldsymbol{w} \in \mathbb{R}^d : \hat{\kappa}(\boldsymbol{w}) \geq 0.$$

Note on Examples

1. *The gaussian-shape function* $\kappa(\boldsymbol{x}) = \exp(-\frac{||\boldsymbol{x}||_2^2}{2\sigma^2})$ *is a kernel function, since* $\hat{\kappa}(\boldsymbol{w}) = (\sqrt{2\pi}\sigma)^d \exp(-\frac{\sigma^2||w||_2^2}{2}) \geq 0$.
2. *The sinc function* $sinc(x) = \frac{\sin(\pi x)}{\pi x}$ *is a kernel function, since*
3. *The box function* $b(x) = \mathbf{1}[-1 \leq x \leq 1]$ *is not a kernel function, because*

6.4 Representer Theorem

Theorem 6.2 (Representer Theorem (Simple Version))

Consider solving the ERM problem with the feature function ϕ *using the gradient descent algorithm with stepsize* γ*, the* $k+1$*th gradient descent update will be*

$$\begin{aligned}\boldsymbol{w}^{(k+1)} &= \boldsymbol{w}^{(k)} - \gamma \nabla_{\boldsymbol{w}} \hat{L}(\boldsymbol{w}^k) \\ &= \boldsymbol{w}^{(k)} - \frac{2\gamma}{n}\sum_{i=1}^{n}(\boldsymbol{w}^{(k)^\top} - \phi(\boldsymbol{x}_i) - y_i)\phi(\boldsymbol{x}_i).\end{aligned}$$

Then, for every k *the* k*th update of the gradient descent method initialized at* $\mathbf{w}^{(0)} = \mathbf{0}$ *will satisfy the following for some scalars* $\alpha_1^{(k)}, \alpha_n^{(k)} \in \mathbb{R}$*:*

$$\mathbf{w}^{(k)} = \sum_{i=1}^{n} \alpha_i^{(k)} \phi(\mathbf{x}_i).$$

Note on *In other words, the above result shows that* $\boldsymbol{w}^{(k)}$ *is a linear combination of* $\phi(\boldsymbol{x}_1), \cdots, \phi(\boldsymbol{x}_n)$ *at every iteration. When* $m \leq n$*, this result is trivial since the span of* n*-dimensional vectors could cover the* $\mathbb{R}^m$ *space. Therefore, if* $m > n$ *then this result will be non-trivial.*

Note on Kernel Trick based on Representer Theorem *If* $\mathbf{w} = \sum_{i=1}^{n} \alpha_i \phi(\mathbf{x}_i)$*, then for every* $\boldsymbol{x}$ *we have*

$$\boldsymbol{w}^\top \phi(\boldsymbol{x}) = \sum_{j=1}^{n} \alpha_j \langle \phi(\boldsymbol{x}_j), \phi(\boldsymbol{x}) \rangle,$$

where $\langle \cdot, \cdot \rangle$ *denotes the standard inner product. Therefore, if we define the kernel function* $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ *as* $k(\boldsymbol{x}, \boldsymbol{x}') = \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{x}') \rangle$ *we can rewrite*

$$\boldsymbol{w}^\top \phi(\boldsymbol{x}) = \sum_{j=1}^{n} \alpha_j k(\boldsymbol{x}_j, \boldsymbol{x}).$$

That is, we rewrite the problem into the following optimization problem where $\boldsymbol{K} \in \mathbb{R}^{n \times n}$ *has* $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ *as* (i, j)*-entry:*

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{n} \alpha_j k(\boldsymbol{x}_j, \boldsymbol{x}) \right)^2 = \frac{1}{n} ||\boldsymbol{y} - \boldsymbol{K}\boldsymbol{\alpha}||_2^2.$$

The kernel trick is useful especially when $m \gg n$*, i.e., the complexity of ERM problem dominates sample size* n*. For example, if* $\phi(\boldsymbol{x})$ *covers all quadratic functions of vector* $\boldsymbol{x}$*, then* $m = d^2 + d$ *Why?. In general, for a degree-*r *polynomial, the number of variables will be* $m = O(d^r)$*. On the other hand, the equivalent optimization problem has only* n *variables, and*

as long as we can compute the $n \times n$ matrix $\boldsymbol{K}$ the complexity of the problem will be independent of m.

Theorem 6.3 (Representer Theorem)

Consider the following ERM problem over an RKHS $\mathcal{H}$ that corresponds to kernel k, and suppose Q is a strictly increasing function:

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(f(\boldsymbol{x}_i), y_i) + Q(||f||_{\mathcal{H}}).$$

Then, every optimal solution $f^* \in \mathcal{H}$ to the above problem satisfies the following for some real coefficients $\alpha_1, \cdots, \alpha_n \in \mathbb{R}$:

$$f^*(\boldsymbol{x}) = \sum_{j=1}^{n} \alpha_j k(\boldsymbol{x}_j, \boldsymbol{x}).$$

That is, this is equal to solve

$$\min_{\alpha \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^{n} \ell(\sum_{j=1}^{n} \alpha_j k(\boldsymbol{x}_j, \boldsymbol{x}_i), y_i) + Q(\sqrt{\boldsymbol{\alpha}^\top \boldsymbol{K} \boldsymbol{\alpha}}).$$

Note on Example of Ridge Regression *Ridge regression is the training of a linear model $f_{\boldsymbol{w}}(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$ via Least Squares with the additive L_2-norm-squared penalty*

$$\min_{\boldsymbol{w} \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{w}^\top \boldsymbol{x}_i - y_i)^2 + \lambda ||\boldsymbol{w}||_2^2.$$

Kernel ridge regression is the training of a kernel-based model $f \in \mathcal{H}$ with squared-error loss and the additive kernel-norm-squared penalty

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (f(\boldsymbol{x}_i) - y_i)^2 + \lambda ||f||_{\mathcal{H}}^2.$$

Representer theorem implies that there is a solution $f^*(\boldsymbol{x}) = \sum_{j=1}^{n} \alpha_j k(\boldsymbol{x}_j, \boldsymbol{x})$ to the problem:

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^{n} (\sum_{j=1}^{n} \alpha_j k(\boldsymbol{x}_j, \boldsymbol{x}) - y_i)^2 + \lambda \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(\boldsymbol{x}_j, \boldsymbol{x}_i).$$

If we denote kernel matrix $K = [k(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{n \times n}]$ and vector $\boldsymbol{y} = [y_1, \cdots, y_n]$, we have the following equivalent problem:

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^n} ||K\boldsymbol{\alpha} - \boldsymbol{y}||_2^2 + \lambda \boldsymbol{\alpha}^\top K \boldsymbol{\alpha}.$$

Note that this is a convex optimization problem Why?. FOC leads to

$$2K(K\boldsymbol{\alpha}^* - \boldsymbol{y}) + 2\lambda K \boldsymbol{\alpha}^* = 0 \to K(K + \lambda I)\boldsymbol{\alpha}^* = K\boldsymbol{y} \to \boldsymbol{\alpha}^* = (K + \lambda I)^{-1} \boldsymbol{y}.$$

Note on Example of SVM *Consider the SVM problem, which is a L_2-regularized ERM problem with the hinge loss $\ell_{hinge}(\hat{y}, y) = \max\{0, 1 - \hat{y}y\}$:*

$$\min_{\boldsymbol{w} \in \mathbb{R}^d} \sum_{i=1}^{n} \ell_{hinge}(\boldsymbol{w}^\top \boldsymbol{x}_i, y_i) + \lambda ||\boldsymbol{w}||_2^2.$$

Note that $\max\{0, 1 - z\} = \max_{0 \le \alpha \le 1} \alpha(1 - z)$, the problem can be rewritten as

$$\min_{\boldsymbol{w} \in \mathbb{R}^d} \max_{\boldsymbol{\alpha} \in [0,1]^n} \sum_{i=1}^{n} \alpha_i (1 - y_i \boldsymbol{w}^\top \boldsymbol{x}_i) + \lambda ||\boldsymbol{w}||_2^2.$$

The minmax theorem implies that we can swap min and max in the above problem to obtain the dual problem:

$$\max_{\boldsymbol{\alpha}\in[0,1]^n} \min_{\boldsymbol{w}\in\mathbb{R}^d} \sum_{i=1}^n \alpha_i(1 - y_i\boldsymbol{w}^\top \boldsymbol{x}_i) + \lambda||\boldsymbol{w}||_2^2.$$

Note that the objective function regarding $\boldsymbol{w}$ is convex, and achieves its minimum at $\boldsymbol{w}^* = \frac{1}{2\lambda}\sum_{i=1}^n \alpha_i y_i \boldsymbol{x}_i$ by FOC. The problem can be rewritten as

$$\begin{aligned}
&\max_{\boldsymbol{\alpha}\in[0,1]^n} (\sum_{i=1}^n \alpha_i) - \frac{2}{\lambda}||\sum_{i=1}^n \alpha_i y_i \boldsymbol{x}_i||_2^2 \\
=&\max_{\boldsymbol{\alpha}\in[0,1]^n} (\sum_{i=1}^n \alpha_i) - \frac{2}{\lambda}\sum_{i=1}^n\sum_{j=1}^n \alpha_i\alpha_j y_i y_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle
\end{aligned}.$$

In other words, if we define $K = [y_i y_j k(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{n\times n}$, the above problem is

$$\max_{\boldsymbol{\alpha}\in[0,1]^n} \mathbf{1}^\top\boldsymbol{\alpha} - \frac{2}{\lambda}\boldsymbol{\alpha}^\top K\boldsymbol{\alpha}.$$

6.5 Random Fourier Features

Definition 6.8 (Random Fourier Features)

If feature maps belong to high(or infinite)-dimenisonal spaces, it would be infeasible to use the map to compute the kernel function, i.e.,

$$k(\boldsymbol{x}, \boldsymbol{x}') = \langle\phi(\boldsymbol{x}), \phi(\boldsymbol{x}')\rangle.$$

However, for a shift-invariant kernel $k(\boldsymbol{x}, \boldsymbol{x}') = \kappa(\boldsymbol{x} - \boldsymbol{x}')$, we have

$$k(\boldsymbol{x}, \boldsymbol{x}') = \frac{1}{(2\pi)^d}\int \phi_w(\boldsymbol{x})\overline{\phi_w(\boldsymbol{x}')}\hat{\kappa(w)}dw,$$

where $\phi_w(\boldsymbol{x}) = \exp(iw^\top\boldsymbol{x})$. Assuming $\kappa(0) = 1$, then $\hat{\kappa}$ is a probability density function. And we can draw independent samples $w_1, \cdots, w_m$ according to $\hat{\kappa}$, the kernel function can be approximated as

$$\hat{k}(\boldsymbol{x}, \boldsymbol{x}') = \frac{1}{m}\sum_{i=1}^m \phi_{\boldsymbol{w}_i}(\boldsymbol{x})\overline{\phi_{\boldsymbol{w}_i}(\boldsymbol{x})}.$$

Then, we can use the random feature $\hat{\phi}(\boldsymbol{x}) = \frac{1}{\sqrt{m}}[\phi_{\boldsymbol{w}_1}(\boldsymbol{x}), \cdots, \phi_{\boldsymbol{w}_m}(\boldsymbol{x})]$ map to approximate the target shift-invariant kernel:

$$\hat{k}(\boldsymbol{x}, \boldsymbol{x}') = \langle\hat{\phi}(\boldsymbol{x}), \hat{\phi}(\boldsymbol{x}')\rangle.$$

Theorem 6.4 (Approximation Error of Random Fourier Features)

Suppose k is a shift-invariant kernel function. We consider a subset of the resulting RKHS $\mathcal{H}$ where the Fourier coefficients are bounded by C as

$$\mathcal{H}_C := \{\int \alpha(\boldsymbol{w})\hat{\kappa}(\boldsymbol{w})\phi_{\boldsymbol{w}}(x)d\boldsymbol{w} : |\alpha(\boldsymbol{w})| \le C, \forall\boldsymbol{w}\}.$$

Consider the norm $||\cdot||$ induced by a distribution q-based inner product $\langle f, g\rangle = \mathbb{E}_{x\sim q}[f(x)g(x)]$. Then, for every $f^* \in \mathcal{H}_C$ and set of samples $\boldsymbol{w}_1, \cdots, \boldsymbol{w}_k$ i.i.d. drawn

from $\hat{\kappa}$, with probability at least $1-\delta$ there exists coefficients $\alpha_1, \cdots, \alpha_m$ such that

$$||\frac{1}{m}\sum_{i=1}^{m}\alpha_i\phi_{\boldsymbol{w}_i} - f^*|| \leq \sqrt{\frac{C^2}{m}} + \sqrt{\frac{2\log(1/\delta)}{m}}.$$

7 Online Learning

7.1 Introduction to Online Learning

Motivation. Previously we assume that training data are drawn randomly from a fixed distribution and are coming as a batch of size n. However, this may not be true in real world. Firstly, the distribution generating the data may evolve as time goes by. Secondly, samples are coming in an online fashion instead of a fixed batch. In this section, we train the model over all the samples from time $t = 0$ till the current $t = T$.

In online learning, we suppose the learning task is formed as a game between the learner and nature players:

1. At every iteration t, Nature reveals the input $\boldsymbol{x}_t \in \mathcal{X}$ to the learner.
2. Learner outpus a prediction $p_t \in \mathcal{Y}$.
3. Nature reveals the true label $y_t \in \mathcal{Y}$, and Learner will suffer a loss $\ell(y_t, p_t)$.
4. Learner updates her prediction model.

Definition 7.1 (Cumulative Loss of an Online Learner)

One potential goal is to minimize the cumulative loss of the learner:

$$\textit{CumulativeLoss}\ (T) = \sum_{t=1}^{T}\ell(y_t, p_t).$$

Note on *The drawback of the cumulative loss-based evaluation is that if the nature player acts as an adversary, every learning algorithm will have the same evaluation score.*

Definition 7.2 (Regret of an Online Learner)

Given an expert $h : \mathcal{X} \to \mathcal{Y}$, the regret of the online learner is defined as the extra cumulative loss of the learner with respect to expert h:

$$\textit{Regret}\ (h) = \sum_{t=1}^{T}[\ell(y_t, p_t)] - \sum_{t=1}^{T}[\ell(y_t, h(\boldsymbol{x}_t))].$$

For a set of experts $\mathcal{H}$, we define the learner's regret as the worst-case regret for any expert $h \in \mathcal{H}$:

$$\begin{aligned}\textit{Regret}\ (\mathcal{H}) &= \max_{h\in\mathcal{H}} \textit{Regret}\ (\langle\rangle) \\ &= \sum_{t=1}^{T}[\ell(y_t, p_t)] - \min_{h\in\mathcal{H}}\sum_{t=1}^{T}[\ell(y_t, h(\boldsymbol{x}_t))]\end{aligned}.$$

Note on Examples

- *Consider a binary prediction task with $\mathcal{Y} = \{0, 1\}$ and an adversary nature which always generating the opposite label of the learner's prediction. The cumulative loss of every learning algorithms will be T at iteration T. Suppose $\mathcal{H} = \{h_0, h_1\}$ where $h_i(\boldsymbol{x}) = i$, i.e., h_i always outputs the assigned label. The regret with respect to $\mathcal{H}$ will be at least $\frac{T}{2}$. Note that*

$$\ell(y_t, 1) + \ell(y_t, 0) = 1,$$

i.e.,

$$\ell(y_t, h_1(\boldsymbol{x}_t)) + \ell(y_t, h_0(\boldsymbol{x}_t)) = 1.$$

This is equivalent to

$$\sum_{t=1}^{T}[\ell(y_t, h_1(\boldsymbol{x}_t))] + \sum_{t=1}^{T}[\ell(y_t, h_0(\boldsymbol{x}_t))] = 1,$$

and there must exists h_i such that $\sum_{t=1}^{T}[\ell(y_t, h_i(\boldsymbol{x}_t))] \leq \frac{T}{2}$.

- *Consider a binary prediction task with $\mathcal{Y} = \{0, 1\}$ and a realizable nature which generates the label according to an expert $h^* \in \mathcal{H} : Y = h^*(\boldsymbol{x})$. In the reliable case, the cumulative loss of a learning algorithm is equal to its regret with respect to $\mathcal{H}$. Why?*
  - *Consider the **follow the best expert algorithm**, where we arbitrarily choose among the experts with the best score up to iteration t. Then the regret with respect to $\mathcal{H}$ could be as large as $card(\mathcal{H}) - 1$.*
  - *Consider the **majority algorithm** where we vote for the label with the majority vote among experts in $\mathcal{H}$. Then the regret to $\mathcal{H}$ is bounded by $\log(card(\mathcal{H}))$.*

7.2 Online Convex Optimization and Online Linear Regression

In the online convex optimization setting,

1. Learner chooses model parameters $\boldsymbol{w}_t \in S$ from a convex set S at iteration t.
2. Nature chooses a convex loss function $f_t \in \mathcal{F}_{convex}$.
3. The regret with respect to model $\boldsymbol{u}$ will be

$$\text{Regret}(\boldsymbol{u}) = \sum_{t=1}^{T} f_t(\boldsymbol{w}_t) - \sum_{t=1}^{T} f_t(\boldsymbol{u}),$$

$$\text{Regret}(S) = \max_{\boldsymbol{u} \in S} \text{Regret}(\boldsymbol{u}) = \sum_{t=1}^{T} f_t(\boldsymbol{w}_t) - \min_{\boldsymbol{u} \in S} \sum_{t=1}^{T} f_t(\boldsymbol{u}).$$

Maximizing a concave function subject to a convex feasible set is a convex optimization problem.

Definition 7.3 (Online Linear Regression)

Consider the squared-error loss $\ell_2(\hat{y}, y) = (\hat{y} - y)^2$. The online linear regression setting is

- *The nature reveals input vector $\boldsymbol{x}_t \in \mathbb{R}^d$.*
- *The online learner chooses model parameters $\boldsymbol{w}_t \in \mathbb{R}^d$.*
- *The nature reveals output $y_t \in \mathbb{R}$, and the loss value at iteration t is*

$$f_t(\boldsymbol{w}_t) = \ell_2(\boldsymbol{w}_t^\top \boldsymbol{x}_t, y_t) = (\boldsymbol{w}_t^\top \boldsymbol{x}_t - y_t)^2.$$

And the regret function in this online learnign problem is

$$\textit{Regret}(\boldsymbol{u}) = \sum_{t=1}^{T}[f_t(\boldsymbol{w}_t) - f_t(\boldsymbol{u})] = \sum_{t=1}^{T}[(\boldsymbol{w}_t^\top \boldsymbol{x}_t - y_t)^2 - (\boldsymbol{u}^\top \boldsymbol{x}_t - y_t)^2].$$

Definition 7.4 (Online Convex Optimization with Expert Advice)

Consider a general lsos $\ell(\hat{y}, y)$ and a set of experts $\mathcal{H} = \{h_1, \cdots, h_m\}$. To convexify the problem, the learner searches for a probability distribution over the m experts which means $\boldsymbol{w}_t \in \Delta_m$ where Δ_m is the set of all probability distributions on $\mathcal{H} = \{h_1, \cdots, h_m\}$. The online learning task is as follows

1. *The nature reveals input vector $\boldsymbol{x}_t \in \mathbb{R}^d$.*
2. *The online learner chooses $\boldsymbol{w}_t \in \mathbb{R}^m$.*
3. *The nature reveals output $y_t \in \mathbb{R}$ and the loss value at iteration t is*

$$f_t(\boldsymbol{w}_t) = \boldsymbol{w}_t^\top L_t, L_t = [\ell(h_1(\boldsymbol{x}_t), y_t), \cdots, \ell(h_m(\boldsymbol{x}_t), y_t)].$$

The regret function in this online learning problem is

$$\textit{Regret}(\boldsymbol{u}) = \sum_{t=1}^{T}[f_t(\boldsymbol{w}_t) - f_t(\boldsymbol{u})] = \sum_{t=1}^{T}(\boldsymbol{w}_t - \boldsymbol{u})^\top L_t.$$

Definition 7.5 (Follow The Leader (FTL) Strategy)

Following the best-performing expert up to the current iteration.

Note on *In an online convex optimization task, the FTL problem is indeed a convex optimization problem.*

Lemma 7.1 (Regret Bound for FTL)

Given that $\boldsymbol{w}_t$ is chosen according to the FTL strategy, we have the following upper-bound on the FTL learner's regret at iteration T:

$$\textit{Regret}(S) \leq \sum_{t=1}^{T}[f_t(\boldsymbol{w}_t) - f_t(\boldsymbol{w}_{t+1})].$$

Proof

■

Note on Examples

Definition 7.6 (Follow The Regularized Leader (FTRL) Strategy)

Lemma 7.2 (Regret Bound for FTRL)

Proof

■

Corollary 7.1

Definition 7.7 (Online Gradient Descent (OGD) Algorithm)

Theorem 7.1 (Regret Bound for OGD Learner)

Definition 7.8 (Online Mirror Descent (OMD) Algorithm)

Bibliography

Bagla, Vandana (2020). *Engineering Mathematics*.

Braverman, Mark (2011). *COS597D: Information Theory in Computer Science*.

Chen, Jason (2013). *COS 511: Theoretical Machine Learning*.

Cuong (Jan. 2019). *Answer to "Confused about the Realizability Assumption and Equations of Upper Bound"*. (Visited on 03/11/2023).

Danica (Nov. 2018). *Answer to "How Is True Risk Equal to the Expected Value of the Empirical Risk?"* (Visited on 03/13/2023).

Farnia, Farzan (2023). *CSCI 5030 Machine Learning Theory*.

Francis Bach (2021). *Learning Theory from First Principles*.

halvorsen, kjetil b (Jan. 2016). *Answer to "Intuition on the Kullback–Leibler (KL) Divergence"*. (Visited on 02/25/2023).

Lacoste-Julien, Simon (2022). *IFT 6269 : Probabilistic Graphical Models*.

Liao, Renjie (2020). *Notes on Rademacher Complexity*.

Mairal, Julien and Jean-Philippe Vert (2020). *Machine Learning with Kernel Methods*.

Mao, Lei (Aug. 2019). *Cross Entropy, KL Divergence, and Maximum Likelihood Estimation*. https://leimao.github.io/blog/Cross-Entropy-KL-Divergence-MLE/. (Visited on 04/22/2023).

Martin Wainwright (2009). *STAT241B / EECS 281B Advanced Topics in Statistical Learning Theory*.

mb- (Aug. 2016). *Answer to "Definition of Absolute Value of a Complex Number"*. (Visited on 04/21/2023).

Meila, Marina (2012). *STAT 538 Statistical Learning: Modeling, Prediction and Computing*.

Ng, Andrew (2022). *CS229: Machine Learning*.

Nowak, Robert (2009). *ECE901 Summer '09: Statistical Learning Theory*.

Pauly, John (2001). *EE102: Signal Processing and Linear Systems I*.

Pillow, Jonathan (2018). *NEU560 Statistical Modeling and Analysis of Neural Data*.

Schramm, Tselil (2022). *STATS214/CS229M Machine Learning Theory*.

"Shattered Set" (July 2021). In: *Wikipedia*. (Visited on 04/26/2023).

Soch, Joram (Aug. 2020). *Convexity of the Kullback-Leibler Divergence*. https://statproofbook.github.io/P/kl-conv.html. (Visited on 04/22/2023).