CSC2532 Winter 2021
Statistical Learning Theory

Murat A. Erdogdu*

March 18, 2022

Lectures

*Department of Computer Science and Department of Statistical Sciences at University of Toronto, and Vector Institute erdogdu@cs.toronto.edu

0 Introduction

Machine learning (ML) is a set of algorithms/tools that learn and improve from data and/or past experience. It has many applications in areas such as computer vision, healthcare, physics, biology, etc. Since ML has become a cruical part of our daily life, it is important that we understand the principles that govern these algorithms. For this reason, practitioners generally use terms such as overfitting, underfitting, fast convergence, model complexity, do well on test data etc. to assess the performance of ML algorithms.

In this course, we will

- focus on formal definitions of the above concepts,
- and explain the behavior of ML algorithms using (mostly) probability theory, calculus, and linear algebra.

Example. Assume that you trained a binary classifier (e.g. logistic regression) on a dataset of n samples $\mathcal{D}_n^{(1)} = \{(y_i, x_i)\}_{i=1}^n$ where $x_i \in \mathbb{R}^d$ (i.e. you had d features), and this classifier achieved training and test errors of err_{train} and err_{test}, respectively.

- If you had another dataset $D_n^{(2)}$ and you train the same model on it, would you achieve the same err_{train} and err_{test}? Do you need them to be from the same distribution?
- What would happen to err_{train} and err_{test} if you combine datasets $D_n^{(1)}$ and $D_n^{(2)}$ and train your model?
- For each sample, assume that you collected d more features, i.e. $x_i \in \mathbb{R}^{2d}$ by keeping the number of samples n constant. How would this affect err_{train} and err_{test}?
- Instead of cross-entropy, you decided to use another loss function which has different smoothness properties. How would the smoothness of loss function affect err_{train} and err_{test}?

We will use theory to answer questions similar to the above ones. In general, the answers we get using theory can explain the observed behavior (or if it cannot, there is room for future research). The insights we gain from these can offer troubleshooting, or even suggest new ML algorithms. However, the theoretical guarantees we derive usually cannot tell if algorithm X is better than algorithm Y. They usually rely on generic tools and on assumptions that are violated in practice.

Our recipe in majority of this course will be

- assume a parametric model on data (data distribution),
- choose a suitable loss function,
- minimize the loss over training data (training error), and hope that you achieve small test error.

1 Warm-up: Gaussian Mean Estimation

Suppose we have i.i.d. random variables $x_1, x_2, \ldots, x_n \sim \mathcal{N}(\theta_*, \sigma^2 I)$ where $\theta_* \in \mathbb{R}^d$ is unknown and σ^2 is known. Our goal is to estimate θ_* with an estimator $\hat{\theta}$ such that, $d(\hat{\theta}, \theta_*) < \epsilon$ for some small ϵ, where $d : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ is some metric measuring the distance between $\hat{\theta}$ and θ_*. It is understood that $\hat{\theta}$ is a random variable whereas θ_* is deterministic.

There are many approaches that we can take to tackle this estimation problem. For example, we can use

- Sample mean estimator: $\hat{\theta} = \frac{1}{n}\sum_{i=1}^n x_i$;
- Maximum Likelihood Estimator (exercise: in fact it reduces to sample mean)
- Maximum A posteriori Probability under some prior on θ_*
- ...

Let's take a look at the sample mean estimator as given by $\hat{\theta} = \frac{1}{n}\sum_{i=1}^n x_i$, and find its performance.

Since x_i's are i.i.d. Gaussian random vectors, their linear combination is also Gaussian. One way to see this is by using the moment generating function (MGF) for Gaussian random vectors.

Lemma 1. *Given $Z_1 \sim \mathcal{N}(0, \Sigma_1), Z_2 \sim \mathcal{N}(0, \Sigma_2)$ independent random vectors, we have*

$$Z_1 + Z_2 \sim \mathcal{N}(0, \Sigma_1 + \Sigma_2).$$

Proof. Recall the definition of the MGF of a random variable X as $m_X(t) = \mathbb{E}[e^{\frac{1}{2}\langle t, X\rangle}]$. We have

$$m_{Z_1+Z_2}(t) = \mathbb{E}[e^{\frac{1}{2}\langle t, Z_1+Z_2\rangle}] = \mathbb{E}[e^{\frac{1}{2}\langle t, Z_1\rangle} e^{\frac{1}{2}\langle t, Z_2\rangle}] = \mathbb{E}[e^{\frac{1}{2}\langle t, Z_1\rangle}]\mathbb{E}[e^{\frac{1}{2}\langle t, Z_2\rangle}] = m_{Z_1}(t) m_{Z_2}(t)$$

by the independence of Z_1 and Z_2. Using the fact that the MGF for a Gaussian random variable $Z \sim \mathcal{N}(0, \Sigma)$ is $m_Z(t) = e^{\frac{1}{2}\langle t, \Sigma t\rangle}$, we have

$$m_{Z_1+Z_2}(t) = m_{Z_1}(t) m_{Z_2}(t) = e^{\frac{1}{2}\langle t, \Sigma_1 t\rangle} e^{\frac{1}{2}\langle t, \Sigma_2 t\rangle} = e^{\frac{1}{2}\langle t, (\Sigma_1+\Sigma_2) t\rangle}$$

which is the MGF of $\mathcal{N}(0, \Sigma_1 + \Sigma_2)$. Therefore, $Z_1 + Z_2 \sim \mathcal{N}(0, \Sigma_1 + \Sigma_2)$.

□

If we look at the difference between the sample mean estimator and the true mean, we have

$$\hat{\theta} - \theta_* = \frac{1}{n}\sum_{i=1}^n (x_i - \theta_*)$$

Since each $x_i - \theta_* \sim \mathcal{N}(0, \sigma^2 I)$, applying Lemma 1 iteratively, we obtain

$$\hat{\theta} - \theta_* \sim \mathcal{N}\left(0, \frac{\sigma^2}{n} I\right). \tag{1.1}$$

Definition 2. *We define the notion of loss and risk as follows.*

- ***Loss*** *measures the distance. We will denote it by $\ell : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$. For example, squared L_2-norm can be a loss $\ell(\theta, \theta') = \|\theta - \theta'\|^2$.*

- ***Risk*** *is the expected loss (so it is a population quantity). Risk between an estimator and true parameter*
$$R(\hat{\theta}, \theta_*) = \mathbb{E}[\ell(\hat{\theta}, \theta_*)].$$
Here, the expectation is over $\hat{\theta}$.

Next, let's choose the loss function as the squared L_2-norm, i.e., $\ell(\theta, \theta') = \|\theta - \theta'\|^2$. Then the risk function is given as $R(\hat{\theta}, \theta_*) = \mathbb{E}[\ell(\hat{\theta}, \theta_*)] = \mathbb{E}[\|\hat{\theta} - \theta_*\|^2]$. For the sample mean estimator, we have

$$\ell(\hat{\theta}, \theta_*) = \|\hat{\theta} - \theta_*\|^2, \quad \text{and} \quad R(\hat{\theta}, \mu) = \mathbb{E}[\|\hat{\theta} - \theta_*\|^2] = \frac{\sigma^2 d}{n}, \tag{1.2}$$

where in the last step we used (1.1). Note that the risk $R(\hat{\theta}, \theta_*)$ increases with dimension d and decreases with the number of samples n. This dependence structure is commonly observed for most loss minimization problems. This intuitively means that estimation is harder in higher dimensions, but gets better with more observations.

Remark. The loss $\ell(\hat{\theta}, \theta_*) \sim \chi^2_d$ where χ^2_d denotes the chi-square distribution.

One concern about this estimator is that $\mathbb{E}[||\hat{\theta}||^2] = ||\theta_*||^2 + \frac{\sigma^2 d}{n} > ||\theta_*||^2$. This means that the second moment of our estimator is always significantly larger than that of the true parameter we are estimating. To resolve this, we can simply multiply $\hat{\theta}$ by a factor $(1-\eta)$ to *shrink* it. This type of estimators called *shrinkage estimator*. In what follows, we show that MLE can be beaten.

1.1 SURE: Stein's Unbiased Risk Estimator

Lemma 3 (Stein's Lemma)**.** *Suppose* $x \sim \mathcal{N}(\mu, \sigma^2 I)$*, and* $g : \mathbb{R}^d \to \mathbb{R}^d$ *is weakly differentiable. Then*
$$\mathbb{E}[\langle x - \mu, g(x)\rangle] = \sigma^2 \mathbb{E}[\mathrm{Tr}(\nabla g(x))].$$

Remark. We are not giving a definition of *weak differentiablity*, but hereby we will assume g is differentiable which is a stronger assumption.

Proof. Let $\phi(x)$ denote the distribution of an isotropic Gaussian random vector. We can write

$$\mathbb{E}[\langle x - \mu, g(x)\rangle] = \int_{-\infty}^{\infty} \langle x - \mu, g(x)\rangle \phi(\tfrac{x-\mu}{\sigma})\, \mathrm{d}x.$$

Using the fact that
$$\mathrm{d}\phi(\tfrac{x-\mu}{\sigma}) = -\frac{x-\mu}{\sigma^2}\phi(\tfrac{x-\mu}{\sigma})\, \mathrm{d}x$$
and integration by parts, we have

$$\begin{aligned}\int_{-\infty}^{\infty} \langle x - \mu, g(x)\rangle \phi(\tfrac{x-\mu}{\sigma})\, \mathrm{d}x &= -\sigma^2 \int_{-\infty}^{\infty} \langle \mathrm{d}\phi(\tfrac{x-\mu}{\sigma}), g(x)\rangle \\ &= \sigma^2 \int_{-\infty}^{\infty} \phi(\tfrac{x-\mu}{\sigma})\, \mathrm{Tr}(\nabla g(x))\, \mathrm{d}x = \sigma^2 \mathbb{E}[\mathrm{Tr}(\nabla g(x))].\end{aligned}$$

□

Remark. The above results is also referred to as Stein's identity, and has remarkable applications ranging from probability theory (non-asymptotic CLTs) to machine learning (Stein's variational gradient descent) and optimization (Newton-Stein method, Scaled Least Squares).

In the following we will consider the risk of estimators of a particular form and show that MLE can be beaten in terms of risk. Let $\hat{\theta}^s$ be an estimator of the form

$$\hat{\theta}^s = \hat{\theta} + g(\hat{\theta}), \tag{1.3}$$

where $\hat{\theta}$ is the sample mean and $g : \mathbb{R}^d \to \mathbb{R}^d$ is any differentiable function. Then

$$\begin{aligned} R(\hat{\theta}^s, \theta_*) =& \mathbb{E}[\|\hat{\theta}^s - \theta_*\|^2] = \mathbb{E}[\|\hat{\theta} + g(\hat{\theta}) - \theta_*\|^2], \\ =& \mathbb{E}[\|\hat{\theta} - \theta_*||^2] + \mathbb{E}[\|g(\hat{\theta})\|^2] + 2\mathbb{E}[\langle \hat{\theta} - \theta_*, g(\hat{\theta})\rangle], \\ =& \frac{\sigma^2 d}{n} + \mathbb{E}[\|g(\hat{\theta})\|^2] + \frac{2\sigma^2}{n}\mathbb{E}[\mathrm{Tr}(\nabla g(\hat{\theta}))], \end{aligned} \tag{1.4}$$

where in the last step, we applied Stein's Lemma 3 on the last term.

This leads to the definition of the Stein's Unbiased Risk Estimator:

Definition 4 (SURE: Stein's Unbiased Risk Estimator)**.** *For an estimator of the form $\hat{\theta}^s = \hat{\theta} + g(\hat{\theta})$, we have the following unbiased estimator of the risk,*

$$SURE(\hat{\theta}) = \frac{\sigma^2 d}{n} + ||g(\hat{\theta})||^2 + \frac{2\sigma^2}{n}\,\mathrm{Tr}(\nabla g(\hat{\theta})).$$

The fact that $SURE(\hat{\mu})$ is an unbiased estimator for $R(\hat{\mu}^s, \mu)$ follows from (1.4). In other words, any estimator of the form (1.3), has the risk $\mathbb{E}[SURE(\hat{\mu})]$. Also note that the first term on the right hand side is the risk of MLE.

In the following, we will specify the function g in (1.3).

1.2 James-Stein Estimator

Definition 5 (James-Stein Estimator)**.** *Define the estimator*

$$\hat{\theta}^{js} = \left(1 - \frac{d-2}{||\hat{\theta}||^2}\frac{\sigma^2}{n}\right)\hat{\theta}.$$

The above estimator is of the form (1.3) with

$$g(x) = -\frac{\sigma^2}{n}\frac{d-2}{||x||^2}x, \quad \text{and} \quad \nabla g(x) = -\frac{\sigma^2}{n}\frac{d-2}{||x||^2}I + 2(d-2)\frac{\sigma^2}{n}\frac{xx^T}{||x||^4}.$$

This gives

$$\|g(x)\|^2 = \frac{\sigma^4}{n^2}\frac{(d-2)^2}{\|x\|^2} \quad \text{and} \quad \mathrm{Tr}(\nabla g(x)) = \frac{-d(d-2) + 2(d-2)}{||x||^2}\frac{\sigma^2}{n} = -\frac{(d-2)^2}{||x||^2}\frac{\sigma^2}{n}.$$

Therefore the risk of the James-Stein estimator is given as

$$\begin{aligned} R(\hat{\theta}^{js}, \theta_*) =& \frac{\sigma^2 d}{n} + \frac{\sigma^4}{n^2}\mathbb{E}\left[\frac{(d-2)^2}{\|\hat{\theta}\|^2}\right] - 2\frac{\sigma^4}{n^2}\mathbb{E}\left[\frac{(d-2)^2}{\|\hat{\theta}\|^2}\right] = \frac{\sigma^2 d}{n} - \frac{\sigma^4}{n^2}\mathbb{E}\left[\frac{(d-2)^2}{\|\hat{\theta}\|^2}\right] \\ <& R(\hat{\theta}, \theta_*), \end{aligned}$$

where the last step follows from $R(\hat{\theta}, \theta_*) = \sigma^2 d/n$ as derived in (1.2) Therefore, the James-Stein estimator is a strictly better estimator than the sample mean estimator based on the measure of the risk function. Note that this result holds for $d > 2$.

If we go one step further by applying Jensen's inequality ($x \to 1/x$ is convex for $x > 0$), we obtain

$$\mathbb{E}\left[\frac{1}{\|\hat{\theta}\|^2}\right] \geq \frac{1}{\mathbb{E}[\|\hat{\theta}\|^2]} = \frac{1}{\|\theta_*\|^2 + \frac{\sigma^2 d}{n}}.$$

Using this in the last step above, our bound for the risk of James-Stein estimator yields

$$\begin{aligned} R(\hat{\theta}^{js}, \theta_*) =& \frac{\sigma^2 d}{n} - \frac{\sigma^4}{n^2}\mathbb{E}[\frac{(d-2)^2}{\|\hat{\theta}\|^2}], \\ \leq& \frac{\sigma^2 d}{n} - \frac{\sigma^4}{n^2}\frac{(d-2)^2}{\|\theta_*\|^2 + \frac{\sigma^2 d}{n}}. \end{aligned}$$

Remark. A more careful treatment yields the following bound

$$R(\hat{\theta}^{js}, \theta_*) \leq \frac{\sigma^2 d}{n} - \frac{\sigma^4}{n^2}\frac{(d-2)^2}{\|\theta_*\|^2 + \frac{\sigma^2 (d-2)}{n}}.$$

- James-Stein is one the most significant advances in statistics.
- It shows that MLE can be beaten (inadmissable) for $d > 2$.
- This phenomenon is also known as Stein's paradox.

2 Exponential Families and Information Inequality

- Exponential families form a basis for many statistical methodology such as generalized linear models (GLMs), undirected graphical models, etc.
- They define a broad class of distributions covering distributions such as Gaussian, Bernouilli, beta, Poisson etc.
- They also arise as the solutions of interesting optimization problems.

Definition 6. *Exponential families are defined as a collection of densities with respect to a base measure ν (either counting or Lebesgue)*

$$\mathcal{P} = \{p_\theta(x) : \theta \in \Theta\} \quad \textit{where} \quad p_\theta(x) = \exp\{\langle \theta, \phi(x)\rangle - \psi(\theta)\}p_0(x).$$

Above,

- $\theta \in \Theta \subset \mathbb{R}^d$*: Natural parameter*
- $\phi : \mathcal{X} \to \mathbb{R}^d$*: Sufficient statistics*
- $\psi : \mathbb{R} \to \mathbb{R}$*: log-partition function, cumulant generating function (CGF)*
- $p_0(x)$*: carrying density w.r.t. carrying measure $\nu(\mathrm{d}x)$ on $\mathcal{X}$. We will ignore this part mostly as it can be combined with the carrying measure ν.*

The natural parameter θ lives in a parametric space where the CGF is finite: $\Theta = \{\theta : \psi(\theta) < \infty\}$. Since p_θ is a density, we have

$$1 = \int p_\theta(x)\,\mathrm{d}\nu(x) \quad \text{and} \quad \psi(\theta) = \log\left\{\int \exp\{\langle \theta, \phi(x)\rangle\}p_0(x)\,\mathrm{d}\nu(x)\right\}.$$

Note that in this class we only consider the measure $d\nu(x)$ either as the Lebesgue measure when the random variable is continuous or as the counting measure when it is discrete.

Example. Let X be a Bernoulli random variable with mean μ, i.e., $\mathbb{P}(X = 1) = \mu$ and $\mathbb{P}(X = 0) = 1 - \mu$. We can write the probability mass function as $p_\theta(x) = \mu^x(1-\mu)^{1-x} = \exp\{x \log \mu + (1-x)\log(1-\mu)\}$ where $x \in \{0, 1\}$. One way to write the Bernoulli distribution as an exponential family is through the following natural parameter and sufficient statistic

$$\theta = \begin{bmatrix} \log \mu \\ \log(1-\mu) \end{bmatrix}, \quad \phi(x) = \begin{bmatrix} x \\ 1-x \end{bmatrix}$$

We say that an exponential family is *minimal* if there is no linear relations/constraints between the entries of the sufficient statistic and the natural parameter vectors. Notice that the above formulation is not minimal. Re-write the PMF, natural parameter, and CGF:

$$p(x) = \exp\left\{x \log \frac{\mu}{1-\mu} + \log(1-\mu)\right\} \quad \text{with}$$
$$\theta = \log \frac{\mu}{1-\mu}, \quad \psi(\theta) = \log(1 + e^\theta), \quad \mu = \frac{e^\theta}{1+e^\theta}.$$

Proposition 7. *Θ is a convex set, and $\psi(\theta)$ is a convex function.*

Proof. Θ is a convex set if for $\theta_1, \theta_2 \in \Theta$, $\theta_\lambda = \lambda\theta_1 + (1-\lambda)\theta_2 \in \Theta$, $\forall \lambda \in [0,1]$.

$$\psi(\theta) < \infty \;\; \Leftrightarrow \;\; e^{\psi(\theta)} < \infty \;\; \Leftrightarrow \;\; \int \exp\{\langle \theta, \phi(x)\rangle\} d\nu(x) < \infty$$

$$\begin{aligned}
\exp\big(\psi(\theta_\lambda)\big) &= \int \exp\{\langle \theta_\lambda, \phi(x)\rangle\} d\nu(x) = \int \Big(e^{\langle \theta_1, \phi(x)\rangle}\Big)^{\lambda} \Big(e^{\langle \theta_2, \phi(x)\rangle}\Big)^{1-\lambda} d\nu(x) \\
&\leq \left(\exp\big(\psi(\theta_1)\big) \int p_{\theta_1}(x) d\nu(x)\right)^{\lambda} \left(\exp\big(\psi(\theta_2)\big) \int p_{\theta_2}(x) d\nu(x)\right)^{1-\lambda} \\
&= \exp\big(\psi(\theta_1)\big)^{\lambda} \exp\big(\psi(\theta_2)\big)^{1-\lambda} < \infty.
\end{aligned}$$

Where the inequality is justified above from Hölder's inequality for integrals: $\int |fg| du \leq (\int |f|^p)^{1/p} (\int |g|^q)^{1/q}$, $p^{-1} + q^{-1} = 1$. This completes the proof of first part. The second part follows applying logs to both sides,

$$\psi(\theta_\lambda) \leq \lambda\psi(\theta_1) + (1-\lambda)\psi(\theta_2).$$

□

2.1 Moments of exponential families

It can be shown that the moments of the sufficient statistic associated with an exponential family can be linked to the corresponding orders of differentiation of that family's CGF.

- **Mean:** We can write

$$\begin{aligned}
1 &= \int p_\theta(x)\, \mathrm{d}\nu(x) = \int e^{\langle \theta, \phi(x)\rangle - \psi(\theta)}\, \mathrm{d}\nu(x) \;\; \text{differentiating both sides w.r.t } \theta \\
0 &= \int e^{\langle \theta, \phi(x)\rangle - \psi(\theta)} \Big(\phi(x) - \nabla\psi(\theta)\Big) d\nu(x) \\
0 &= \mathbb{E}[\phi(x)] - \nabla\psi(\theta) \underbrace{\int e^{\langle \theta, \phi(x)\rangle - \psi(\theta)} d\nu(x)}_{1} \;\; \Leftrightarrow \;\; \mathbb{E}[\phi(x)] = \nabla\psi(\theta) \coloneqq \mu
\end{aligned}$$

- **Variance:** Taking one more derivative yields that $\mathrm{Cov}(\phi(x)) = \nabla^2\psi(\theta)$.
- **Higher-order moments:** Similarly, by taking more derivatives of the above equation, we can obtain higher-order moments.

Proposition 8 (Invertibility)**.** *If ψ is strictly convex, then $\nabla\psi : \Theta \to \mathcal{M}$ is invertible for $\mathcal{M} = \{\mu : \mu = \nabla\psi(\theta) \;\; \text{for} \;\; \theta \in \Theta\}$.*

Proof. We need to show that for $\theta_1, \theta_2 \in \Theta$,

$$\theta_1 = \theta_2 \Leftrightarrow \nabla\psi(\theta_1) = \nabla\psi(\theta_2).$$

One side is trivial. For the other side, we write

$$\nabla\psi(\theta_2) = \nabla\psi(\theta_1) + \int_0^1 \nabla^2\psi(\theta_1 + \tau(\theta_2 - \theta_1))(\theta_2 - \theta_1) d\tau.$$

Suppose it was the case that $\exists \theta_1, \theta_2$, $\theta_1 \neq \theta_2$ such that $\nabla\psi(\theta_1) = \nabla\psi(\theta_2)$, then it would be that $0 = \int_0^1 \nabla^2\psi(\theta_1 + \tau(\theta_2 + \theta_1))(\theta_2 - \theta_1)d\tau$. However, because ψ is strictly convex, $\nabla^2\psi > 0$, so the previous integral must be greater than zero and therefore $\nabla\psi(\theta_1) \neq \nabla\psi(\theta_2)$. □

Since the mapping $\nabla\psi : \Theta \to \mathcal{M}$ is invertible, we can write

1. $(\nabla\psi)^{-1}(\mu) = \theta$
2. $\Sigma = \nabla^2_\theta\psi(\theta) = \nabla_\theta\nabla_\theta\psi(\theta) = \nabla_\theta\mu$ or equivalently, $\Sigma = \frac{\mathrm{d}\mu}{\mathrm{d}\theta}$ where Σ is the covariance matrix of $\phi(X)$. Intuitively (from Leibniz notation), we have $\frac{\mathrm{d}\theta}{\mathrm{d}\mu} = \Sigma^{-1}$. This can be shown using chain rule.

$$\mu = \nabla_\eta\psi(\eta) \implies \frac{d\mu}{d\mu} = \frac{d\eta}{d\mu}\nabla^2_\eta\psi(\eta) \implies \frac{d\eta}{d\mu} = \Sigma^{-1}.$$

2.2 MLE, Score, Information

In this section, we consider the basic MLE setup where we assume $\mathbf{x} = [x_1, \ldots, x_n]$ where $x_i \overset{iid}{\sim} p_\theta(x)$. Using the iid assumption, we can write the joint density as

$$\begin{aligned} p_\theta(\mathbf{x}) = \prod_{i=1}^n p_\theta(x_i) &= \exp\left\{\left\langle \theta, \sum_{i=1}^n \phi(x_i)\right\rangle - n\psi(\theta)\right\}\prod_{i=1}^n p_0(x_i) \\ &= \exp\left\{n[\langle\theta, \bar{\phi}\rangle - \psi(\theta)]\right\}p_0(\mathbf{x}), \quad \text{where} \quad \bar{\phi} = \frac{1}{n}\sum_{i=1}^n \phi(x_i). \end{aligned}$$

The corresponding log-likelihood, and score with respect to θ and μ are therefore:

- **Log-likelihood:** $\ell_\theta(\mathbf{x}) = n[\langle\theta, \bar{\phi}\rangle - \psi(\theta)] + \text{ const}$
- **Score w.r.t.** θ: $\nabla_\theta\ell_\theta(\mathbf{x}) = n[\bar{\phi} - \nabla\psi(\theta)]$
- **Score w.r.t.** μ: $\nabla_\mu\ell_\theta(\mathbf{x}) = \Sigma^{-1}n[\bar{\phi} - \nabla\psi(\theta)]$
- **Information w.r.t.** θ: $\mathcal{I}_\theta = \mathbb{E}[\nabla\ell_\theta\nabla\ell_\theta^T] = -\mathbb{E}[\nabla^2\ell_\theta] = n\Sigma$.
- **Information w.r.t.** μ: $\mathcal{I}_\mu = n\Sigma^{-1}$.

Remark. Information matrix quantifies how much information the observable statistic $\phi(\mathbf{x})$ contains about the parameter of interest.

We compute the **MLE** of natural parameter θ by solving the following equation for θ.

$$\begin{aligned} &\nabla\ell_\theta(\mathbf{x}) = 0 \Leftrightarrow \bar{\phi} = \nabla\psi(\hat{\theta}^{\mathrm{MLE}}) \quad \Leftrightarrow \\ &\hat{\theta}^{\mathrm{MLE}} = (\nabla\psi)^{-1}(\bar{\phi}) \text{ by the invertibility of } \nabla\psi. \end{aligned}$$

Similarly, we can also find the MLE for the mean μ by differentiating the log-likelihood w.r.t. μ and setting it to 0. Since we focus on strictly convex CGFs (which imply $\Sigma \succ 0$), the MLE can be computed to be $\hat{\mu}^{\mathrm{MLE}} = \bar{\phi}$. Therefore, we notice that the mapping $\nabla\psi$ also maps the MLEs.
Remark. As a side note, we can see that the score function has 0 expectation:

$$\mathbb{E}[\nabla\ell_\theta(\mathbf{x})] = \mathbb{E}[\bar{\phi}] - \nabla\psi(\theta) = \mu - \mu = 0.$$

Asymptotics of MLE: We can leverage the sample average structure of $\bar{\phi}$ and obtain its asymptotic distribution using Central Limit Theorem (CLT). That is, with a slight abuse of notation

$$\hat{\mu}^{\text{MLE}} = \bar{\phi} = \frac{1}{n}\sum_{i=1}^{n}\phi(x_i) \approx\sim \mathcal{N}(\mu, \Sigma/n).$$

Here, it is worth noting that the distribution $\mathcal{N}$ is approximate, but the mean and the variance are exact. The correct way to state this result is

$$\sqrt{n}(\hat{\mu}^{\text{MLE}} - \mu) \xrightarrow{d} \mathcal{N}(0, \Sigma). \tag{2.1}$$

The asymptotic distribution of $\hat{\theta}^{\text{MLE}}$ requires an extra derivation. Notice that there is a non-linearity $(\nabla\psi)^{-1}$ applied to the sample average form $\bar{\phi}$. We know, by CLT, that $\bar{\phi}$ will be Gaussian, but we need a way of dealing with the non-linear function applied to it.

Proposition 9 (Delta Method)**.** *Assume that a random variable is asymptotically normal, i.e.,* $\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{d} \mathcal{N}(0, \Sigma)$. *Then, for a differentiable function* f*, we have*

$$\sqrt{n}(f(\hat{\mu}) - f(\mu)) \xrightarrow{d} \mathcal{N}(0, \nabla f(\mu)^\top \Sigma \nabla f(\mu)).$$

Using the Delta method on the asymptotic result obtained for $\hat{\mu}^{\text{MLE}}$ in (2.1), and also recalling that $\hat{\theta}^{\text{MLE}} = \nabla\psi^{-1}(\hat{\mu}^{\text{MLE}})$, we can write

$$\sqrt{n}(\hat{\theta}^{\text{MLE}} - \theta) \xrightarrow{d} \mathcal{N}(0, \nabla_\mu(\nabla\psi)^{-1}(\mu)^\top \Sigma \nabla_\mu(\nabla\psi)^{-1}(\mu)). \tag{2.2}$$

We can compute the quantity $\nabla_\mu(\nabla\psi)^{-1}(\mu)$ using the chain rule (left as exercise), but below we just use the Leibniz notation.

$$\nabla_\mu(\nabla\psi)^{-1}(\mu) = \frac{\mathrm{d}\theta}{\mathrm{d}\mu} = \left[\frac{\mathrm{d}\mu}{\mathrm{d}\theta}\right]^{-1} = \Sigma^{-1}.$$

Therefore, the variance term in (2.2) becomes

$$\nabla_\mu(\nabla\psi)^{-1}(\mu)^\top \Sigma \nabla_\mu(\nabla\psi)^{-1}(\mu) = \Sigma^{-1}.$$

Thus,

$$\hat{\theta}^{\text{MLE}} \approx\sim \mathcal{N}(\theta, \Sigma^{-1}/n) \tag{2.3}$$

or equivalently $\sqrt{n}(\hat{\theta}^{\text{MLE}} - \theta) \xrightarrow{d} \mathcal{N}(0, \Sigma^{-1})$. In (2.3), distribution $\mathcal{N}$, as well as the mean and the variance are approximate.

Remark. Proof for the delta method was hinted using the Taylor Series expansion of the function under consideration. This is a very handy theorem.

2.3 Information inequality

In this section, we will derive a lower bound on the variance of a generic estimator which we call as the information inequality. Later, we will use our main result here to derive the celebrated Cramer-Rao lower bound. Information inequality is very much related to the Fisher information which is where it get its name from. It is a classical concept and defines the notion of *efficiency* for estimators.

As in the previous section, suppose that we have data from an exponential family and we have a statistic of the form $T : \mathbb{R}^d \to \mathbb{R}^p$ with

$$\mathbb{E}_\theta[T(\bar{\phi})] = \xi(\theta)$$

for some differentiable function ξ.

Remark. If we have two matrices A and B, we write $A \succeq B$ if $A - B \succeq 0$, i.e., $A - B$ is positive semi-definite. This is equivalent to saying $\forall u$, $\langle u, (A-B)u\rangle \geq 0$.

Theorem 10 (Information Inequality)**.** *Variance of any estimator of the above form can be lower bounded as*

$$\mathrm{Var}(T(\bar{\phi})) \succeq \frac{1}{n}\nabla\xi(\theta)^\top \Sigma^{-1}\nabla\xi(\theta).$$

Proof. In the first step of the proof, we compute a useful expression for the $\nabla\xi$. We write

$$\begin{aligned}
\nabla\xi(\theta) &= \int \nabla p_\theta(x_1, .., x_n) T(\bar{\phi})^\top \,\mathrm{d}\nu \\
&= \int n[\bar{\phi} - \nabla\psi(\theta)] T(\bar{\phi})^\top p_\theta(x_1, .., x_n)\,\mathrm{d}\nu \\
&= n\mathbb{E}_\theta\Big[(\bar{\phi} - \nabla\psi(\theta))T(\bar{\phi})^\top\Big] \\
&= n\mathbb{E}_\theta\Big[(\bar{\phi} - \nabla\psi(\theta))(T(\bar{\phi}) - \xi(\theta))^\top\Big]
\end{aligned}$$

The first term inside the expectation is in $\mathbb{R}^d$ and the second term belongs to $\mathbb{R}^p$. Therefore, the above expectation is a $d \times p$ matrix.

Next, choose any vector $u \in \mathbb{R}^p$ and compute the quantity,

$$\begin{aligned}
\frac{1}{n}\langle \nabla\xi(\theta)u, \Sigma^{-1}\nabla\xi(\theta)u\rangle &= u^\top \mathbb{E}_\theta\Big[(T(\bar{\phi}) - \xi(\theta))(\bar{\phi} - \nabla\psi(\theta))^\top\Big]\Sigma^{-1}\nabla\xi(\theta)u \\
&= \mathbb{E}_\theta\big[\langle T(\bar{\phi}) - \xi(\theta), u\rangle\langle \bar{\phi} - \nabla\psi(\theta), \Sigma^{-1}\nabla\xi(\theta)u\rangle\big] \\
\text{(by Cauchy-Schwartz)} &\leq \mathbb{E}_\theta\Big[\langle T(\bar{\phi}) - \xi(\theta), u\rangle^2\Big]^{1/2}\mathbb{E}_\theta\Big[\langle \bar{\phi} - \nabla\psi(\theta), \Sigma^{-1}\nabla\xi(\theta)u\rangle^2\Big]^{1/2} \\
&\leq \langle u, \mathrm{Var}(T(\bar{\phi}))u\rangle^{1/2}\big[\langle \Sigma^{-1}\nabla\xi(\theta)u, \mathrm{Var}(\bar{\phi})\Sigma^{-1}\nabla\xi(\theta)u\rangle\big]^{1/2} \\
&= \langle u, \mathrm{Var}(T(\bar{\phi}))u\rangle^{1/2}\left[\frac{1}{n}\langle \nabla\xi(\theta)u, \Sigma^{-1}\nabla\xi(\theta)u\rangle\right]^{1/2}
\end{aligned}$$

where in the last step we used the fact that $\mathrm{Var}(\bar{\phi}) = \frac{1}{n}$. We notice that the second term on the last line is the square root of the left hand side. Canceling these and squaring both sides concludes the proof. □

An immediate corollary of this result is the celebrated Cramer-Rao lower bound.

Corollary 11 (Cramer-Rao Lower Bound)**.** *If $T(\bar{\phi})$ is an unbiased estimator for θ, i.e., $\mathbb{E}_\theta[T(\bar{\phi})] = \theta$, then*

$$\mathrm{Var}(T(\bar{\phi})) \succeq \frac{1}{n}\Sigma^{-1}.$$

The lower bound in the above corollary is the inverse Fisher information with respect to the parameter being estimated. That is, the bound reads $\mathrm{Var}(T(\bar{\phi})) \succeq \mathcal{I}_\theta^{-1}$.

If we were to estimate another parameter such as μ, we can derive a similar bound using the information inequality. In this case, our unbiased estimator $T(\bar{\phi})$ (for μ) has an expectation

$$\mathbb{E}_\theta[T(\bar{\phi})] = \xi(\theta) = \mu.$$

Notice that in this case $\xi = \nabla\psi$ and consequently $\nabla\xi = \nabla^2\psi = \Sigma$. Plugging this into the information inequality yields a lower bound on the variance of the estimator as

$$\mathrm{Var}(T(\bar{\phi})) \succeq \frac{1}{n}\Sigma\Sigma^{-1}\Sigma = \frac{1}{n}\Sigma = \mathcal{I}_\mu^{-1}.$$

Remarkably in this case, information inequality yields a lower bound which is again the inverse Fisher information with respect to the parameter being estimated.

Remark. Estimators that achieve Cramer-Rao lower bound are called *efficient*. For example, MLE for μ, $\bar{\phi}$, has the variance $\frac{1}{n}\Sigma$ which is the Cramer-Rao lower bound! So MLE already achieves this bound in this case. Although it is worth noting that MLE in general may not be efficient, yet it is asymptotically efficient.

3 Asymptotic Statistics

In this section, we discuss the asymptotic properties of the parametric models. We will start with describing the supervised learning setup which will be the focus of next few lectures.

3.1 Supervised learning setting

We assume that we observed n pairs of feature/response pairs $(x_i, y_i) \sim p(x, y)$ for $i = 1, 2, ..., n$ where $x \in \mathcal{X} \subset \mathbb{R}^d$ and $y \in \mathcal{Y} \subset \mathbb{R}$ (which could be a real number or discrete class label). Data pairs are i.i.d. and underlying joint distribution $\sim p(x, y)$ is unknown to us. Our goal is to learn some function $\hat{f} : \mathcal{X} \to \mathcal{Y}$ using the observed data that will help us predict y_i given features x_i, i.e., $y_i \approx \hat{f}(x_i)$.

We will need to define a measure to evaluate the quality of learned function $\hat{f}$.

- **Loss:** For this, we choose a loss function $\ell(y, f(x)) : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$. For example, a commonly used loss function is the squared error loss function $\ell(y, f(x)) = (y - f(x))^2$, or another one is the absolute value of the error $\ell(y, f(x)) = |y - f(x)|$. Loss function evaluates the error on only one sample.
- **Risk:** However, we would like to measure the error on average which is why we define the risk $R : \mathcal{F} \to \mathbb{R}_+$ of this function to be $R(f) = \mathbb{E}[\ell(y, f(x))]$. Hereby, the expectation will be implicitly over all random variables inside brackets, and $\mathcal{F}$ denotes the set of functions. The risk is a function of f and it also depends on the joint density $p(x, y)$, and loss ℓ.
- **Goal (revised):** Find $f \in \mathcal{F}$ such that $R(f)$ is small (to be revised again).

Example. [Bias-Variance Decomposition (first step)] We choose the loss as the squared error loss, $\ell(y, f(x)) = (y - f(x))^2$ and write the risk as

$$\begin{aligned}
R(f) &= \mathbb{E}[(y - f(x))^2] \\
&= \mathbb{E}[\mathbb{E}[(y - f(x))^2 | x]] \quad \text{(Law of iterated expectations)} \\
&= \mathbb{E}\Big[\mathbb{E}[(y - \mathbb{E}[y|x] + \mathbb{E}[y|x] - f(x))^2 | x]\Big] \\
&= \mathbb{E}\Big[\mathbb{E}[(y - \mathbb{E}[y|x])^2 | x] + \mathbb{E}[(f(x) - \mathbb{E}[y|x])^2 | x] + 2\underbrace{\mathbb{E}[(y - \mathbb{E}[y|x])(\mathbb{E}[y|x] - f(x))|x]}_{=0}\Big] \\
&= \underbrace{\mathbb{E}[\mathrm{Var}(y|x)]}_{\text{Irreducible error}} + \mathbb{E}[(f(x) - \mathbb{E}(y|x))^2] \quad = \text{Irreducible error + Variance + Bias}^2
\end{aligned}$$

Since the irreducible error is not a function of f, the lower bound on the risk of f is obtained when $f_*(x) = \mathbb{E}[y|x]$. This is attainable if $f_* \in \mathcal{F}$. That is,

$$\inf_{f \in \mathcal{F}} R(f) = \mathbb{E}[\mathrm{Var}(y|x)] + \inf_{f \in \mathcal{F}} E[(f(x) - E(y|x))^2].$$

We will return to bias-variance decomposition later.

3.1.1 Parametric Models

When we are searching for a function f satisfying $y_i \approx f(x_i)$ for $i = 1, ..., n$, we need to restrict ourselves to a specific set of functions $\mathcal{F}$ to avoid overfitting. Otherwise, we can simply choose any function satisfying $y_i = f(x_i)$ for $\forall i$.

In this subsection, we focus our attention on a parametric function class $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ where f_θ is a function (or hypothesis) and Θ is the parameter space.

Example. Consider the set of linear functions that have weights constrained in a ball of radius λ,

$$\mathcal{F} = \{f_\theta(x) = \langle x, \theta \rangle \ : \ \|\theta\|_2 \leq \lambda\}$$

Notice that the parameter space is given by $\Theta = \{\theta \ : \ \|\theta\|_2 \leq \lambda\}$.

In the case of parametric models, it is generally redundant to write the function f_θ, instead we will simply use the parameter θ to describe it. For example,

$$\ell(y, f_\theta(x)) \triangleq \ell((y, x), \theta) \quad \text{and} \quad R(f_\theta) \triangleq R(\theta),$$

is more compact and conveys the same information for parametric function classes.

We would like to minimize the population risk $R(\theta)$, that is, we want

$$\theta_* \in \arg\min_{\theta \in \Theta} R(\theta) = \mathbb{E}[\ell((x, y), \theta)]$$

for $(x, y) \sim p$. But we don't have access to the joint density p, therefore we cannot minimize this objective. Instead, what we can estimate the population risk with the empirical risk using our n training samples. The empirical risk is just a sample mean estimator for the population risk and given as

$$\hat{\theta} \in \arg\min_{\theta} \hat{R}(\theta) \coloneqq \frac{1}{n} \sum_{i=1}^{n} \ell((x_i, y_i), \theta).$$

Notice the hat in $\hat{R}$ and $\hat{\theta}$ which indicates that these are estimators that depend on data. These quantities are random variables (or vectors) whereas $R(\theta)$ and θ_* are deterministic values ($R(\hat{\theta})$ is also random).

- The quantity $\hat{R}(\hat{\theta})$ is the **training error**.
- The quantity $R(\hat{\theta})$ is simply **test error**. It is worth noting that in machine learning courses, we define test error as an estimator to this quantity.

Notice that when n is large, we expect to have $R(\theta) \approx \hat{R}(\theta)$; thus, it would makes sense to have the minimizers of these functions close together $\hat{\theta} \approx \theta_*$.

The following quantity will be used repeatedly as a notion of generalization error.

Definition 12 (Excess risk)**.** *We define the excess risk of an estimator $\hat{\theta}$ as the distance between the test error and the minimum achievable error*

$$\textit{Excess risk} = R(\hat{\theta}) - R(\theta_*).$$

3.2 MLE Framework

In the MLE framework, we assume that data pairs are sampled in the following hierarchical way

$$\begin{aligned} y_i | x_i \sim & p_{\theta_*}(y|x) \\ x_i \sim & p(x) \end{aligned}$$

where θ_* is the true but unknown parameter. However, we make the very strong assumption that the parametric form $p_\theta(x)$ is known. This is like assuming that we know a random variable z is Gaussian $z \sim \mathcal{N}(\theta_*, 1)$ but we don't know the value of θ_*.

The MLE is motivated as a finding "the most likely" parameter. If we translate this to our framework, we simply choose a loss function that is the negative of the log-likelihood, i.e.

$$\ell((y, x), \theta) = -\log p_\theta(y|x).$$

We give two examples below.

Example. Parametric distribution is normal with mean $\langle x, \theta \rangle$ and some variance σ^2 (which doesn't matter). That is

$$\begin{aligned} y|x &\sim \mathcal{N}(\langle x, \theta \rangle, \sigma^2) \\ p_\theta(y|x) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(y - \langle x, \theta \rangle)^2}{2\sigma^2} \right\} \\ \ell((x, y), \theta) &= -\log p_\theta(y|x) = (y - \langle x, \theta \rangle)^2 + \text{ const.} \end{aligned}$$

which is the squared error loss yielding linear regression.

Example. Parametric distribution is Bernoulli with mean $\sigma(\langle x, \theta \rangle)$ where σ, in this case, is the sigmoid function. That is

$$\begin{aligned} y|x &\sim \text{Ber}(\sigma(\langle x, \theta \rangle)) \\ p_\theta(y|x) &= \sigma(\langle x, \theta \rangle)^y (1 - \sigma(\langle x, \theta \rangle))^{1-y} \\ \ell((x, y), \theta) &= -\log p_\theta(y|x) = -y \log(\sigma(\langle x, \theta \rangle)) - (1 - y) \log(1 - \sigma(\langle x, \theta \rangle)) \end{aligned}$$

which is the cross-entropy loss yielding logistic regression. Both of above settings belong to large class of regression models called generalized linear models (GLMs). They are obtained by modeling the natural parameter in exponential families with a linear function of feature vector. As seen above, Gaussian leads to linear regression whereas Bernoulli leads to logistic regression.

MLE problem: We observe n data point: $(y_i, x_i) \sim p_{\theta_*}(y|x)p(x)$, $i = 1, \ldots, n$. Our goal here is to estimate the true parameter θ_*, by minimizing the empirical risk:

$$\hat{\theta} = \arg\min_\theta \ \hat{R}(\theta) = \frac{1}{n} \sum_{i=1}^n \ell((x_i, y_i), \theta) = -\frac{1}{n} \sum_{i=1}^n \log p_\theta(y_i|x_i).$$

Let's see how this is related to population risk minimizer. We start investigating by writing out the gradient and the Hessian of $R(\theta)$.

- $\nabla R(\theta_*) = \mathbb{E}[-\nabla \log p_{\theta_*}(y|x)] = 0$. Therefore, the true parameter is a critical point of the population risk.
- $\nabla^2 R(\theta_*) = \mathbb{E}[-\nabla^2 \log p_{\theta_*}(y|x)] = E[\nabla \log p_{\theta_*}(y|x) \nabla \log p_{\theta_*}(y|x)^T] = \mathcal{I}_{\theta_*} \succeq 0$. This still doesn't prove that θ_* is a local minimum. Note that the Hessian of the risk must be positive semi-definite (PSD) since zz^T is PSD as $u^T zz^T u = (u^T z)^2$.

In what follows, for simplicity we assume that $\mathcal{I}_{\theta_*} \succ 0$ which clearly implies that true parameter θ_* is a local minimum. Actually, if we assume identifiability of our parametric family, that is $\theta \neq \theta' \implies p_\theta \neq p_{\theta'}$, then θ_* can be shown to be a unique global minimum.

3.3 Asymptotics of MLE

First, we need a few definitions.

Definition 13 (Convergence of random variables)**.**

(a) ***Convergence in probability****: We write $\hat{\theta}_n \xrightarrow{p} \theta_*$, if for every $\epsilon > 0$ we have*

$$\mathbb{P}(|\hat{\theta}_n - \theta_*| > \epsilon) \to 0 \quad as \quad n \to \infty.$$

(b) ***Convergence in distribution****: We write $\hat{\theta}_n \xrightarrow{d} \theta_*$, if X_n and X have CDFs $F_n(x)$ and $F(x)$, respectively and for every continuity point of $F(x)$, we have $\lim_{n\to\infty} F_n(x) = F(x)$. This is also called weak convergence as it is a very weak notion of convergence. The letter d in the symbol $\xrightarrow{d}$ is to specify that the convergence is in distribution, and it should not be confused be the dimension d.*

(c) ***Consistency****: We say $\hat{\theta}_n$ is a consistent estimator for θ_* if $\hat{\theta}_n \xrightarrow{p} \theta_*$.*

In our asymptotic setting, we fix the dimension d and let number of samples $n \to \infty$. We drop the subscript n to ease the notation.

3.3.1 Asymptotic normality

The following theorem is charactering the asymptotics of the MLE.

Theorem 14 (Asymptotics of MLE)**.** *Assume that $\hat{\theta}$ is consistent for θ_*, and the Fisher information satisfies $\mathcal{I}_{\theta_*} \succ 0$, and that $\sup_\theta \|\nabla^3 \log p_\theta\|_{op} < B$. Then,*

1. $\sqrt{n}(\hat{\theta} - \theta_*) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}_{\theta_*}^{-1})$.

2. $n(R(\hat{\theta}) - R(\theta_*)) \xrightarrow{d} \frac{1}{2}\chi^2_d$.

Remark. We make two important remarks about the above theorem.

1. The first result is giving us the asymptotic distribution of the MLE. We observe that the variance of this distribution is the inverse Fisher information which validates its name: larger the Fisher information is, lower the variance of this distributions. Therefore, the estimator gives more information about the true parameter.

   It is worth noting that these types of distributional results are useful in constructing confidence intervals; hence quantifying uncertainty in models.

2. The second item above is the asymptotic distribution of the excess risk. Since χ^2_d is a random variable with mean d and variance $2d$, the right hand side is roughly of order d/n. That is,

   $$R(\hat{\theta}) - R(\theta_*) \approx \mathcal{O}\left(\frac{d}{n}\right).$$

   The excess risk gets worse with increased dimension buy gets better with increased number of samples. We should emphasize that this is an asymptotic rate and it is quite fast compared to the non-asymptotic rates that we will obtain in the future lectures. It is also worth noting that this is an equality rather than a upper bound.

Proof sketch.

We start by proving the first item, the normality of the MLE. The distribution of excess risk will follow. Our proof outline is 1- we apply Taylor's theorem, 2- identify a term that is an iid sum which converges to a Gaussian random variable by central limit theorem (CLT), 3- show that the other quantities converge in probability to deterministic quantities. We finally apply the Slutsky's theorem to conclude the proof.

Lemma 15 (Slutsky's Theorem)**.** *For a sequence of random variables* $\{x_n, y_n, z_n\}_{n\in\mathbb{N}}$ *satisfying* $x_n \xrightarrow{d} x$, $y_n \xrightarrow{p} a$ *and* $z_n \xrightarrow{p} b$ *where* a, b *are constants, then we have* $x_n y_n + z_n \xrightarrow{d} ax + b$.

We omit the proof as it is a simple application of the continuous mapping theorem.

We notice that $\nabla R(\theta_*) = \nabla \hat{R}(\hat{\theta}) = 0$. We expand the latter using the Taylor's theorem around the true parameter θ_*.

$$\nabla \hat{R}(\hat{\theta}) = \nabla \hat{R}(\theta_*) + \nabla^2 \hat{R}(\theta_*)(\hat{\theta} - \theta_*) + \frac{1}{2}\nabla^3 \hat{R}(\bar{\theta})[\hat{\theta} - \theta_*, \hat{\theta} - \theta_*].$$

Above, the last term is a tensor in $\nabla^3 \hat{R}(\bar{\theta}) \in \mathbb{R}^{d\times d\times d}$, when multiplied by a vector (e.g. $\hat{\theta} - \theta_*$) it reduces to a $d \times d$ matrix. Also, $\bar{\theta}$ is chosen somewhere on the line of and between $\hat{\theta}$ and θ_* (it is worth noting that mean value theorem doesn't hold for vector valued functions which can be easily fixed by using the integral form Taylor's theorem).

We notice that the left hand side is zero. Rearranging terms, we get

$$-\nabla \hat{R}(\theta_*) = [\nabla^2 \hat{R}(\theta_*) + \frac{1}{2}\nabla^3 \hat{R}(\bar{\theta})(\hat{\theta} - \theta_*)](\hat{\theta} - \theta_*) \tag{3.1}$$

Multiplying both sides with $\sqrt{n}$, we obtain

$$\underbrace{-\sqrt{n}\nabla \hat{R}(\theta_*)}_{\text{iid sum}/\sqrt{n}} = [\underbrace{\nabla^2 \hat{R}(\theta_*)}_{\text{iid sum}/n} + \underbrace{\frac{1}{2}\nabla^3 \hat{R}(\bar{\theta})(\hat{\theta} - \theta_*)]}_{\xrightarrow{p} 0} \underbrace{\sqrt{n}(\hat{\theta} - \theta_*)}_{\text{of interest}}$$

We observe that the left hand side of (3.1) is a iid sum divided by $\sqrt{n}$. By the CLT, we obtain

$$-\sqrt{n}\nabla \hat{R}(\theta_*) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \nabla \log p_{\theta_*}(y_i|x_i) \xrightarrow{d} \mathcal{N}(0, \mathrm{Cov}(\nabla \log p_{\theta_*}(y_i|x_i))).$$

Here, the expected value is 0 since $\mathbb{E}[\nabla \log p_{\theta_*}(y_i|x_i)] = 0$, and $\mathrm{Cov}(\nabla \log p_{\theta_*}(y_i|x_i)) = \mathcal{I}_{\theta_*}$.

For the first term on the right hand side of (3.1), we have another iid sum but this time divided by n. We use law of large numbers (LLN) to obtain

$$\nabla^2 \hat{R}(\theta_*) = \frac{1}{n}\sum_{i=1}^{n} \nabla^2 \log p_{\theta_*}(y_i|x_i) \xrightarrow{p} \nabla^2 R(\theta_*) = \mathcal{I}_{\theta_*}.$$

The second term on the right hand side of (3.1) converges to 0 in probability by the consistency assumption. Therefore, multiplying both sides with $\mathcal{I}_{\theta_*}^{-1}$, we obtain that

$$\sqrt{n}(\hat{\theta} - \theta_*) \xrightarrow{d} \mathcal{I}_{\theta_*}^{-1}\mathcal{N}(0, \mathcal{I}_{\theta_*}). \tag{3.2}$$

We proceed by using a very useful property of Gaussian random vectors.

Lemma 16. *Let $z \sim \mathcal{N}(\mu, \Sigma)$ be a d-dimensional Gaussian random vector. Then for a matrix $A \in \mathbb{R}^{l \times d}$ we have $Az \sim \mathcal{N}(A\mu, A\Sigma A^\top)$.*

Using the above lemma together with (3.2) and obtain

$$\sqrt{n}(\hat{\theta} - \theta_*) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}_{\theta_*}^{-1}). \tag{3.3}$$

This concludes the proof of the first part. For the proof of second part, we again use Taylor's theorem and write

$$R(\hat{\theta}) - R(\theta_*) = \Big\langle \nabla R(\theta_*), \hat{\theta} - \theta_* \Big\rangle + \frac{1}{2}\Big\langle \nabla^2 R(\theta_*)(\hat{\theta} - \theta_*), \hat{\theta} - \theta_* \Big\rangle + \frac{1}{6}\nabla^3 \hat{R}(\bar{\theta})[\hat{\theta} - \theta_*, \hat{\theta} - \theta_*, \hat{\theta} - \theta_*],$$

where again the first term on the right hand side disappears, and $\bar{\theta}$ is in between θ_* and $\hat{\theta}$ (this time without any issue since R is real-valued). Multiplying both sides with n and rearranging, we obtain

$$n\{R(\hat{\theta}) - R(\theta_*)\} = \frac{1}{2}\Big\langle \sqrt{n}(\hat{\theta} - \theta_*), \{\nabla^2 R(\theta_*) + \frac{1}{3}\nabla^3 R(\bar{\theta})[\hat{\theta} - \theta_*]\}\sqrt{n}(\hat{\theta} - \theta_*) \Big\rangle.$$

Using the previous result (3.3), we know that $\sqrt{n}(\hat{\theta} - \theta_*) \xrightarrow{d} z$ where $z \sim \mathcal{N}(0, \mathcal{I}_{\theta_*}^{-1})$, and the term multiplying $\nabla^3 R$ vanishes due to consistency. Therefore, as $n \to \infty$, the right hand side converges in distribution to

$$n\{R(\hat{\theta}) - R(\theta_*)\} \xrightarrow{d} \frac{1}{2}\langle z, \mathcal{I}_{\theta_*} z \rangle.$$

We use Lemma 16 to deduce that

$$\frac{1}{2}\langle z, \mathcal{I}_{\theta_*} z \rangle = \frac{1}{2}\langle \mathcal{I}_{\theta_*}^{1/2} z, \mathcal{I}_{\theta_*}^{1/2} z \rangle = \frac{1}{2}\|\tilde{z}\|^2 \sim \frac{1}{2}\chi_d^2$$

where $\tilde{z} \sim \mathcal{N}(0, I)$ with I denoting the identity matrix. This concludes the proof of the second statement.

□

It is important to identify the contribution of each assumption. It is obvious that the CLT follows from the iid average structure of the MLE problem (also there for many learning tasks). The bounded third derivative is needed to control higher-order terms. Lastly, consistency is needed to kill the third-order term which reduces everything to a quadratic problem in the asymptotic limit.

3.3.2 Consistency

It turns out that the consistency assumption is actually true for the MLE, under certain assumptions (Note that the below assumptions are stronger than what is in fact needed).

Theorem 17 (MLE is consistent)**.** *Assume that the following assumptions are satisfied.*

(a) **Uniform convergence:** *The empirical process satisfies* $\sup_{\theta \in \Theta} |\hat{R}(\theta) - R(\theta)| \xrightarrow{p} 0$.

(b) **Identifiability:** *For every* $\epsilon > 0$, $\inf_{\theta : \|\theta - \theta_*\| \geq \epsilon} R(\theta) > R(\theta_*)$.

(c) **Compactness***: Θ is non-empty and compact.*

Then, $\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}} \hat{R}(\theta)$ *is consistent.*

Remark. The first assumption above is a very strong notion of convergence and it will be quite handy when we talk about generalization. The second assumption simple means that we can identify the function has a unique minimizer θ_* and around that point, R grows. The last assumption is only needed to ensure that θ_* and $\hat{\theta}$ belong to the set Θ.

Proof. By the compactness assumption, we have $\hat{\theta}, \theta_* \in \Theta$. Next, notice that since $\hat{\theta}$ minimizes $\hat{R}$ in Θ, we have $\hat{R}(\hat{\theta}) \leq \hat{R}(\theta_*)$. We can write

$$\begin{aligned}
\hat{R}(\hat{\theta}) \leq & \hat{R}(\theta_*) \\
= & \hat{R}(\theta_*) - R(\theta_*) + R(\theta_*) \\
\leq & \sup_{\theta \in \Theta} \left| \hat{R}(\theta) - R(\theta) \right| + R(\theta_*) \xrightarrow{p} R(\theta_*) \quad \text{by assump (a).}
\end{aligned} \tag{3.4}$$

Also, since θ_* minimizes R, we write

$$\begin{aligned}
0 \leq R(\hat{\theta}) - R(\theta_*) \leq & R(\hat{\theta}) - \hat{R}(\hat{\theta}) \quad \text{as } n \to \infty \text{ by (3.4)}, \\
\leq & \sup_{\theta \in \Theta} \left| \hat{R}(\theta) - R(\theta) \right| \xrightarrow{p} 0, \quad \text{by assump (a).}
\end{aligned} \tag{3.5}$$

Notice that we squeezed the excess risk between zeros. So for every $\epsilon > 0$, the following holds for the events

$$\left\{ \| \hat{\theta} - \theta_* \| \geq \epsilon \right\} \underset{\text{by assumption (b)}}{\subseteq} \left\{ R(\hat{\theta}) - R(\theta_*) > \delta_\epsilon \right\}$$

Probability of the right hand side above goes to 0 as we let $n \to \infty$ due to (3.5).

□

4 Uniform Convergence $\implies$ Generalization

Most of this section will rely on the notation introduced in Section 3.1. Our objective is to relate the generalization performance of a learning algorithm to certain properties of the problem at hand. We have already done this in the case MLE, where we characterized the behavior of the excess risk as $R(\hat{\theta}) - R(\theta_*) \approx d/n$ where d is the dimension of the features and n is the number of samples. This characterization tells us that as the number of samples increase, the excess risk decrease with a rate of n^{-1}, and as the dimension of the features increase, the excess risk also increase with a rate of d. But there were a couple of limitations of this result. First, this result was asymptotic, i.e. it only holds when $n \to \infty$. Second, the entire MLE setup assumes that we know the true parametric form of the data distribution. These are very strong assumptions which do not hold in practice.

In the sequel, our objective is modified. We do not assume that the data distribution is known anymore. We only assume that data samples are iid from some distribution. The problem we consider can be summarized as follows.

$$f_* = \operatorname*{argmin}_{f \in \mathcal{F}} \; R(f) \coloneqq \mathbb{E}[\ell((y, x), f)]$$

As before the expectation is over the true unknown distribution $(y, x) \sim p(y, x)$; thus, we cannot compute this expectation. Luckily, we can estimate this risk with a sample mean estimator (aka empirical risk). That is,

$$\hat{f} = \operatorname*{argmin}_{f \in \mathcal{F}} \; \hat{R}(f) \coloneqq \frac{1}{n} \sum_{i=1}^{n} \ell((y_i, x_i), f)$$

Notice that $\hat{R}(f)$ is an estimator to $R(f)$ and when n is large, they will be close to each other. The hope is that, their minimizers are also close, and we will show that they actually are!

Two quantities have a big impact on the generalization performance of our learning algorithm: 1- the complexity of the function class $\mathcal{F}$, and 2- the number data points used in training. We would like to characterize the behavior of the excess risk in the following way.

$$R(\hat{f}) - R(f_*) \leq \frac{\text{a func of comp of } \mathcal{F}}{\text{a func of } n}. \tag{4.1}$$

In the case of MLE, we sort of achieved this (d is not really a complexity measure of the function class but ...).

We notice that the left hand side of (4.1) is a random variable. Therefore, we need to make a probabilistic argument for this statement to make sense (e.g. almost sure, or high probability etc). We choose high probability. More formally:

$$\mathbb{P}\big(\underbrace{R(\hat{f}) - R(f_*) \geq \epsilon}_{\text{bad event}} \big) < \underbrace{\delta}_{\text{small probability}}$$

Here, ϵ and δ are ideally smaller numbers.

4.1 From excess risk to empirical process

We can decompose the excess risk in three terms.

$$R(\hat{f}) - R(f_*) = \underbrace{[R(\hat{f}) - \hat{R}(\hat{f})]}_{\text{not iid sum}} + \underbrace{[\hat{R}(\hat{f}) - \hat{R}(f_*)]}_{\leq 0} + \underbrace{[\hat{R}(f_*) - R(f_*)]}_{\text{iid sum}/n}.$$

The first term above is the main term we need heavy lifting. This is because $\hat{f}$ is a random variable, and it breaks the iid sum structure of the empirical risk (as we will see soon, iid structure is very handy). The second term is less than or equal to 0 since $\hat{f}$ minimizes the empirical risk. The last term is an iid sum since f_* is deterministic (not random). As before, we can consider uniform bounds over the feasible set to solve issues that come from non-iid structure.

$$\begin{aligned} R(\hat{f}) - R(f_*) &\le |\hat{R}(\hat{f}) - R(\hat{f})| + 0 + |\hat{R}(f_*) - R(f_*)| \\ &\le \sup_{f\in\mathcal{F}} |\hat{R}(f) - R(f)| + 0 + \sup_{f\in\mathcal{F}} |\hat{R}(f) - R(f)| \\ \underbrace{R(\hat{f}) - R(f_*)}_{\text{excess risk}} &\le \underbrace{2 \cdot \sup_{f\in\mathcal{F}} |\hat{R}(f) - R(f)|}_{\text{empirical process}} \end{aligned} \tag{4.2}$$

The right hand side is called empirical process in statistics. If we can control the empirical process, we can control the generalization error. Intuitively we have bounded the risk of the empirical estimator by the "worst-case" function possible from the function class. We see that the bound in (4.2) translates immediately to

$$\mathbb{P}\big(R(\hat{f}) - R(f_*) \ge \epsilon\big) \le \mathbb{P}\left(\sup_{f\in\mathcal{F}} \Big|\hat{R}(f) - R(f)\Big| \ge \frac{\epsilon}{2} \right), \tag{4.3}$$

where the inequality in (4.3) is due to the fact that if the event $R(\hat{f}) - R(f_*) \ge \epsilon$ happens, since we have (4.2), the event $\sup_{f\in\mathcal{F}} \Big|\hat{R}(f) - R(f)\Big| \ge \frac{\epsilon}{2}$ will also happen.

Uniform convergence generally refers to that the empirical process $\sup_{f\in\mathcal{F}} \Big|\hat{R}(f) - R(f)\Big|$ converges to 0 in probability. Because of (4.3), we see that uniform convergence implies generalization. But we can also talk about explicit convergence rates.

4.2 Finite function classes, $|\mathcal{F}| < \infty$

Our first result in this direction is for finite function classes. Denoting the number of elements in a set with $|\cdot|$, in the following, we assume $|\mathcal{F}| < \infty$.

Theorem 18 (Generalization of Finite Function Classes)**.** *If the function class is finite (i.e. $|\mathcal{F}| < \infty$) and loss is bounded $\ell \le B$, then we have,*

$$\mathbb{P}\left(R(\hat{f}) - R(f_*) < B\sqrt{\frac{2\log(2|\mathcal{F}|) + 2\log\delta^{-1}}{n}} \right) > 1 - \delta. \tag{4.4}$$

Remark.

- The above theorem reads, with probability at least $1-\delta$, we have

  $$R(\hat{f}) - R(f_*) < B\sqrt{\frac{2\log(2|\mathcal{F}|) + 2\log\delta^{-1}}{n}},$$

  This is true in a non-asymptotic sense.

- The complexity measure of the function class $\mathcal{F}$ turns out to be very intuitive in this case, simply the number of functions. The generalization error depends on this quantity in a logarithmic way. This is a good dependence since log grows very slow.

- δ is the confidence level for the bad event. Smaller it is, more risk averse the bound is. It should be chosen in a way that the convergence rate is not affected. In the above bound we observe that $\delta^{-1} = 2|\mathcal{F}|$ is a good choice. The resulting convergence rate is $\mathcal{O}\left(\sqrt{\frac{\log(|\mathcal{F}|)}{n}}\right)$.

- We see that the dependence on number of sample dropped to $\sqrt{n}$ from n (compared to MLE). This is the price we paid to make this result very general, i.e. non-asymptotic and unknown distribution.

- Clearly, this setup doesn't cover any interesting function class since $|\mathcal{F}| < \infty$ almost never holds. For example, think of class of linear functions. How many functions are there in that set?

- The assumption on the loss is also restrictive. It doesn't cover, for example, square loss; yet it does cover 0-1 loss.

Proof. The proof will follow from three steps. In the first step we use a concentration of measure argument for iid averages. In the second, we use the uniform convergence argument derived in (4.3). In the last step, we control the empirical process to obtain a bound on the generalization error. We start with a classical concentration result that will be handy.

Lemma 19 (Hoeffding's inequality)**.** *Suppose $z_1, z_2, \ldots, z_n$ are independent random variables (not necessarily iid) where $a_i \leq X_i \leq b_i$ almost surely. For the partial sums $S_n = n^{-1}\sum_i z_i$ and $\forall \epsilon > 0$, we have*

$$\mathbb{P}\big(|S_n - \mathbb{E}[S_n]| > \epsilon\big) \leq 2 \cdot \exp\left\{ - \frac{2n^2\epsilon^2}{\sum_i (b_i - a_i)^2} \right\}.$$

Remark. The one sided version is also holds without the factor 2 on the right hand side.

1. **Concentration**: We notice that for a non-random f (this excludes $\hat{f}$), $\hat{R}(f) - R(f)$ is the same as $S_n - \mathbb{E}[S_n]$ if we let $z_i \coloneqq \ell((y_i, x_i), f)$. Since loss is bounded by B, by the Hoeffding's inequality, the sample average is concentrating around the true average. That is,

$$\begin{aligned}\mathbb{P}\big(|\hat{R}(f) - R(f)| \geq \epsilon/2\big) &\leq 2 \cdot \exp\left\{ - \frac{n^2\epsilon^2}{2\sum_i B^2} \right\} \\ &\leq 2 \cdot \exp\left\{ - \frac{n\epsilon^2}{2B^2} \right\}\end{aligned}$$

2. **Union bound**: Next, we make use of the finite function class assumption to handle the empirical process.

$$\begin{aligned}\mathbb{P}\left(\sup_{f \in \mathcal{F}} |\hat{R}(f) - R(f)| \geq \epsilon/2 \right) &= \mathbb{P}\left(\bigcup_{f \in \mathcal{F}} \Big\{ |\hat{R}(f) - R(f)| \geq \epsilon/2 \Big\} \right) \\ &\leq \sum_{f \in \mathcal{F}} \mathbb{P}\big(|\hat{R}(f) - R(f)| \geq \epsilon/2\big) \quad \text{(by the union bound)} \\ &\leq 2|\mathcal{F}| \exp\Big\{ - \frac{n\epsilon^2}{2B^2} \Big\}.\end{aligned}$$

3. **Uniform convergence** $\implies$ **Generalization**: Finally, we use the inequality derived in (4.3) to conclude

$$\begin{aligned}\mathbb{P}\big(R(\hat{f}) - R(f_*) \geq \epsilon\big) &\leq \mathbb{P}\left(\sup_{f\in\mathcal{F}} |\hat{R}(f) - R(f)| \geq \epsilon/2\right) \\ &\leq 2|\mathcal{F}| \exp\Big\{-\frac{n\epsilon^2}{2B^2}\Big\} := \delta.\end{aligned}$$

Solving for the δ in the above equation, we obtain

$$\epsilon^2 = \frac{2B^2}{n}\log(2|\mathcal{F}|\delta^{-1}).$$

By substituting $\epsilon(\delta)$ we recover (4.4).

□

5 Covering with ε-nets

The main objective in this lecture is to relax the strong and impractical assumption of Theorem 18, namely the finite function class condition. This assumption is valid only if the practitioner is allowed to choose among a finite number of functions.

In the majority of machine learning methodology, we train our models (weights) by assuming a parametric structure on the function class and the parameter space is infinitely rich (more like uncountably rich).

Model Setup: Suppose that we have a family of functions parametrized in the following sense

$$\mathcal{F} = \{f_\theta : \theta \in \Theta \subseteq \mathbb{R}^d\}$$

and a loss function $\ell((x, y), f_\theta) \coloneqq \ell((x, y), \theta)$. In the sequel, we assume that the parameter space is a d-dimensional ball with radius R, i.e. $\Theta = \{\theta : ||\theta|| \leq R\}$, while the loss function ℓ is bounded by B and is L-Lipschitz continuous in θ. It is worth noting that Θ can be any set that is compact, which can be confined inside a ball with some radius, so all the arguments we will soon be making still are valid in general.

Definition 20 (Lipschitz continuous)**.** *A function f is L-Lipschitz continuous if*

$$\forall \theta, \theta', \quad |f(\theta) - f(\theta')| \leq L\|\theta - \theta'\|.$$

Functions that satisfy this condition have stable fluctuations, i.e., if θ and θ' are points that are close to each other, the function values f evaluated at θ and θ' should also be close to each other. For differentiable functions, the above assumption is equivalent to having a uniformly bounded gradient $\|\nabla f(\theta)\| \leq L$. We also notice that it enforces function to have at most linear growth, i.e., let $\theta' = 0$. This rules out the options such as quadratic functions like $f(\theta) = \theta^2$. Functions that are strongly convex cannot be Lipschitz continuous.

5.1 ε-covers of sets in $\mathbb{R}^d$

Remember in the proof of Theorem 18, we have used a union bound over the finite set of functions. This is the main obstacle in our new setup where we have an uncountable set of parameter space Θ. We simply cannot apply union bound over an uncountable sets! But what we can do is to discretize this uncountable set in a way so that it is a *good* representation of the original set, but we can apply union bound. We introduce the following notion of set covers for this task.

Definition 21 (ε-Net)**.** *For $\varepsilon > 0$, $\mathcal{N}_\varepsilon$ is an ε-net (or an ε-cover) over the set $\Theta \subseteq \mathbb{R}^d$ if for all $\theta \in \Theta$, there exists $\theta' \in \mathcal{N}_\varepsilon$ such that $\|\theta - \theta'\| \leq \varepsilon$. That is,*

$$\forall \theta \in \Theta, \ \exists \theta' \in \mathcal{N}_\varepsilon \quad \text{such that} \quad \|\theta - \theta'\| \leq \varepsilon.$$

The size of the ε-net with smallest size $|\mathcal{N}_\varepsilon|$ is called the covering number.

In our applications, we are ideally looking for ε-nets over our parameter space Θ, that have small number of points. But it is worth noting that we are not looking for the optimal ε-net. The following examples will demonstrate the level of optimality we require for our purposes.

Example: Suppose $\Theta = [0, 1]$. To find the ε-net of Θ, one can divide the interval into shorter intervals, each with length 2ε. By defining the set N_ϵ to include all the endpoints of each small interval, for any point θ in the $[0, 1]$ interval, we can always find a point $\theta' \in \mathcal{N}_\varepsilon$ such that $|\theta - \theta'| \leq \varepsilon$. This is demonstrated in Figure 1, and it gives us a valid $\mathcal{N}_\varepsilon$ with $|\mathcal{N}_\varepsilon| = \frac{1}{2\varepsilon} + 1$.

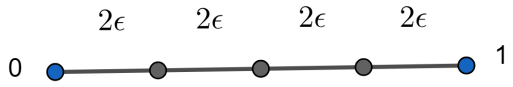


Figure 1: ε-Net for $\Theta = [0, 1]$

Example: Suppose $\Theta = \{\theta \in \mathbb{R}^d : \|\theta\| \leq R\}$. We can do something similar by dividing the ball into grids of size a. Since the number of points required in 1 dimension is $\frac{2R}{a} + 1$, the total number of points required in the d-dimensional space is $(\frac{2R}{a} + 1)^d$. Within each d-dimensional cube of edge length a, the largest distance between the interior points and the vertices comes from the center of the cube, which is $\frac{a\sqrt{d}}{2}$. This is demonstrated in Figure 2. Therefore, to guarantee a full cover of all the points in the ball, the largest grid size should satisfy $\varepsilon = \frac{a\sqrt{d}}{2}$. This leads to the upper bound of the size of an ϵ-net for θ, which is

$$|\mathcal{N}_\varepsilon| \leq \left(\frac{R\sqrt{d}}{\varepsilon} + 1\right)^d \leq \left(\frac{2R\sqrt{d}}{\varepsilon}\right)^d. \tag{5.1}$$

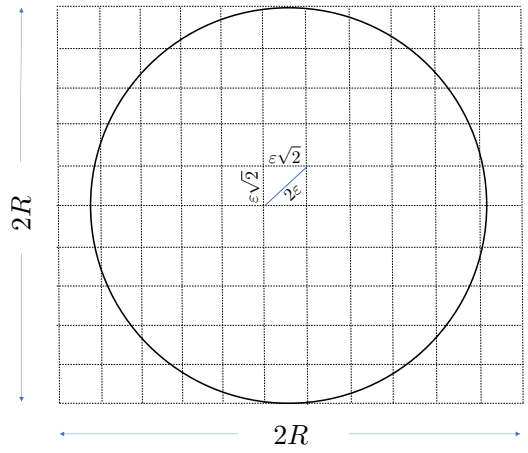


Figure 2: ϵ-Net for $\Theta = \{\theta \in \mathbb{R}^2 : \|\theta\| \leq R\}$

.

The above dependence $\mathcal{O}(d^d)$ is not looking good. But we will next see that the exponential decay in Hoeffding's inequality will be able to compensate for this.

5.2 Generalization for parametrized function classes

We state the following generalization bound for parametrized function classes.

Theorem 22 (Generalization by covering)**.** *Assume that the loss function is bounded by B and L-Lipschitz continuous in its second argument θ. For the parametric function class $\mathcal{F} = \{f_\theta : \|\theta\| \leq R\}$, and the corresponding empirical and population risk minimizers*

$$\hat{\theta} = \operatorname*{argmin}_{\theta\in\Theta} \hat{R}(\theta) \quad \textit{and} \quad \theta_* = \operatorname*{argmin}_{\theta\in\Theta} R(\theta),$$

we have with probability at least $1 - 2e^{-d/2}$

$$R(\hat{\theta}) - R(\theta_*) \leq c\sqrt{\frac{d\log(n)}{n}} \quad \textit{where} \quad c = 2(B \vee 8RL),$$

whenever $n \geq 16$.

Remark. We make a few remarks before proving the above theorem.

- Convergence rate is $\sqrt{\frac{d\log(n)}{n}}$. This is slower than the previous result by a factor of $\log(n)$, which is due to the covering argument we are about to make.
- Note that the function class is parametrized over a ball of radius R. This is not at all needed and our proof would still follow for any compact set Θ by simply replacing R with $\text{diam}(\Theta)/2$.
- The above probability is decaying exponentially fast with dimension. The constants are arbitrary and can be improved with a more careful treatment.

Proof. The proof will be similar to the finite function class case, with an additional step where we discretize the uncountably rich parameter space Θ.

1. **Concentration**: Since loss is bounded by B, for a non-random θ, by the Hoeffding's inequality applied on $\hat{R}(\theta) - R(\theta)$, we obtain

$$\mathbb{P}\big(|\hat{R}(\theta) - R(\theta)| \geq \epsilon/4\big) \leq 2 \cdot \exp\left\{-\frac{n\epsilon^2}{8B^2}\right\}.$$

2. **Discretization**: In order to apply union bound, we first discretize our uncountable parameter space using an ε-net argument. Before we introduce the ε-net, we first derive a few useful inequalities using the Lipschitz continuity of loss. By the definition of $R(\theta)$ and $\hat{R}(\theta)$, if $l((x,y),\theta)$ is L-Lipschitz, then both $R(\theta)$ and $\hat{R}(\theta)$ would also be L-Lipschitz.

   First, we notice that since ℓ is L-Lipschitz, R and $\hat{R}$ are both L-Lipschitz continuous. Next, by the triangle inequality,

$$\begin{aligned}
|\hat{R}(\theta) - R(\theta)| =& |\hat{R}(\theta') - R(\theta') + \hat{R}(\theta) - \hat{R}(\theta') - R(\theta) + R(\theta')| \\
\leq& |\hat{R}(\theta') - R(\theta')| + |\hat{R}(\theta) - \hat{R}(\theta')| + |R(\theta) - R(\theta')| \\
\leq& |\hat{R}(\theta') - R(\theta')| + 2L\|\theta - \theta'\|.
\end{aligned}$$

   Now, let $\mathcal{N}_\Delta$ be a Δ-net over $\Theta \subseteq \mathbb{R}^d$. For any $\theta \in \Theta$, there exists $\theta' \in \mathcal{N}_\Delta$ such that $\|\theta - \theta'\| \leq \Delta$. Using this and together with the previous inequality, we obtain that

$$|\hat{R}(\theta) - R(\theta)| \leq |\hat{R}(\theta') - R(\theta')| + 2L\Delta.$$

   By first taking maximum over the Δ-net over the right hand side, and next taking a supremum on the left hand side, we obtain

$$\sup_{\theta\in\Theta} |\hat{R}(\theta) - R(\theta)| \leq \max_{\theta\in\mathcal{N}_\Delta} |\hat{R}(\theta) - R(\theta)| + 2L\Delta.$$

3. **Union bound:** Now that we discretized the parameter space, we can apply the union bound. Using the previous display, we write

$$\begin{aligned}\mathbb{P}\bigg(\sup_{\theta\in\Theta}|\hat{R}(\theta)-R(\theta)|\geq\epsilon/2\bigg)\leq&\mathbb{P}\bigg(\max_{\theta\in\mathcal{N}_\Delta}|\hat{R}(\theta)-R(\theta)|\geq\epsilon/2-2L\Delta\bigg)\\ \leq&\mathbb{P}\bigg(\max_{\theta\in\mathcal{N}_\Delta}|\hat{R}(\theta)-R(\theta)|\geq\epsilon/4\bigg)=(*)\end{aligned}$$

where we let $\Delta=\epsilon/8L$ in the last step. By the union bound, we get

$$\begin{aligned}(*)=\mathbb{P}\left(\bigcup_{\theta\in\mathcal{N}_\Delta}\{|\hat{R}(\theta)-R(\theta)|\geq\epsilon/4\}\right)\leq&\sum_{\theta\in\mathcal{N}_\Delta}\mathbb{P}\Big(|\hat{R}(\theta')-R(\theta')|\geq\epsilon/4\Big),\\ \leq&2|\mathcal{N}_\Delta|\exp\left\{-\frac{n\epsilon^2}{8B^2}\right\}.\end{aligned}$$

The bound on the right hand side is quite explicit and depends on the covering number of the parameter space Θ. But we already have a bound on this covering number from the previous example as given in (5.1), that is, $|\mathcal{N}_\Delta|\leq(\frac{2R\sqrt{d}}{\Delta})^d$ where $\Delta=\epsilon/8L$. Hence, the above inequality suggests that

$$\mathbb{P}\bigg(\sup_{\theta\in\Theta}|\hat{R}(\theta)-R(\theta)|\geq\epsilon/2\bigg)\leq2\left(\frac{16RL\sqrt{d}}{\epsilon}\right)^d\exp\bigg\{-\frac{n\epsilon^2}{8B^2}\bigg\}=(**)$$

By moving all terms inside the exponent, we get

$$(**)\leq2\exp\bigg\{-\frac{n\epsilon^2}{8B^2}+d\log(16RL\sqrt{d})+d\log(\epsilon^{-1})\bigg\}.$$

At this point, we are ready to identify the convergence rate. We start by trying out $\epsilon=c\sqrt{\frac{d}{n}}$, which yields

$$(**)\leq2\exp\bigg\{-\frac{d}{8B^2}+d\log(16RL\sqrt{d})-d\log(c\sqrt{d})+d\log(c\sqrt{n})\bigg\}.$$

Notice that the second and third terms can cancel each other with a right choice of c, but the first and last terms cannot, as one is decaying with d and other is growing with $d\log(n)$. In fact, any rate slower that $\sqrt{n}$ would work here. But we can also get away with only losing a log factor.

By choosing $\epsilon=c\sqrt{\frac{d\log(n)}{n}}$, we have

$$\begin{aligned}(**)\leq&2\exp\bigg\{-\frac{c^2d\log(n)}{8B^2}+d\log(16RL\sqrt{d})-\frac{d}{2}\log\log(n)+\frac{d}{2}\log(n)-d\log(c\sqrt{d})\bigg\},\\ \leq&2\exp\{-d/2\},\end{aligned}$$

where we let $c=2(B\vee8RL)$ and $n\geq16$.

4. **Uniform convergence $\implies$ generalization:** The last step is to convert the bound on the empirical process to a bound on the excess risk. We have

$$\mathbb{P}\big(R(\hat{\theta}) - R(\theta_*) \geq \epsilon\big) \leq \mathbb{P}\left(\sup_{\theta \in \Theta} |\hat{R}(\theta) - R(\theta)| \geq \epsilon/2\right),$$

combining this with the previous result, we obtain

$$\mathbb{P}\left(R(\hat{\theta}) - R(\theta_*) \geq c\sqrt{\frac{d\log(n)}{n}}\right) \leq 2\exp\{-d/2\},$$

whenever $c = 2(B \vee 8RL)$ and $n \geq 16$.

□

6 Rademacher Complexity: Definition

So far, our quest to achieve generalization involves three key steps: 1-concentration, 2- union bound, and 3- uniform conv $\implies$ generalization. That is, we used Hoeffding's lemma to obtain a concentration result for the empirical risk. We then establish that an empirical process is small, by either handling the supremum through a union bound over either a finite function class, or using an ε-net argument to obtain a generalization bound. Lastly, using that unif. conv $\implies$ generalization, we get

$$\mathbb{P}\big(R(\hat{f}) - R(f^*) \geq \epsilon\big) \leq \mathbb{P}\Bigg(\underbrace{\sup_{f\in\mathcal{F}} |\hat{R}(f) - R(f)| \geq \frac{\epsilon}{2}}_{\text{empirical process}}\Bigg) = (*) \tag{6.1}$$

$$\leq \text{ func. of } \Bigg(\epsilon, \text{ complexity of } \mathcal{F}, n\Bigg).$$

In the following, all of the above steps will be modified– steps 1 and 3 change slightly, but 2 entirely. In the concentration step, we will obtain a concentration result directly for the empirical process (not for empirical risk), showing that it is close to its expectation. Then we will use a technique called "symmetrization" to show that the expectation of the empirical process depends on the complexity of the function class. Lastly, by using a slightly modified version of "uniform conv. $\implies$ generalization" we will obtain a generalization bound.

Rademacher complexity of the function class $\mathcal{F}$ over n samples will be denoted by $\mathfrak{R}_n(\mathcal{G})$. We will be formally defining the Rademacher complexity **later** in this section, but in the sequel it should be understood as a measure of complexity of the function class $\mathcal{G}$ over n data points.

6.1 Generalization based on Rademacher complexity

Theorem 23 (Generalization based on Rademacher complexity)**.** *Define*

$$\mathcal{G} = \{(y, x) \to \ell((y, x), f) \text{ where } f \in \mathcal{F}\},$$

and assume ℓ is bounded, $\ell \in [0, B]$, and $(x_i, y_i) \overset{iid}{\sim} p$. Then with probability at least $1 - \delta$,

$$R(\hat{f}) - R(f) \leq 4\mathfrak{R}_n(\mathcal{G}) + B\sqrt{\frac{2\log(2/\delta)}{n}}. \tag{6.2}$$

Before proving the above theorem, we make a few remarks.

Remarks

- As stated before, Rademacher complexity measures the complexity of the function class $\mathcal{G}$ over n data points. It should converge to zero as n gets large, and this determines the generalization error rate.
- It is important to note that in the bound (6.2), $\mathfrak{R}_n(\mathcal{G})$ is the Rademacher complexity of the function class $\mathcal{G}$, not $\mathcal{F}$. We will connect this to $\mathcal{F}$ later.
- Although what we care about is bounding the generalization error, the above bound is obtained for the empirical process, and the technique used here has applications beyond generalization.

Proof. Our proof strategy is as follows.

We will conclude our proof with a modified version of the "uniform conv. $\implies$ generalization". For this, we start by splitting the inequality (6.1) into two components:

$$* \leq \mathbb{P}\left(\sup_{f \in \mathcal{F}} \; \hat{R}(f) - R(f) \geq \frac{\epsilon}{2} \right) + \mathbb{P}\left(\sup_{f \in \mathcal{F}} \; R(f) - \hat{R}(f) \geq \frac{\epsilon}{2} \right). \tag{6.3}$$

In the remainder of this proof, we will focus on bounding the first term on the right hand side above. But an equivalent bound can be shown for the second term.

The proof relies on three key steps as before: 1-concentration, 2-symmetrization, and 3- uniform conv. $\implies$ generalization.

1. **Concentration**: Previously, we relied on Hoeffding's inequality to obtain a concentration bound for the empirical risk. In the sequel, we will use a stronger theorem in order to obtain a concentration result directly for the empirical process.

   Lemma 24 (McDiarmid's inequality)**.** *Let $g : \mathcal{Z} \times ... \times \mathcal{Z} \to \mathbb{R}$ be a function satisfying the bounded difference property*

   $$|g(z_1, \ldots, z_j, \ldots, z_n) - g(z_1, \ldots, z_j', \ldots, z_n)| \leq c_j$$

   Then for independent random variables $z_1, z_2 \ldots, z_n$, we have

   $$\mathbb{P}\Big(g(z_1, \ldots, z_n) - \mathbb{E}[g(z_1, \ldots, z_n)] \geq \epsilon \Big) \leq \exp\left\{ \frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2} \right\}.$$

   This lemma is stronger than the previously used concentration arguments. Indeed, Hoeffding's inequality can be derived by using the above lemma.

   Example: [Hoeffding's inequality] Suppose $z_1, \ldots, z_n$ are independent random variables that are bounded almost surely $a_i \leq z_i \leq b_i$. We define g as their average and verify the bounded difference property

   $$\begin{aligned} g(z_1, ..., z_n) = S_n &= \frac{1}{n} \sum_{i=1}^n z_i \\ |g(z_1, \ldots, z_j, \ldots, z_n) - g(z_1, \ldots, z_j', \ldots, z_n)| &\leq \frac{1}{n} |z_j - z_j'| \\ &\leq \frac{b_j - a_j}{n}. \end{aligned}$$

   By the McDiarmid's inequality, we obtain

   $$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n z_i - \mathbb{E}[z_1] \geq \epsilon \right) \leq \exp\left\{ \frac{-2\epsilon^2 n}{\sum_j (b_j - a_j)^2} \right\}.$$

   We continue the proof by recalling our goal: We need to bound the empirical process in (6.3).

For this, we let the g function from McDiarmid's inequality be the function of interest.

$$\begin{aligned} g(z_1, \ldots, z_n) &= \sup_{f \in \mathcal{F}} \hat{R}(f) - R(f) \\ &= \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(\underbrace{(x_i, y_i)}_{z_i}, f) - \mathbb{E}[\ell(\underbrace{(x, y)}_{z}, f)] \\ &= \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(z_i, f) - E[\ell(z, f)] \end{aligned}$$

Notice that, in order to ease the notation, we denoted the data pairs (x_i, y_i) as z_i. We first verify the bounded difference property.

$$\begin{aligned} &|g(z_1, \ldots, z_j, \ldots, z_n) - g(z_1, \ldots, z'_j, \ldots, z_n)| \\ &= \left| \sup_{f \in \mathcal{F}} \left[\frac{1}{n} \sum_{i=1}^{n} \ell(z_i, f) - E[\ell(z, f)] \right] - \sup_{f \in \mathcal{F}} \left[\frac{1}{n} \sum_{i=1}^{n} \ell(z_i, f) - E[\ell(z, f)] - \frac{1}{n} \Big\{ \ell(z_j, f) - \ell(z'_j, f) \Big\} \right] \right| \\ &\underbrace{\leq}_{\text{(i)}} \sup_{f \in \mathcal{F}} \frac{1}{n} \big| \ell(z_j, f) - \ell(z'_j, f) \big| \\ &\leq \frac{B}{n} \end{aligned}$$

where the inequality (i) follows from the following simple fact. For function F, G, we have

Fact 25.

$$\Big| \sup_x F(x) - \sup_x G(x) \Big| \leq \sup_x \Big| F(x) - G(x) \Big|.$$

Hence, by the McDiarmid's inequality, we obtain

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} \hat{R}(f) - R(f) \geq t + \overbrace{\mathbb{E}\Big[\sup_{f \in \mathcal{F}} \hat{R}(f) - R(f) \Big]}^{\text{Need to show is small}} \right) \underbrace{\leq}_{\text{McDiarmid's}} \exp\left\{ \frac{-2nt^2}{B^2} \right\}.$$

It is worth highlighting the following again. Previously, we have focused on the concentration over $n^{-1} \sum_i \ell(z_i, f) - E[\ell(z, f)]$, followed by a union bound, but now we are looking at the concentration of the supremum directly $\sup_{f \in \mathcal{F}} \; n^{-1} \sum_i \ell(z_i, f) - E[\ell(z, f)]$. The above bound is looking good for our goal except that we need to control the additional term that is the expected value of the empirical process. This will be done by the symmetrization argument.

6.2 Symmetrization

We start with a simple argument. If X, X' are iid r.v.'s then $X \overset{d}{=} X'$. If g is a function then $g(X) \overset{d}{=} g(X')$. Further,

$$\begin{aligned} g(X) - g(X') &\overset{d}{=} g(X') - g(X) \\ &\overset{d}{=} -1 \cdot [g(X) - g(X')] \\ &\overset{d}{=} \sigma \cdot (g(X) - g(X')), \end{aligned}$$

where σ is a Rademacher random variable, i.e. $\mathbb{P}(\sigma = +1) = \mathbb{P}(\sigma = -1) = 1/2$, which is independent of X, X'. This argument is very useful (as we will see soon), and termed as symmetrization.

2. **Symmetrization**: Denote our dataset as: $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\} = \{z_1, \ldots, z_n\}$. Introduce a random copy of the dataset $\mathcal{D}' = \{z'_1, \ldots, z'_n\}$, where z_i and z'_i are iid. This new dataset is called the ghost dataset. Now that we have two datasets $\mathcal{D}, \mathcal{D}'$, there are also two empirical risks $R(f; \mathcal{D})$ and $R(f; \mathcal{D}')$ where we denote their dependence on the corresponding dataset. That is,

$$\hat{R}(f; \mathcal{D}) = \frac{1}{n} \sum_{i=1}^{n} \ell(z_i, f) \quad \text{and} \quad \hat{R}(f; \mathcal{D}') = \frac{1}{n} \sum_{i=1}^{n} \ell(z'_i, f).$$

The population risk will be identical for these datasets,

$$R(f) = E[\ell(z, f)] = E[\hat{R}(f, \mathcal{D})] = E[\hat{R}(f, \mathcal{D}')].$$

We write,

$$\begin{aligned}
\mathbb{E}\left[\sup_{f \in \mathcal{F}} \; \hat{R}(f) - R(f)\right] &= \mathbb{E}\left[\sup_{f \in \mathcal{F}} \; \hat{R}(f; \mathcal{D}) - \mathbb{E}[\hat{R}(f; \mathcal{D}')]\right] \\
&= \mathbb{E}\left[\sup_{f \in \mathcal{F}} \; \Big\{\hat{R}(f; \mathcal{D}) - \underbrace{\mathbb{E}[\hat{R}(f; \mathcal{D}')|D]}_{D \overset{\text{indep}}{\sim} D'}\Big\}\right] \\
&= \mathbb{E}\left[\sup_{f \in \mathcal{F}} \; \Big\{\mathbb{E}\Big[\hat{R}(f; \mathcal{D}) - \hat{R}(f; \mathcal{D}')|D\Big]\Big\}\right] \\
&\underbrace{\leq}_{\text{(i)}} \mathbb{E}\left[\mathbb{E}\Big[\sup_{f \in \mathcal{F}} \; \Big\{\hat{R}(f; \mathcal{D}) - \hat{R}(f; \mathcal{D}')\Big\}|D\Big]\right] \\
&\underbrace{=}_{\text{(ii)}} \mathbb{E}\left[\sup_{f \in \mathcal{F}} \; \hat{R}(f; \mathcal{D}) - \hat{R}(f; \mathcal{D}')\right] \\
&= \mathbb{E}\left[\sup_{f \in \mathcal{F}} \; \frac{1}{n} \sum_{i=1}^{n} \Big(\ell(z_i, f) - \ell(z'_i, f)\Big)\right] \\
&= \mathbb{E}\left[\sup_{f \in \mathcal{F}} \; \frac{1}{n} \sum_{i=1}^{n} \sigma_i \Big(\ell(z_i, f) - \ell(z'_i, f)\Big)\right] \\
&\underbrace{\leq}_{\text{(iii)}} \mathbb{E}\left[\sup_{f \in \mathcal{F}} \; \frac{1}{n} \sum_{i=1}^{n} \sigma_i \ell(z_i, f) + \sup_{f \in \mathcal{F}} \; \frac{1}{n} \sum_{i=1}^{n} -\sigma_i \ell(z'_i, f)\right] \\
&\underbrace{=}_{\sigma_i \overset{d}{=} -\sigma_i} 2 \cdot \mathbb{E}\left[\sup_{f \in \mathcal{F}} \; \frac{1}{n} \sum_{i=1}^{n} \sigma_i \ell(z_i, f)\right]
\end{aligned}$$

where (i) follows from $\mathbb{E}[\sup] \geq \sup \mathbb{E}$, (ii) follows from the law of iterated expectation, and (iii) follows from the following fact.

Fact 26. $\sup_x \{F(x) + G(x)\} \leq \sup_x F(x) + \sup_x G(x)$.

The bound we obtained through the above steps is simply,

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \hat{R}(f) - R(f)\right] \leq 2\mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \ell(z_i, f)\right], \tag{6.4}$$

and the final bound doesn't include the ghost dataset at all!

Next, we define the Rademacher complexity.

Definition 27 (Rademacher complexity)**.** *For a function class $\mathcal{G} = \{g : \mathcal{Z} \to \mathbb{R}\}$, the Rademacher complexity is defined as,*

$$\mathfrak{R}_n(\mathcal{G}) = \mathbb{E}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i g(z_i)\right],$$

where $z_i \overset{iid}{\sim} p$ are data points, and $\sigma_i \overset{iid}{\sim}$ Rademacher r.v.'s independent from the dataset. Furthermore, the empirical Rademacher complexity is defined as,

$$\hat{\mathfrak{R}}_n(\mathcal{G}) = \mathbb{E}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i g(z_i) | z_{1:n}\right].$$

Therefore, defining the function class $\mathcal{G}$ as

$$\mathcal{G} = \{g : z \to \ell(z, f) \text{ such that } f \in \mathcal{F}\},$$

the bound in (6.4) can be written as

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \hat{R}(f) - R(f)\right] \leq 2\mathfrak{R}_n(\mathcal{G}).$$

3. **Uniform convergence $\implies$ generalization** (yet again): We now have the necessary building blocks to construct our goal of generalization. We write out generalization bound as

$$\mathbb{P}\big(R(\hat{f}) - R(f^*) \geq \epsilon\big) \leq \mathbb{P}\left(\sup_{f\in\mathcal{F}} \hat{R}(f) - R(f) \geq \frac{\epsilon}{2}\right) + \mathbb{P}\left(\sup_{f\in\mathcal{F}} -\hat{R}(f) + R(f) \geq \frac{\epsilon}{2}\right).$$

We will obtain (already obtained) a bound on the first term on the right hand side. Similar argument yields the same bound for the second term.

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} \hat{R}(f) - R(f) \geq t + \mathbb{E}\Big[\sup_{f\in\mathcal{F}} \hat{R}(f) - R(f)\Big]\right) \leq \exp\left\{\frac{-2nt^2}{B^2}\right\}$$

Using that

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \hat{R}(f) - R(f)\right] \leq 2 \cdot \mathfrak{R}_n(\mathcal{G})$$
$$\text{where } \mathcal{G} = \{z \to \ell(z, f) \text{ where } f \in \mathcal{F}\},$$

we can write,

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} \hat{R}(f) - R(f) \geq \underbrace{t + 2\mathfrak{R}_n(\mathcal{G})}_{\epsilon/2} \right) \leq \underbrace{\exp\left\{\frac{-2nt^2}{B^2}\right\}}_{\delta}$$

Then with probability at least $1-\delta \coloneqq 1 - 2\exp\left\{\frac{-2nt^2}{B^2}\right\}$, we have

$$\hat{R}(f) - R(f) \leq \epsilon/2 = t + 2\mathfrak{R}_n(\mathcal{G})$$
$$\text{for} \quad t = B\sqrt{\frac{\log(2/\delta)}{2n}}.$$

Similar argument holds for $\mathbb{P}\Big(\sup_{f\in\mathcal{F}} -\hat{R}(f) + R(f) \geq \frac{\epsilon}{2}\Big)$. Therefore, we obtain

$$\mathbb{P}\Big(R(\hat{f}) - R(f^*) \geq 4\mathfrak{R}_n(\mathcal{G}) + B\sqrt{\frac{2\log(2/\delta)}{n}}\Big) \leq 1-\delta,$$

which concludes the proof.

□

Outline of the above proof is as follows.

1. Concentration of the empirical process

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} \hat{R}(f) - R(f) \leq t + \mathbb{E}\Big[\sup_{f\in\mathcal{F}} \hat{R}(f) - R(f)\Big] \right) \underbrace{\leq}_{\text{McDiarmid's}} \exp\Big\{\frac{-2nt^2}{B^2}\Big\}.$$

2. Symmetrization: For $\mathcal{G} = \{z \to \ell(z,f) \text{ where } f \in \mathcal{F}\}$,

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \hat{R}(f) - R(f) \right] \leq 2\cdot\mathbb{E}\left[\sup_{f\in\mathcal{F}} n^{-1}\sum_{i=1}^{n} \sigma_i \ell(z_i, f) \right] = 2\cdot\mathfrak{R}_n(\mathcal{G}).$$

3. Uniform convergence implies generalization

$$\mathbb{P}\big(R(\hat{f}) - R(f^*) \geq \epsilon\big) \leq \mathbb{P}\left(\sup_{f\in\mathcal{F}} \hat{R}(f) - R(f) \geq \frac{\epsilon}{2}\right) + \mathbb{P}\left(\sup_{f\in\mathcal{F}} -\hat{R}(f) + R(f) \geq \frac{\epsilon}{2}\right).$$

where we set $\epsilon/2 = t + 2\mathfrak{R}_n(\mathcal{G})$ and $\delta = \exp\left\{\frac{-2nt^2}{B^2}\right\}$ and solve for these quantities.

7 Rademacher Complexity: Properties & Applications

From now on, we will rely on the following (informal) inequality to establish generalization. With probability at least $1-\delta$,

$$R(\hat{f}) - R(f) \leq 4\mathfrak{R}_n(\mathcal{G}) + B\sqrt{\frac{2\log(2/\delta)}{n}}. \tag{7.1}$$

where $\mathcal{G} = \{(y,x) \to \ell((y,x),f) \text{ where } f \in \mathcal{F}\}$. Formal statement is given in Theorem 23. Key observation is that, in order to achieve generalization, we only need to find an upper bound to Rademacher complexity $\mathfrak{R}_n(\mathcal{G})$ that decays with n.

7.1 Properties of Rademacher complexity

Below, we state some properties of Rademacher complexity.

1. Monotonicity: if $\mathcal{F}_1 \subseteq \mathcal{F}_2$ then $\mathfrak{R}_n(\mathcal{F}_1) \leq \mathfrak{R}_n(\mathcal{F}_2)$
2. Linear combination: if $\mathcal{F}_1 + \mathcal{F}_2 = \{f_1 + f_2 : f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2\}$ then $\mathfrak{R}_n(\mathcal{F}_1 + \mathcal{F}_2) = \mathfrak{R}_n(\mathcal{F}_1) + \mathfrak{R}_n(\mathcal{F}_2)$
3. Scaling: if $c \in \mathbb{R}$ and $c\mathcal{F} = \{cf : f \in \mathcal{F}\}$ then $\mathfrak{R}_n(c\mathcal{F}) = |c|\mathfrak{R}_n(\mathcal{F})$
4. Convex Hull of $\mathcal{F}$: if $|\mathcal{F}| < \infty$ then $\mathfrak{R}_n(\text{convex-hull}(\mathcal{F})) = \mathfrak{R}_n(\mathcal{F})$

The above properties follow from the definition of Rademacher complexity, and their proof is left to reader as an exercise.

We notice that in the generalization bound 7.1, the Rademacher complexity of the function class $\mathcal{G}$ plays a key role. Our objective is to connect this bound to the complexity of the hypothesis functions $\mathcal{F}$. The following strong result serves to that purpose.

Lemma 28 (**Talagrand's contraction principal**). *Let g be an L-Lipschitz continuous function, and $\mathcal{F}$ is a function class. Then,*

$$\mathfrak{R}_n(g \circ \mathcal{F}) \leq L \cdot \mathfrak{R}_n(\mathcal{F}).$$

Proof of the above lemma is involved and skipped in class. We emphasize that Talagrand's lemma can be used to map the Rademacher complexity of $\mathcal{G}$, to that of $\mathcal{F}$ which is known for certain function classes. We first go over an example to demonstrate how to use the above result.

Example. [Support Vector Machines] In our first example, we visit a classical learning algorithm. As before, we denote our data pairs with $z = (x,y)$ and $y \in \{\pm 1\}$, $x \in \mathbb{R}^d$ and loss $\ell(z,f) = \max\{0, 1 - y \cdot f(x)\}$ which is often called as the hinge-loss. Let's define the function $\phi(s) = \max\{0, 1-s\}$, and notice that $\ell(z,f) = \phi(yf(x))$ and ϕ is 1-Lipschitz continuous.

Recall that the generalization bound we obtained in (7.1) relies on the Rademacher complexity of the loss class $\mathfrak{R}_n(\mathcal{G})$, where $\mathcal{G} = \{z = (y,x) \to \phi(yf(x)), f \in \mathcal{F}\}$. In order to connect this to the complexity of $\mathcal{F}$, we define $\mathcal{H} = \{z = (y,x) \to yf(x), f \in \mathcal{F}\}$, and we notice that $\mathcal{G} = \phi \circ \mathcal{H}$. By the Talagrand's contraction principal, we can bound the Rademacher complexity of $\mathcal{H}$ as

$$\mathfrak{R}_n(\mathcal{G}) \leq 1 \cdot \mathfrak{R}_n(\mathcal{H}),$$

since ϕ is 1-Lipschitz. But

$$\begin{aligned}
\mathfrak{R}_n(\mathcal{H}) &= \mathbb{E}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i h(z_i)\right] \\
&= \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i y_i f(x_i)\right], \quad \sigma_i y_i \overset{d}{=} \sigma_i \ \ \{**\} \\
&= \mathbb{E}\left[\sup_{f\in\mathcal{F}} n^{-1}\sum_{i}\sigma_i f(x_i)\right] \\
&= \mathfrak{R}_n(\mathcal{F}).
\end{aligned}$$

Note that in the second line of equality, $y_i \overset{d}{=} \sigma_i y_i$ comes from the fact that $\sigma_i y_i \perp\!\!\!\perp x_i$ even though $y_i \not\perp\!\!\!\perp x_i$ (verify this).

Therefore we can conclude that if we characterize $\mathfrak{R}_n(\mathcal{F})$, then we can characterize $\mathfrak{R}_n(\mathcal{G})$. This still doesn't complete the whole picture, but we are making progress.

Example. [Smooth relaxations to 0-1 loss] Smooth surrogate relaxations to 0-1 loss are commonly employed in machine learning. The basic idea is that we would like to minimize the misclassification error which is based on the 0-1 loss, but in practice we cannot minimize this ill-behaved loss function due to its discontinuous behavior. For this reason, we consider surrogate losses to 0-1 loss which are its smoothed versions. In this example, we will see how using a surrogate loss may lead to a worsened generalization error.

We consider a binary classification problem where we denote the data as $z = (y, x)$, with class labels $y \in \{\pm 1\}$, and $f \in \mathcal{F}$, then the 0-1 loss function is given as $\mathbb{1}_{\{yf(x)\leq 0\}}$ and can be equivalently written as follows.

$$\ell_0(z, f) \triangleq \ell_0(yf(x)) \ \text{ where } \ \ell_0(s) = \begin{cases} 1 & \text{if } s < 0, \\ 0 & \text{if } s \geq 0. \end{cases}$$

We will assume that the product of response and prediction satisfies the following property.

Assumption 1. *We assume the following holds*

$$\exists C > 0, \forall f \in \mathcal{F} \ \ \mathbb{P}(0 \leq yf(x) \leq \tau) \leq C\tau.$$

for small τ. The assumption is simply stating that the probability of misclassifying a sample gets smaller with smaller margin. This assumption is not very transparent as is and it can unpacked for certain data distributions. For example for $x \sim \mathcal{N}(0, 1)$ and $y = \text{sign}(x)$, we get $\mathbb{P}(0 \leq yf(x) \leq \tau) \leq \sqrt{2/\pi}\tau + O(\tau^3)$. But for now, let's work with this assumption to obtain our result.

Let's introduce a surrogate loss function which will serve as a relaxation to 0-1 loss.

$$\ell_\tau(s) = \begin{cases} 1 & \text{if } s < 0 \\ 1 - \frac{s}{\tau} & \text{if } 0 \leq s < \tau \\ 0 & \text{if } s \geq \tau \end{cases}$$

Another motivation for using the above loss is that the 0-1 loss function assignes the same penalty for low and high confidence predictions. Instead we would like to encourage higher confidence predictions with τ-margin sensitivity.

The loss function $\ell_\tau(s)$ is Lipschitz continuous with constant $L = 1/\tau$. Denote

$$\hat{f}_\tau = \operatorname*{argmin}_{f \in \mathcal{F}} \ \hat{R}_\tau(f)$$
$$f^*_\tau = \operatorname*{argmin}_{f \in \mathcal{F}} \ R_\tau(f)$$

With this notation, we have f^*_0 as the minimizer of the population risk $R_0(f)$. We make the following observations:

1. By Theorem 7.1, with probability at least $1 - \delta$, we have

$$R_\tau(\hat{f}_\tau) \leq R_\tau(f^*_\tau) + 4\mathfrak{R}_n(\mathcal{G}) + \sqrt{\frac{2\log(2/\delta)}{n}}$$

   where $\mathcal{G} = \{z \to \ell_\tau(z, f) \ : \ f \in \mathcal{F}\}$.

2. By an argument similar to the one in previous example, we have $\mathfrak{R}_n(\mathcal{G}) \leq (1/\tau)\mathfrak{R}_n(\mathcal{F})$ (by Talagrand's contraction principal).

3. Since $\ell_\tau \geq \ell_0$, we have $R_\tau \geq R_0$ and $\hat{R}_\tau \geq \hat{R}_0$.

4. Also, by the Assumption 1, we have

$$\sup_{f \in \mathcal{F}} R_\tau(f) - R_0(f) \leq \sup_{f \in \mathcal{F}} \mathbb{P}(0 \leq yf(x) \leq \tau) \leq C\tau$$

   for small τ. This allows us to write

$$R_\tau(f^*_\tau) \leq R_\tau(f^*_0) \leq R_0(f^*_0) + C\tau.$$

Combining the above observations, we can write with probability $1 - \delta$

$$R_0(\hat{f}_\tau) - R_0(f^*_0) \leq C\tau + \frac{4}{\tau}\mathfrak{R}_n(\mathcal{F}) + \sqrt{\frac{2\log(2/\delta)}{n}}.$$

Notice the trade-off on τ in the above bound. As we will see in the next result, we typically have $\mathfrak{R}_n(\mathcal{F}) = \mathcal{O}(1/\sqrt{n})$. Therefore the above is of order

$$R_0(\hat{f}_\tau) - R_0(f^*_0) \lesssim \tau + \frac{1}{\tau\sqrt{n}},$$

which yields a rate of $1/n^{1/4}$ after optimizing the bound over τ. Notice the sharp drop in the convergence rate, from $1/n^{1/2}$ to $1/n^{1/4}$, which is due to using a surrogate loss.

7.2 Rademacher complexity of constrained linear models

So far, we have shown that the generalization bounds can be written in terms of $\mathfrak{R}_n(\mathcal{F})$. In the following, we will show that $\mathfrak{R}_n(\mathcal{F})$ decays with n which completes the picture in terms of achieving a generalization bound.

Theorem 29 (**Rademacher Complexity of linear models**). *Define the function class of ball constrained linear models as* $\mathcal{F} = \{f(x) = \langle x, \theta\rangle, \ \|\theta\| \leq r\}$. *We have*

1. $\hat{\mathfrak{R}}_n(\mathcal{F}) \le \frac{r}{n}\sqrt{\sum_{i=1}^n \|x_i\|^2}$
2. *If* $\mathbb{E}[\|x_i\|^2] \le \kappa^2$, *then* $\mathfrak{R}_n(\mathcal{F}) \le \frac{r\kappa}{\sqrt{n}}$.

Remark. The above bound tells us that the Rademacher complexity of decays with a rate $1/\sqrt{n}$. Plugging this back in the bound (7.1), we can achieve generalization. For example, using this for linear SVMs, we obtain a generalization bound of $\mathcal{O}(1/\sqrt{n})$.

Proof. We first prove the first result. We write

$$\begin{aligned}
\widehat{\mathfrak{R}}_n(\mathcal{F}) &= \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_i \sigma_i f(x_i)\Big| x_{1:n}\right] \\
&= \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_i \sigma_i \langle\theta, x_i\rangle\Big| x_{1:n}\right] \\
&= \mathbb{E}\left[\sup_{\|\theta\|\le r} \langle\theta, \frac{1}{n}\sum_i \sigma_i x_i\rangle\Big| x_{1:n}\right], \\
&\underset{(i)}{=} r\cdot\mathbb{E}\Big[\Big\|\frac{1}{n}\sum_i \sigma_i x_i\Big\|\Big| x_{1:n}\Big] \\
&\underset{(ii)}{\le} r\cdot\mathbb{E}\Big[\Big\|\frac{1}{n}\sum_i \sigma_i x_i\Big\|^2\Big| x_{1:n}\Big]^{1/2} \\
&= \frac{r}{n}\mathbb{E}\left[\sum_i \sigma_i^2\|x_i\|^2 + \sum_{i\ne j}\sigma_i\sigma_j\langle x_i, x_j\rangle | x_{1:n}\right]^{1/2}, \\
&= \frac{r}{n}\left(\sum_i \|x_i\|^2\right)^{1/2}
\end{aligned}$$

where step (i) follows from the dual formulation of ℓ_2-norm, i.e., $\sup_{\|\theta\|=1} \langle\theta, u\rangle = \|u\|$, step (ii) follows from Jensen's inequality.

For the second part, we write

$$\begin{aligned}
\mathfrak{R}_n(\mathcal{F}) = \mathbb{E}\big[\widehat{\mathfrak{R}}_n(\mathcal{F})\big] &\le \mathbb{E}\left[\frac{r}{n}\sqrt{\sum_i \|x_i\|^2}\right] \\
&\le \frac{r}{n}\sqrt{\sum_i \mathbb{E}[\|x_i\|^2]} \\
&\le \frac{r}{\sqrt{n}}\kappa,
\end{aligned}$$

where the second inequality follows from Jensen's inequality, and the last one follows from the assumption $\mathbb{E}[\|x_i\|^2] \le \kappa^2$. □

We should remark that κ is typically of order $\sqrt{d}$, so the generalization bound we get is like $\sqrt{d/n}$ as expected.

7.3 Massart's Finite Lemma

We have already worked out the generalization performance of finite function classes in Section 4.2. But in this section, we would like to use our new tool, the Rademacher complexity for the same

purpose. This will allow us to compare bounds obtained through different techniques. The following result is very useful in that respect.

We introduce the Massart's Lemma that will be used throughout next few lectures.

Lemma 30 (Massart's Finite Lemma). *Suppose that $\mathcal{F}$ satisfies $\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n f(z_i)^2 \le \kappa^2$, then the empirical Rademacher complexity of the function class is bounded, i.e.*

$$\hat{\mathfrak{R}}_n(\mathcal{F}) \le \kappa\sqrt{\frac{2\log|\mathcal{F}|}{n}}$$

Remark.

1. The above bound is only useful (for now) when $|\mathcal{F}| < \infty$.
2. When the loss is bounded by B, the condition above is immediately satisfied for $\kappa = B$.
3. Plugging this into (7.1), we get a generalization bound with probability at least $1-\delta$
   - by Rademacher Complexity: $4B\sqrt{\frac{2\log(|\mathcal{F}|)}{n}} + B\sqrt{\frac{2\log(2/\delta)}{n}}$
   - by union bound: $B\sqrt{\frac{2\log(|\mathcal{F}|)}{n} + \frac{2\log(1/\delta)}{n}}$.

   Although these two bounds have the same rate of convergence, we notice that the latter bound is slightly tighter.
4. Perhaps the most important observation we can make is that, the function class $\mathcal{F}$ enters the above bound only through function evaluations over the data points $z_{1:n}$. This observation will be crucial in the next section.

Proof. Note that throughout this proof, we denote data with $z_{1:n} = \{z_1, \dots, z_n\}$. We will first obtain a bound for $\exp\{t\cdot\hat{\mathfrak{R}}_n(\mathcal{F})\}$, and convert this to a bound on $\hat{\mathfrak{R}}_n(\mathcal{F})$.

$$\begin{aligned}
\exp\{t\cdot\hat{\mathfrak{R}}_n(\mathcal{F})\} &= \exp\left\{t\cdot\mathbb{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n\sigma_i f(z_i)\Big|z_{1:n}\right]\right\} \qquad (7.2)\\
&\le \mathbb{E}\left[\exp\left\{t\cdot\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n\sigma_i f(z_i)\right\}\Big|z_{1:n}\right] \quad \text{(by Jensens's inequality)}\\
&= \mathbb{E}\left[\sup_{f\in\mathcal{F}}\exp\left\{t\cdot\frac{1}{n}\sum_{i=1}^n\sigma_i f(z_i)\right\}\Big|z_{1:n}\right] \quad \text{(sup on a monotone transformation)}\\
&\underset{(*)}{\le} \sum_{f\in\mathcal{F}}\mathbb{E}\left[\exp\left\{t\cdot\frac{1}{n}\sum_{i=1}^n\sigma_i f(z_i)\right\}\Big|z_{1:n}\right] \quad (\mathcal{F}\text{ is finite and exp() is positive})\\
&= \sum_{f\in\mathcal{F}}\prod_{i=1}^n M\Big(\tfrac{t}{n}f(z_i)\Big) \qquad M_\sigma(t)\text{ is the MGF of }\sigma\text{, i.e.,}\\
M_\sigma(t) &= \mathbb{E}\big[\exp\{t\sigma\}\big|z_{1:n}\big] = \cosh(t).
\end{aligned}$$

In the above derivation, in step $(*)$, we replaced sup over a set with a summation over that set. It is important to pay attention to this step as in the next section, we will obtain a general bound by simply tightening this inequality.

We proceed by noticing that $x^2/2 \geq \log\cosh(x)$ (check this!) which implies $\exp\{x^2/2\} \geq \cosh(x)$. Therefore, we can write

$$\begin{aligned}
\sum_{f\in\mathcal{F}}\prod_{i=1}^{n} M\Big(\tfrac{t}{n}f(z_i)\Big) &\leq \sum_{f\in\mathcal{F}}\prod_{i=1}^{n}\exp\left\{\frac{t^2}{2n^2}f(z_i)^2\right\} \\
&= \sum_{f\in\mathcal{F}}\exp\left\{\frac{t^2}{2n}\underbrace{\frac{1}{n}\sum_{i=1}^{n}f(z_i)^2}_{\leq\kappa^2}\right\} \\
&\leq |\mathcal{F}|\exp\left\{\frac{t^2}{2n}\kappa^2\right\}.
\end{aligned}$$

The final bound we obtained can be written as

$$\begin{aligned}
\exp\left\{t\cdot\hat{\mathfrak{R}}_n(\mathcal{F})\right\} &\leq |\mathcal{F}|\exp\left\{\frac{t^2\kappa^2}{2n}\right\} \\
\implies \hat{\mathfrak{R}}_n(\mathcal{F}) &\leq \frac{\log|\mathcal{F}|}{t} + \frac{t\kappa^2}{2n}
\end{aligned}$$

which holds for all $t \geq 0$. By optimizing over t, we will obtain the final result. That is, differentiating the RHS above with respect to t and solving for the optimal value gives $2\kappa\sqrt{\log|\mathcal{F}|/2n}$. $\square$

8 Combinatorial Measures of Complexity

By a careful inspection of the Massart's Finite Lemma and its proof, we notice that the functions that belong to our function class $\mathcal{F}$ enter the bounds only through their evaluation at the data points. That is, if the functions have bounded second moment under the empirical distribution over the data set, then Rademacher complexity decays with n. We will make use of this observation throughout this section.

8.1 Shattering Coefficient

Above, in our proof of Massart's Lemma, we flagged the inequality in (7.2), the step $(*)$ as the point at which we appealed to $|\mathcal{F}| < \infty$ to convert $\sup_{f\in\mathcal{F}}$ to a summation $\sum_{f\in\mathcal{F}}$.

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \exp\left\{t\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(z_i)\right\}\Big|z_{1:n}\right] \leq \mathbb{E}\left[\sum_{f\in\mathcal{F}} \exp\left\{t\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(z_i)\right\}\Big|z_{1:n}\right]. \tag{8.1}$$

We notice that $f \in \mathcal{F}$ enters this bound only through $f(z)$ for $z \in \mathcal{Z}$. As in our next example, $\mathcal{F}$ can be infinitely large as long as it has finite behavior over $\mathcal{Z}$.

Example. Let's assume that we have integer data points, $z \in \mathcal{Z} \subset \mathbb{Z}$, and the function class is given as $\mathcal{F} = \{z \to \sin(z\pi k), k \in \mathbb{N}\}$. We notice that even though $|\mathcal{F}| = \infty$, clearly $f(z) = 0$ for $\forall f \in \mathcal{F}$ and $z \in \mathcal{Z}$.

This behavior is not at all uncommon. Especially when we are working with loss functions with finite range, we always have finitely many function behavior over data. An example to this case is the 0-1 loss.

Example. Let's assume we are working with 0-1 loss function, and the loss class that enter the Rademacher complexity-based generalization bound is given as $\mathcal{G} = \{z \to \ell(z, f), f \in \mathcal{F}\}$. Then over data $z_1, \ldots, z_n$ we can have at most $|\mathcal{G}| = 2^n$ different assignments for the vectors $[f(z_1), ..., f(z_n)]$.

The above argument is not enough. Assume that we can replace $|\mathcal{F}|$ in the Massart's Finite Lemma, with the above exponential number 2^n. The bound on Rademacher complexity becomes $\mathcal{O}(1)$, which doesn't yield generalization. Of course, we would require sub-exponential behaviour over the data to have a useful bound which in turn yields generalization.

Let's modify the inequality (8.1) a little bit. We write

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \exp\left\{t\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(z_i)\right\}\Big|z_{1:n}\right] = \mathbb{E}\left[\sup_{[f_1\ldots f_n]\in\{[f(z_1)\ldots f(z_n)]:f\in\mathcal{F}\}} \exp\left\{t\frac{1}{n}\sum_{i=1}^{n}\sigma_i f_i\right\}\Big|z_{1:n}\right]. \tag{8.2}$$

We define the shattering coefficient as follows.

Definition 31 (Shattering Coefficient). *For $\mathcal{F} = \{f : \mathcal{Z} \to \mathcal{Y}\}$, define*

$$s(\mathcal{F}, n) = \max_{z_1,\ldots,z_n\in\mathcal{Z}} \Big|\Big\{\big[f(z_1)\cdots f(z_n)\big] : f \in \mathcal{F}\Big\}\Big|.$$

The term inside our set $\{\cdot\}$ is counting how many different configurations of the vector $[f(z_1)\ldots f(z_n)]$ are possible.

We pick up from the inequality (8.2) and write

$$\begin{aligned}
\mathbb{E}\left[\sup_{f\in\mathcal{F}}\exp\left\{t\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(z_i)\right\}|z_{1:n}\right] &= \mathbb{E}\left[\sup_{[f_1\dots f_n]\in\{[f(z_1)\dots f(z_n)]:f\in\mathcal{F}\}}\exp\left\{t\frac{1}{n}\sum_{i=1}^{n}\sigma_i f_i\right\}|z_{1:n}\right] \\
&\leq \sum_{[f_1\dots f_n]\in\{[f(z_1)\dots f(z_n)]:f\in\mathcal{F}\}}\mathbb{E}\left[\exp\left\{t\cdot\frac{1}{n}\sum_{i=1}^{n}\sigma_i f_i\right\}\right] \\
&= \sum_{[f_1\dots f_n]\in\{[f(z_1)\dots f(z_n)]:f\in\mathcal{F}\}}\prod_{i=1}^{n}\underbrace{M_\sigma\left(\frac{tf_i}{n}\right)}_{\leq\exp(t^2f_i^2/(2n^2))} \\
&\leq |\{[f(z_1)\dots f(z_n)]:f\in\mathcal{F}\}|\exp\left\{\frac{t^2\kappa^2}{2n}\right\} \\
&\leq \max_{z_1,\dots,z_n}|\{[f(z_1)\dots f(z_n)]:f\in\mathcal{F}\}|\exp\left\{\frac{t^2\kappa^2}{2n}\right\} \\
&= s(\mathcal{F},n)\exp\left\{\frac{t^2\kappa^2}{2n}\right\}.
\end{aligned}$$

These steps are exactly the same as before. The only difference is that instead of summing over the entire $\mathcal{F}$, this time we sum over different function evaluations. The max argument was applied over the data points to remove their dependence so we take an expectation on the empirical Rademacher complexity which would give us (population) Rademacher complexity.

We can write the following upgraded version of Massart's Finite Lemma.

Lemma 32 (**Modified Massart's Lemma**). *If* $\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}f(z_i)^2\leq\kappa^2$, *then*

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

Remark. Compared to Massart's Finite Lemma, $|\mathcal{F}|$ is replaced by $s(\mathcal{F},n)$.

For the 0-1 loss, we obtained a bound on shattering coefficient that grows exponentially in n, i.e. $s(\mathcal{F},n)=2^n$. Notice that exponentially growing shattering coefficient doesn't yield generalization based on the above theorem. We need at most sub-exponential growth to achieve generalization. To make this more concrete, we define the notion of a "shattered" set next.

In the sequel, we only consider Boolean functions, $f:\mathcal{Z}\to\{0,1\}$**.**

Definition 33. *Let* $\mathcal{F}$ *be a class of Boolean functions on a domain* $\mathcal{Z}$*. We say that* $\mathcal{F}$ *shatters a subset* $\mathcal{D}\subset\mathcal{Z}$ *if any function* $g:\mathcal{D}\to\{0,1\}$ *can be obtained by restricting some function* $f\in\mathcal{F}$ *to* $\mathcal{D}$*.*

Example. For the data $\mathcal{D}=\{z_1,...,z_n\}$, and $f\in\mathcal{F}$, consider the n-dimensional vectors $[f(z_1),...,f(z_n)]$. These are Boolean vectors, and if we can get every possible 2^n Boolean vectors by varying $f\in\mathcal{F}$, then $\mathcal{F}$ shatters $\mathcal{D}$. Notice that, here $\mathcal{D}$ is fixed and the Boolean vectors are changing since we change f.

For Boolean functions, if the shattering coefficient satisfies $s(\mathcal{F},n)=2^n$, this ultimately means that $\exists\mathcal{D}\subset\mathcal{Z}$ such that $\mathcal{F}$ shatters $\mathcal{D}$. It is worth restating that whenever this happens, Massart's Lemma doesn't yield generalization.

Similar to the previous section, we first justify the move from the Rademacher complexity of the loss class $\mathfrak{R}_n(\mathcal{G})$ to that of hypothesis class $\mathfrak{R}_n(\mathcal{F})$. Our next example serves as a demonstration for this.

Example. Assume that we use 0-1 loss $\ell((y,x),f) = \mathbb{1}_{\{y\neq f(x)\}} \in \{0,1\}$ and let $y \in \{\pm 1\}$ and $f : \mathcal{X} \to \{\pm 1\}$. This is not Boolean, but mapping to that case is trivial. The loss class in this case is given as $\mathcal{G} = \{(y,x) \to \mathbb{1}_{\{y\neq f(x)\}}\}$. Let (y_i, x_i) for $i = 1, 2, ..., n$ denote the samples in the data. Then notice that there is a bijection from the set of vectors $\{[f(x_1), ..., f(x_n)] : \ f \in \mathcal{F}\}$ to $\{[\ell((y_1,x_1),f), ..., \ell((y_n,x_n),f)], \ f \in \mathcal{F}\}$. This can be seen by considering the mapping from $f(x_i) \to (1 - y_i f(x_i))/2 = \ell((y_i, x_i), f)$.

Next example is a demonstration to how we calculate shattering coefficient for simple function classes.

Example. [Indicators of rays] Let's consider the function class $\mathcal{F} = \{z \to \mathbb{1}_{\{z\geq t\}} \ t \in \mathbb{R}\}$. Clearly, this function class has $|F| = |\mathbb{R}|$. But we can easily verify that $s(\mathcal{F}, n) = n + 1$. This shattering coefficient is sub-exponential and thus, the Massart's lemma will provide us with generalization. It is also worth noting that $s(\mathcal{F}, n) = 2^n$ only if $2^n = n + 1$; therefore, for $n > 1$ $\mathcal{F}$ cannot shatter any subset of size n.

8.2 Vapnik-Chervonenkis Dimension

Definition 34 (**VC-dimension of a boolean $\mathcal{F}$**). *VC dimension of $\mathcal{F}$, denoted by VC($\mathcal{F}$), is the largest cardinality of a subset $\mathcal{D} \subset \mathcal{Z}$ that can be shattered by $\mathcal{F}$.*

Remark. Notice that since we are concerned only with Boolean function classes, we can equivalently write $\mathrm{VC}(\mathcal{F})$ as

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

If the VC dimension of a function class $\mathcal{F}$ is d, i.e. $\mathrm{VC}(\mathcal{F}) = d$, this means that there exists $\mathcal{D} \subset \mathcal{Z}$ with $|\mathcal{D}| = d$ such that $\mathcal{F}$ shatters $\mathcal{D}$, i.e. $s(\mathcal{F}, d) = 2^d$, and no subset $\mathcal{D} \subset \mathcal{Z}$ of size $|\mathcal{D}| > d$ can be shattered by $\mathcal{F}$, i.e. $s(\mathcal{F}, d+1) < 2^{d+1}$.

Example. If we revisit the example for the indicators of rays, we found that $s(\mathcal{F}, n) = n + 1$ for every n. Therefore, in order to get $s(\mathcal{F}, n) = 2^n$, we need $n = 1$. Also, for any $n > 1$, we have $s(\mathcal{F}, n) < 2^n$ which proves that $\mathrm{VC}(\mathcal{F}) = 1$.

Example. [Indicators of closed intervals] Consider the following boolean function class

$$\mathcal{F} = \{z \to \mathbb{1}_{\{z\in[a,b]\}}, a < b, \ a, b \in \mathbb{R}\}.$$

We can show that for $n = 1, 2$, we have $s(\mathcal{F}, n) = 2^n$. This can be done by considering every possible 2^n cases. However for $n = 3$, for z_1, z_2, z_3, we cannot obtain $[f(z_1), f(z_2), f(z_3)] = [1, 0, 1]$ using the above function class. In fact, any other configuration is achievable which makes the shattering coefficient $s(\mathcal{F}, 3) = 7$. Therefore we conclude that $\mathrm{VC}(\mathcal{F}) = 2$.

So far we consider simple function classes where it is simple to reason about their shattering coefficient. The following lemma however, can be used together with Massart's lemma and yield a generalization bound directly related to the VC-dimension of the function class $\mathcal{F}$.

Lemma 35 (**Sauer-Shelah's Lemma**). *If VC($\mathcal{F}$) = d, then*

$$s(\mathcal{F}, n) \leq \begin{cases} 2^n & \text{if } n \leq d, \\ \left(\frac{en}{d}\right)^d & \text{if } n > d. \end{cases}$$

Remark. $\mathrm{VC}(\mathcal{F})$ is the n at which the shattering coefficient stops being exponential and starts becoming polynomial (and useful for generalization). In fact, whenever $n > \mathrm{VC}(\mathcal{F})$, by the Massart's and Sauer-Shelah's lemmas, we can write

$$\begin{aligned}\mathfrak{R}_n(\mathcal{F}) \leq & \sqrt{\frac{2\log s(\mathcal{F},n)}{n}} \leq \sqrt{\frac{2\mathrm{VC}(\mathcal{F})\log(en/\mathrm{VC}(\mathcal{F}))}{n}} \\ \leq & \sqrt{\frac{3\mathrm{VC}(\mathcal{F})\log(n)}{n}}.\end{aligned}$$

We have also seen examples that the Rademacher complexity of loss class $\mathcal{G}$, can be upper bounded by that of function class $\mathcal{F}$. Plugging this into the generalization bound obtained through Rademacher complexity (7.1), we get with probability at least $1-\delta$

$$R(\hat{f}) - R(f_*) \leq 4\sqrt{\frac{3\mathrm{VC}(\mathcal{F})\log(n)}{n}} + \sqrt{\frac{2\log(2/\delta)}{n}}.$$

Note that $\mathrm{VC}(\mathcal{G})$ in the bound can be replaced with $\mathrm{VC}(\mathcal{F})$ for binary classification problems (See the previous example).

Proof. Let $\mathcal{Z}^* = \{z_1^*, z_2^*, ..., z_n^*\}$ be such that $s(\mathcal{F},n) = |\{[f(z_1^*), ..., f(z_n^*)] \ : \ f \in \mathcal{F}\}|$, restrict $\mathcal{F}$ onto $\mathcal{Z}^*$ and call it $\mathcal{F}^*$. We notice that $\mathcal{F}^*$ is finite and its size is equal to $s(\mathcal{F},n)$ by construction, i.e. $|\mathcal{F}^*| = s(\mathcal{F},n)$. We state the following lemma due to Pajor.

Lemma 36 (Pajor's lemma)**.** *If $\mathcal{F}^*$ is a class of Boolean functions on a finite domain $\mathcal{Z}^*$, then*

$$|\mathcal{F}^*| \leq |\{\Lambda \subset \mathcal{Z}^* \ : \ \Lambda \textit{ is shattered by } \mathcal{F}^*\}|.$$

We prove the above lemma in the homework. Now, let $d^* = \mathrm{VC}(\mathcal{F}^*)$, and by Pajor's lemma, we obtain

$$s(\mathcal{F},n) \leq \sum_{i=0}^{d^*} \binom{n}{i} \tag{8.3}$$

where there right hand side above is the number of subsets of $\mathcal{Z}^*$ of size at most d^*.

But if $\Lambda \subset \mathcal{Z}^* \subset \mathcal{Z}$ is shattered by $\mathcal{F}^*$, it is also shattered by $\mathcal{F}$ since former is a restriction of the latter. Therefore, $\mathrm{VC}(\mathcal{F}^*) \leq \mathrm{VC}(\mathcal{F})$.

Now, if $d \geq n$, the right hand side of (8.3) is easily bounded by 2^n since $\mathcal{F}^*$ is class from domain of size n and it can shatter at most a set of size n. If $d < n$, then we get

$$\begin{aligned}s(\mathcal{F},n) \leq & \sum_{i=0}^{d} \binom{n}{i}, \\ = & \sum_{i=0}^{d} \binom{n}{i} \left(\frac{n}{d}\right)^i \left(\frac{d}{n}\right)^i, \\ \leq & \left(\frac{n}{d}\right)^d \sum_{i=0}^{d} \binom{n}{i} \left(\frac{d}{n}\right)^i, \\ \leq & \left(\frac{n}{d}\right)^d \left(1+\frac{d}{n}\right)^n, \\ \leq & \left(\frac{en}{d}\right)^d,\end{aligned}$$

which concludes the proof.

□

9 Chaining and Dudley's Theorem

In this lecture, we revisit some of the techniques we covered in Section 5, but there is one key difference. Before, we used the ϵ-nets to discretize the uncountable function class to be able to apply union bound and obtain a generalization bound. In this section, we will use this technique to bound the Rademacher complexity of the function class which in turn will imply generalization.

9.1 ϵ-Nets revisited

Previously in Section 5, we covered the parameter space Θ of a parametric family $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$. In this section, we cover the function class $\mathcal{F}$ directly without parametrizing it. For this, though we need to measure the difference between two different functions f and g. However, in the previous section, we also noticed that in terms of generalization we only care about the function behavior on data. Therefore, if two functions behave the same over data and differently on other points, we treat these two functions as the same. The following difference metric makes this idea concrete.

Definition 37 (Difference metric)**.** *Given a dataset $\{z_1, ..., z_n\}$, we use the following to measure the difference between two functions.*

$$d(f, g) = \left(\frac{1}{n}\sum_{i=1}^{n}(f(z_i) - g(z_i))^2\right)^{1/2}$$

New notation: The following notation will simplify the statements and will be used throughout this section. Since we only care about function behavior on data, we can encode a function $f \in \mathcal{F}$ as a n-dimensional vector, i.e.,

$$\mathbf{f} = \tfrac{1}{\sqrt{n}}[f(z_1), f(z_2), ..., f(z_n)]^\top \in \mathbb{R}^n.$$

Using this new notation, we can simplify the following quantities as

$$\|\mathbf{f}\|^2 = \frac{1}{n}\sum_{i=1}^{n} f(z_i)^2 \quad \text{and} \quad d(f, g) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}[f(z_i) - g(z_i)]^2} = \|\mathbf{f} - \mathbf{g}\|,$$

where the norms are understood to be Euclidean. We can also restate the Massart's Finite Lemma in this notation in a very compact form

$$\hat{\mathfrak{R}}_n(\mathcal{F}) \leq (\sup_{f\in\mathcal{F}} \|f\|)\sqrt{\frac{2\log(|\mathcal{F}|)}{n}}.$$

Reader should convince themselves that the above inequality is equivalent to the statement in Massart's Finite Lemma. We recall some of the key definitions of Section 5.

Definition 38. *We recall the following notions related to covering.*

- *ϵ-cover of $\mathcal{F}$ with respect to distance metric d is a set $\mathcal{N}_\epsilon = \{g_1, g_2, ...\}$ satisfying $\forall f \in \mathcal{F}$, $\exists g \in \mathcal{N}_\epsilon$ such that $d(f, g) \leq \epsilon$.*
- *Covering number of $\mathcal{F}$ is given by $N(\epsilon, \mathcal{F}, d) = \min\{|\mathcal{N}_\epsilon| : \mathcal{N}_\epsilon$ is a ϵ-cover of $\mathcal{F}\}$.*
- *Metric entropy of $\mathcal{F}$ is given by $\log N(\epsilon, \mathcal{F}, d)$.*

In general, one only needs an upper bound on the covering number. Therefore, our strategy will be to first construct a reasonable ϵ-cover of $\mathcal{F}$, then find an upper bound on its size which in turn upper bounds the covering number of $\mathcal{F}$. The following two examples demonstrate how to do this.

Example. [All functions $\mathbb{R} \to [0,1]$] Consider the function class $\mathcal{F} = \{f : \mathbb{R} \to [0,1]\}$. In order to cover this function class, we consider the 2d-grid defined by the points on the x-axis $z_{(1)}, z_{(2)}, ..., z_{(n)}$, the ordered data points, and the points on the y-axis $\{0, 2\epsilon, 4\epsilon, ...\}$. For each function $f \in \mathcal{F}$ and for each data point $z_{(i)}$ on the x-axis, we find the closest point on the grid and define a function g that passes on these points. It is easy to show that $d(f,g) \leq \epsilon$. Therefore if we include such functions g that pass on the points on this grid, we can obtain an ϵ-cover of $\mathcal{F}$. This suggests that we only need at most as many points as the number of points on this grid which can be upper bounded by

$$N(\epsilon, \mathcal{F}, d) \leq |\mathcal{N}_\epsilon| \leq (1 + 1/(2\epsilon))^n \leq (1/\epsilon)^n$$

for small ϵ. Notice that this number is exponential in n.

Example. [Non-decreasing function $\mathbb{R} \to [0,1]$] This time, consider the function class $\mathcal{F} = \{f : \mathbb{R} \to [0,1]$, and f non-decreasing$\}$. Using the same grid as before, we only need to count the number of non-decreasing functions that can be defined on this grid. This number can be upper bounded with $n^{1/\epsilon}$ which in turn implies

$$N(\epsilon, \mathcal{F}, d) \leq n^{1/\epsilon}.$$

We note that this bound is polynomial in n.

9.2 Simple discretization

In this section, we use an argument similar to Section 5 to obtain an upper bound on Rademacher complexity.

Theorem 39 (Discretization)**.** *For a function class $\mathcal{F} \subset \{f : \mathcal{Z} \to \mathbb{R}\}$, let $\kappa = \sup_{f \in \mathcal{F}} \|f\|$. Then,*

$$\forall \epsilon > 0, \quad \hat{\mathfrak{R}}_n(\mathcal{F}) \leq \kappa \sqrt{\frac{2 \log N(\epsilon, \mathcal{F}, d)}{n}} + \epsilon.$$

Remark. Before moving to proof, we make the following remarks.

1. Notice that with increasing ϵ, the first term in the right hand side of above bound decreases whereas the second term increases. This shows that there is a trade-off involving the parameter ϵ, the bound can be optimized over this parameter.

2. The above bound looks quite familiar. The first term above simply follows from the Massart's Finite Lemma whereas the second term is the discretization error.

Proof. Let $\boldsymbol{\sigma} = \frac{1}{\sqrt{n}}[\sigma_1, ..., \sigma_n]^\top$ be the vector of Rademacher random variables. We have $\|\boldsymbol{\sigma}\| = 1$, and also the empirical Rademacher complexity can be written as

$$\hat{\mathfrak{R}}_n(\mathcal{F}) = \mathbb{E}\left[\sup_{f \in \mathcal{F}} \langle \boldsymbol{\sigma}, \mathbf{f} \rangle \big| z_{1:n}\right].$$

Let $\mathcal{N}_\epsilon$ be an ϵ-net over $\mathcal{F}$. Then $\forall f \in \mathcal{F}$, $\exists g \in \mathcal{N}_\epsilon$ such that $\|\mathbf{f} - \mathbf{g}\| \leq \epsilon$. Hence, we can write for any $f \in \mathcal{F}$

$$\begin{aligned}\langle \boldsymbol{\sigma}, \mathbf{f} \rangle =& \langle \boldsymbol{\sigma}, \mathbf{g} \rangle + \langle \boldsymbol{\sigma}, \mathbf{f} - \mathbf{g} \rangle \\ \leq& \langle \boldsymbol{\sigma}, \mathbf{g} \rangle + \|\boldsymbol{\sigma}\| \|\mathbf{f} - \mathbf{g}\| \quad \text{by Cauchy-Schwartz} \\ \leq& \max_{g \in \mathcal{N}_\epsilon} \langle \boldsymbol{\sigma}, \mathbf{g} \rangle + \epsilon.\end{aligned}$$

Now that the right hand side above doesn't depend on the choice of f, we can take supremum on the left hand side and obtain

$$\sup_{f \in \mathcal{F}} \langle \boldsymbol{\sigma}, \mathbf{f} \rangle \leq \max_{g \in \mathcal{N}_\epsilon} \langle \boldsymbol{\sigma}, \mathbf{g} \rangle + \epsilon.$$

Hence, we can write

$$\begin{aligned}\hat{\mathfrak{R}}_n(\mathcal{F}) \leq& \mathbb{E}\left[\max_{g \in \mathcal{N}_\epsilon} \langle \boldsymbol{\sigma}, \mathbf{g} \rangle\right] + \epsilon, \\ =& \mathfrak{R}_n(\mathcal{N}_\epsilon) + \epsilon, \\ \leq& (\sup_{g \in \mathcal{N}_\epsilon} \|g\|) \sqrt{\frac{2 \log |\mathcal{N}_\epsilon|}{n}} + \epsilon,\end{aligned}$$

which holds for all ϵ-covers of $\mathcal{F}$. Hence we can use the best cover and conclude the proof.

□

Using this theorem on the previous examples that we calculated an upper bound on the covering number, we can obtain an explicit rate.

Example.

1. All functions $\mathbb{R} \to [0, 1]$: We had the bound $N(\epsilon, \mathcal{F}, d) \leq (1/\epsilon)^n$. By the above theorem, we obtain

   $$\hat{\mathfrak{R}}_n(\mathcal{F}) \leq \sqrt{\frac{n \log(1/\epsilon)}{n}} + \epsilon = \mathcal{O}(1).$$

   We don't get generalization in this case.

2. Non-decreasing functions $\mathbb{R} \to [0, 1]$: We had the bound $N(\epsilon, \mathcal{F}, d) \leq n^{1/\epsilon})$. By the above theorem, we obtain

   $$\hat{\mathfrak{R}}_n(\mathcal{F}) \leq \sqrt{\frac{\log(n)}{\epsilon n}} + \epsilon.$$

   Optimizing over ϵ yields the rate $\mathcal{O}\big((\log(n)/n)^{1/3}\big)$. We do get generalization in this case, but this rate is slow, and interestingly it is just an artifact of the proof technique and can be improved.

9.3 Chaining

Next, we will see a more powerful technique called "chaining" which will improve the above rate significantly. We first state the main result of this section.

Theorem 40 (Dudley's Theorem)**.** *Let $\mathcal{F}$ be a set of functions $f : \mathcal{Z} \to \mathbb{R}$. Then,*

$$\hat{\mathfrak{R}}_n(\mathcal{F}) \leq 12 \int_0^\infty \sqrt{\frac{\log N(\epsilon, \mathcal{F}, d)}{n}} d\epsilon.$$

Remark. Before proving this theorem, we make a few remarks.

1. When the function class is composed of functions with finite norm, i.e. $\sup_{f \in \mathcal{F}} \|f\| = \kappa < \infty$, then the upper boundary of the above integral is κ since beyond that point covering number $N(\epsilon, \mathcal{F}, d) = 1$.

2. We notice that the discretization error in the result of Theorem 39 is gone!

3. For the above example on non-decreasing functions, since $\sup_{f \in \mathcal{F}} \|f\| = 1$, using the Dudley's theorem, we obtain

$$\hat{\mathfrak{R}}_n(\mathcal{F}) \leq 12 \int_0^1 \sqrt{\frac{\log(n)}{\epsilon n}} d\epsilon = \mathcal{O}\left(\sqrt{\frac{\log(n)}{n}}\right).$$

   This improves the previous rate of $\mathcal{O}\big((\log(n)/n)^{1/3}\big)$ significantly.

Proof. [by chaining]

Figure 3: Chaining idea (to be added)

Let's start by the most crude ϵ-cover for our function class, i.e. set $\epsilon_0 = \sup_{f \in \mathcal{F}} \|\mathbf{f}\|$ and note that we can set $\mathcal{N}_{\epsilon_0} = \{g_0\}$ for $g_0 = 0$ which implies $N(\epsilon_0, \mathcal{F}, d) = 1$. Next, define the sequence of epsilon covers $\mathcal{N}_{\epsilon_j}$ by setting $\epsilon_j = 2^{-j}\epsilon_0$. By definition, $\forall f \in \mathcal{F}$ we can find $g_j \in \mathcal{N}_{\epsilon_j}$ that depends on the choice of f such that $\|\mathbf{f} - \mathbf{g}_j\| \leq \epsilon$.

For any $m \in \mathbb{N}$, we can write the telescopic sum

$$\mathbf{f} = \mathbf{f} - \mathbf{g}_m + \sum_{j=1}^{m} \mathbf{g}_j - \mathbf{g}_{j-1} \tag{9.1}$$

since we have $\mathbf{g}_0 = 0$. By construction, the difference sequence $\mathbf{g}_j - \mathbf{g}_{j-1}$ forms a chain that gets smaller with j (since they also get closer to $\mathbf{f}$). That is, by triangle inequality, we have

$$\|\mathbf{g}_j - \mathbf{g}_{j-1}\| \leq \|\mathbf{g}_j - \mathbf{f}\| + \|\mathbf{f} - \mathbf{g}_{j-1}\| \leq \epsilon_j + \epsilon_{j-1} = 3\epsilon_j.$$

We have

$$\begin{aligned}
\hat{\mathfrak{R}}_n(\mathcal{F}) =& \mathbb{E}\Big[\sup_{f\in\mathcal{F}}\langle\boldsymbol{\sigma},\mathbf{f}\rangle\big|z_{1:n}\Big] = \mathbb{E}\Big[\sup_{f\in\mathcal{F}}\Big\{\underbrace{\langle\boldsymbol{\sigma},\mathbf{f}-\mathbf{g}_m\rangle}_{\leq\|\boldsymbol{\sigma}\|\|\mathbf{f}-\mathbf{g}_m\|\text{ by CS}} + \langle\boldsymbol{\sigma},\sum_{j=1}^m \mathbf{g}_j-\mathbf{g}_{j-1}\rangle\Big\}\big|z_{1:n}\Big] \quad \text{by (9.1)}\\
\leq& \epsilon_m + \mathbb{E}\Big[\sup_{f\in\mathcal{F}}\langle\boldsymbol{\sigma},\sum_{j=1}^m \mathbf{g}_j-\mathbf{g}_{j-1}\rangle\big|z_{1:n}\Big] \quad \text{by CS and } \mathcal{N}_{\epsilon_m}\text{'s net property}\\
\leq& \epsilon_m + \sum_{j=1}^m \mathbb{E}\Big[\sup_{f\in\mathcal{F}}\langle\boldsymbol{\sigma},\mathbf{g}_j-\mathbf{g}_{j-1}\rangle\big|z_{1:n}\Big] \quad \text{by } \sup\Sigma\leq\Sigma\sup\\
\leq& \epsilon_m + \sum_{j=1}^m \mathbb{E}\Big[\sup_{h\in\mathcal{H}_j}\langle\boldsymbol{\sigma},\mathbf{h}\rangle\big|z_{1:n}\Big] \quad \text{where } \mathcal{H}_j=\{g_j-g_{j-1}: g_j\in\mathcal{N}_{\epsilon_j}, g_{j-1}\in\mathcal{N}_{\epsilon_{j-1}}, \|\mathbf{g}_j-\mathbf{g}_{j-1}\|\leq 3\epsilon_j\}\\
\leq& \epsilon_m + \sum_{j=1}^m \Big(\sup_{h\in\mathcal{H}_j}\|\mathbf{h}\|\Big)\sqrt{\frac{2\log\left(|\mathcal{N}_{\epsilon_j}|^2\right)}{n}} \quad \text{by Massart's lemma and } |\mathcal{H}_j|\leq|\mathcal{N}_{\epsilon_j}||\mathcal{N}_{\epsilon_{j-1}}|\leq|\mathcal{N}_{\epsilon_j}|^2\\
\leq& \epsilon_m + 12\sum_{j=1}^m(\epsilon_j-\epsilon_{j+1})\sqrt{\frac{\log|\mathcal{N}_{\epsilon_j}|}{n}} \quad \text{since } \Big(\sup_{h\in\mathcal{H}_j}\|\mathbf{h}\|\Big)\leq 3\epsilon_j\leq 6(\epsilon_j-\epsilon_{j+1})\\
=& \epsilon_m + 12\sum_{j=1}^m\int_{\epsilon_{j+1}}^{\epsilon_j}\sqrt{\frac{\log|\mathcal{N}_{\epsilon_j}|}{n}}dt\\
\leq& \epsilon_m + 12\sum_{j=1}^m\int_{\epsilon_{j+1}}^{\epsilon_j}\sqrt{\frac{\log|\mathcal{N}_t|}{n}}dt \quad \text{since } t\in[\epsilon_{j+1},\epsilon_j]\\
\leq& \epsilon_m + 12\int_{\epsilon_m}^{\epsilon_0}\sqrt{\frac{\log|\mathcal{N}_t|}{n}}dt.
\end{aligned}$$

The result follows by letting $m\to\infty$ and noticing that the above bound holds for every ϵ_j-cover. □

10 Stability and PAC-Bayes Bounds

In this lecture, we will cover two different types for generalization bounds. The first one is based on uniform stability which is based on a small modification of the proof we did for Rademacher complexity.

10.1 Stability based generalization bounds

We define the algorithmic stability as follows.

Definition 41 (Uniform stability)**.** *We say that an empirical risk minimization algorithm given as*

$$\hat{f}_{\mathcal{D}} = \operatorname*{argmin}_{f \in \mathcal{F}} \hat{R}(f; \mathcal{D}) \coloneqq \frac{1}{n} \sum_{i=1}^{n} \ell(z_i, f) \quad \text{for } \mathcal{D} = \{z_1, z_2, ..., z_n\} \in \mathcal{Z}^n,$$

is uniformly β-stable if for all training sets $\mathcal{D} \in \mathcal{Z}^n$, and their j-th sample perturbations denoted by $\mathcal{D}_j = \{z_1, .., z'_j, .., z_n\}$, we have

$$\sup_{z \in \mathcal{Z}} \left| \ell(z, \hat{f}_{\mathcal{D}}) - \ell(z, \hat{f}_{\mathcal{D}_j}) \right| \leq \beta. \tag{10.1}$$

Remark. It should be understood that smaller β corresponds to a more stable algorithm.

- We emphasize that the above notion is not for a specific empirical risk minimizer, rather for the **minimization algorithm** which is why we refer to it as algorithmic stability. The difference is that $\hat{f}$ is data specific whereas an algorithm outputs different minimizers for different data inputs. We make this dependence explicit by using the same notation $\hat{f}_{\mathcal{D}}$.
- Moreover, the above condition (10.1) is uniform over data $z \in \mathcal{Z}$, and all possible datasets $\mathcal{D}$ and their perturbations $\mathcal{D}'_j$, for all j. Needless to say, it is a very strong assumption, but can be easily verified for several algorithms of interest.

Example. [Revisiting Gaussian mean estimation]

- Consider the Gaussian mean estimation problem where we observe n data points $\mathcal{D} = \{z_1, z_2, ..., z_n\}$. Standard assumption in this problem is $z_i \sim \mathcal{N}(\mu, \sigma^2 I)$, when coupled with an ℓ_2-regularization, the MLE yields the following algorithm

$$\hat{\mu}_{\mathcal{D}} = \operatorname*{argmin}_{\mu \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \|z_i - \mu\|^2 = \frac{1}{n} \sum_{i=1}^{n} z_i \triangleq \bar{z},$$

  where we denote the sample mean estimator with $\bar{z}$.

- In this problem, we notice that the loss function is given as $\ell(z, \mu) = \|z - \mu\|^2$. For simplicity, lets assume that data points are uniformly bounded, i.e.

$$\|z_i\| \leq \kappa \ \text{ almost surely.}$$

This assumptions is clearly violated for Gaussian data; however, similar bounds can be obtained under high probability. Denoting the sample mean estimator over the perturbed data $\mathcal{D}'_j$ with $\bar{z}'_j$, we verify the uniform stability condition as follows. For $z \in \mathcal{Z}$, we write

$$\begin{aligned}\big|\ell(z,\hat{\mu}_{\mathcal{D}}) - \ell(z,\hat{\mu}_{\mathcal{D}_j})\big| =& |\|z-\hat{\mu}_{\mathcal{D}}\|^2 - \|z-\hat{\mu}_{\mathcal{D}_j}\|^2|,\\ =& |\|z-\bar{z}\|^2 - \|z-\bar{z}'_j\|^2|,\\ =& |\langle 2z-\bar{z}-\bar{z}'_j, \underbrace{\bar{z}-\bar{z}'_j}_{=(z_j-z'_j)/n}\rangle|, \quad \text{by Cauchy-Schwartz} \downarrow\\ \leq& \tfrac{1}{n}\underbrace{\|2z-\bar{z}-\bar{z}'_j\|}_{\leq 4\kappa}\underbrace{\|z_j-z'_j\|}_{\leq 2\kappa} \leq \frac{8\kappa^2}{n} \coloneqq \beta.\end{aligned}$$

- We observe that larger the sample size n, smaller the parameter β; thus, more stable the algorithm. Another observation we can make is that the radius of the support κ has a negative effect on the stability of an algorithm.

Example. [Stability of Lipschitz loss & linear functions]

- We assume that the loss is Lipschitz in its second argument, i.e.

$$\big|\ell(z,f) - \ell(z,f')\big| \leq L\|f-f'\|_\infty \triangleq L \sup_{x\in\mathbb{R}^d}\big|f(x)-f'(x)\big|.$$

  If we consider an SVM classifier where $y \in \{\pm 1\}$ and the loss is Hinge loss $\ell(z=(y,x),f) = \max\{0, 1-yf(x)\}$, we have

$$\begin{aligned}\big|\ell(z,f) - \ell(z,f')\big| =& \big|\max\{0,1-yf(x)\} - \max\{0,1-yf'(x)\}\big|\\ \leq& \big|yf(x)-yf'(x)\big| \leq \sup_{x\in\mathbb{R}^d}\big|f(x)-f'(x)\big|.\end{aligned}$$

- Now let's focus our attention to the class of linear functions $\mathcal{F} = \{x \to \langle x,\theta\rangle : \theta \in \mathbb{R}^d\}$. Any function $f \in \mathcal{F}$ can be characterized by the parameter θ; so let's switch notation $f \to \theta$.

- SVMs are generally coupled with ℓ_2-regularization; thus the resulting empirical risk minimization algorithm reduces to

$$\hat{\theta}_{\mathcal{D}} = \operatorname*{argmin}_{\theta\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n \max\{0, 1-y_i\langle\theta,x_i\rangle\} + \frac{\lambda}{2}\|\theta\|^2$$

- Therefore the resulting loss function becomes

$$\ell(\underbrace{z}_{=(y,x)}, f) = \max\{0, 1-y\underbrace{\langle\theta,x\rangle}_{=f(x)}\} + \frac{\lambda}{2n}\|\theta\|^2.$$

- If we assume $\|x_i\| \leq \kappa$, Bousquet and Elisseeff showed that this algorithm has uniform stability with parameter

$$\beta = \frac{\kappa^2}{\lambda n}.$$

  This is nontrivial, and skipped in class. Similar to the Gaussian mean estimation example, stability gets better with the number of samples. But another important observation we can make is that stability gets better with more regularization.

The following result provides a generalization bound based on uniform β-stability.

Theorem 42 (Generalization based on Uniform Stability)**.** *Assume that an empirical risk minimization algorithm is uniformly β-stable, and the loss is bounded, i.e., $0 \leq \ell(z, f) \leq B$. Then with probability at least $1 - \delta$, we have*

$$R(\hat{f}) - R(f_*) \leq \beta + (\beta n + 3B)\sqrt{\frac{2\log(1/\delta)}{n}}.$$

Remark. We make the following remarks.

- Notice that for above bound to be useful, one needs $\beta = o(1/\sqrt{n})$. This is because of the term $\beta\sqrt{n}$ in the coefficient of the second term on the right hand side.
- In general, we have $\beta = \mathcal{O}(1/n)$ which gives the familiar rate of generalization error, $\mathcal{O}(1/\sqrt{n})$.
- In the case of linear SVMs (previous example), we have $\beta = \frac{\kappa^2}{\lambda n}$. This yields a bound of order

$$\mathcal{O}\left(\frac{\kappa^2}{\lambda n} + (\kappa^2 + B)\sqrt{\frac{2\log(1/\delta)}{n}}\right) = \mathcal{O}\left(\frac{(\kappa^2 + B)\sqrt{\log(1/\delta)}}{\lambda\sqrt{n}}\right).$$

  This bound is the same order as previous generalization bounds we obtained, but it is worse in terms of dependence on κ.

Proof. The proof of this theorem is very similar to that of Theorem 23, the generalization results based on Rademacher complexity. Recall the notation $\hat{R}(f; \mathcal{D})$ which means the empirical risk of f over the dataset $\mathcal{D}$. For example, $\hat{R}(\hat{f}_{\mathcal{D}}; \mathcal{D}_j)$ is the empirical risk of $\hat{f}_{\mathcal{D}}$ over the single-data perturbed dataset $\mathcal{D}_j$.

The main observation is again to write the following decomposition of the excess risk

$$\begin{aligned} R(\hat{f}_{\mathcal{D}}) - R(f_*) &= \underbrace{[R(\hat{f}_{\mathcal{D}}) - \hat{R}(\hat{f}_{\mathcal{D}}; \mathcal{D})]}_{\text{not iid sum}} + \underbrace{[\hat{R}(\hat{f}_{\mathcal{D}}; \mathcal{D}) - \hat{R}(f_*; \mathcal{D})]}_{\leq 0} + \underbrace{[\hat{R}(f_*; \mathcal{D}) - R(f_*)]}_{\text{iid sum}/n}, \quad (10.2) \\ &\leq 2 \sup_{f \in \mathcal{F}} |\hat{R}(f; \mathcal{D}) - R(f)| \quad \text{which is what we did previously.} \end{aligned}$$

Before, we proceeded by bounding both of the above nontrivial terms with the supremum of the empirical process, $\sup_{f \in \mathcal{F}} |\hat{R}(f; \mathcal{D}) - R(f)|$. This time though, we will handle them separately. Bounding the second term above is quite easy since f_* is deterministic, and therefore it becomes an iid average, i.e.,

$$\hat{R}(f_*; \mathcal{D}) - R(f_*) = \frac{1}{n}\sum_{i=1}^{n} \ell(z_i, f_*) - \mathbb{E}[\ell(z_i, f_*)],$$

which we know how to deal with.

For the first term $R(\hat{f}_{\mathcal{D}}) - \hat{R}(\hat{f}_{\mathcal{D}}; \mathcal{D})$, we will invoke the uniform stability together with McDiarmid's inequality.

The proof relies on three key steps as before: 1-Concentration, 2-Control over expectation, and 3- Uniform conv. (10.2) $\implies$ generalization.

1. **Concentration**: Let's recall the main concentration tool that we will relied on in our efforts to derive a generalization bound based on Rademacher complexity.

Lemma 43 (Recall: McDiarmid's inequality (Lemma 24)). *Let $g : \mathcal{Z} \times ... \times \mathcal{Z} \to \mathbb{R}$ be a function satisfying the bounded difference property*

$$|g(z_1, \ldots, z_j, \ldots, z_n)| - g(z_1, \ldots, z_j', \ldots, z_n)| \leq c_j$$

Then for independent random variables $z_1, z_2 \ldots, z_n$, we have

$$\mathbb{P}\Big(g(z_1, \ldots, z_n) - \mathbb{E}[g(z_1, \ldots, z_n)] \geq \epsilon\Big) \leq \exp\left\{\frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2}\right\}.$$

Recall that Hoeffding's inequality is an application of the above lemma. We can invoke either and immediately obtain a bound on the second term. Let's get that out of the way.

Warm-up: Getting the third term in (10.2) out of way. By McDiarmid's (or by Hoeffding's) inequality, we have

$$\mathbb{P}\Big(\hat{R}(f_*; \mathcal{D}) - R(f_*) \geq \frac{\epsilon}{2}\Big) \leq \exp\Big\{-\frac{n\epsilon^2}{2B^2}\Big\} \triangleq \frac{\delta}{2}.$$

This translates to, with probability at least $1 - \delta/2$, we have

$$\hat{R}(f_*; \mathcal{D}) - R(f_*) \leq B\sqrt{\frac{2\log(2/\delta)}{n}}.$$

Bounding the first term in (10.2). Recall that previously, we needed to bound the empirical process in (10.2). For this, we'd let the g function from McDiarmid's inequality be the function of interest. That is,

$$\text{Previously: } \; g(z_1, \ldots, z_n) = \sup_{f \in \mathcal{F}} R(f) - \hat{R}(f).$$

This time though, we are dealing with another function, so we let

$$\text{This time: } \; g(z_1, \ldots, z_n) = R(\hat{f}_{\mathcal{D}}) - \hat{R}(\hat{f}_{\mathcal{D}}; \mathcal{D}).$$

Notice that, by the uniform β-stability assumption, we have

$$\Big|\hat{R}(\hat{f}_{\mathcal{D}}; \mathcal{D}) - \hat{R}(\hat{f}_{\mathcal{D}_j}; \mathcal{D})\Big| \leq \beta \;\text{ and }\; \Big|R(\hat{f}_{\mathcal{D}}) - R(\hat{f}_{\mathcal{D}_j})\Big| \leq \beta.$$

Let's verify the second one as the first one follows from the same argument.

$$\begin{aligned}
\Big|R(\hat{f}_{\mathcal{D}}) - R(\hat{f}_{\mathcal{D}_j})\Big| =& \Big|\mathbb{E}[\ell(z, \hat{f}_{\mathcal{D}}) - \ell(z, \hat{f}_{\mathcal{D}_j})]\Big| \\
\leq& \mathbb{E}[|\ell(z, \hat{f}_{\mathcal{D}}) - \ell(z, \hat{f}_{\mathcal{D}_j})|] \;\text{ by triangle ineq.} \\
\leq& \beta \;\text{ by uniform } \beta\text{-stability.}
\end{aligned}$$

We proceed by first verifying the bounded difference property which is needed by McDiarmid's inequality.

$$
\begin{aligned}
&|g(z_1,\dots,z_j,\dots,z_n) - g(z_1,\dots,z'_j,\dots,z_n)| \\
&= \left|R(\hat{f}_{\mathcal{D}}) - \hat{R}(\hat{f}_{\mathcal{D}};\mathcal{D}) - \left[R(\hat{f}_{\mathcal{D}_j}) - \hat{R}(\hat{f}_{\mathcal{D}_j};\mathcal{D}_j)\right]\right| \\
&\leq \underbrace{\left|R(\hat{f}_{\mathcal{D}}) - R(\hat{f}_{\mathcal{D}_j})\right|}_{\leq \beta \text{ by stability}} + \left|\hat{R}(\hat{f}_{\mathcal{D}};\mathcal{D}) - \hat{R}(\hat{f}_{\mathcal{D}_j};\mathcal{D}_j) \pm \hat{R}(\hat{f}_{\mathcal{D}_j},\mathcal{D})\right| \quad \text{by triangle ineq.} \\
&\leq \beta + \underbrace{\left|\hat{R}(\hat{f}_{\mathcal{D}};\mathcal{D}) - \hat{R}(\hat{f}_{\mathcal{D}_j};\mathcal{D})\right|}_{\leq \beta \text{ by stability}} + \underbrace{\left|\hat{R}(\hat{f}_{\mathcal{D}_j};\mathcal{D}) - \hat{R}(\hat{f}_{\mathcal{D}_j};\mathcal{D}_j)\right|}_{= \frac{1}{n}|\ell(z_j,\hat{f}) - \ell(z'_j,\hat{f})| \leq \frac{2B}{n}} \quad \text{by triangle ineq.} \\
&\leq 2\beta + \frac{2B}{n} \triangleq c_j \quad \text{in McDiarmid's inequality.}
\end{aligned}
$$

Hence, by the McDiarmid's inequality, we obtain

$$
\begin{aligned}
\mathbb{P}\left(R(\hat{f}_{\mathcal{D}}) - \hat{R}(\hat{f}_{\mathcal{D}};\mathcal{D}) \geq \epsilon + \overbrace{\mathbb{E}\left[R(\hat{f}_{\mathcal{D}}) - \hat{R}(\hat{f}_{\mathcal{D}};\mathcal{D})\right]}^{\text{Need to control}}\right) &\leq \exp\left\{\frac{-2\epsilon^2}{n(2\beta + 2B/n)^2}\right\} \qquad (10.3) \\
&\leq \exp\left\{\frac{-n\epsilon^2}{2(\beta n + B)^2}\right\} \triangleq \frac{\delta}{2}.
\end{aligned}
$$

The above bound is obtained under uniform stability; yet, it is not surprising at all given the McDiarmid's inequality. We still need to control the additional expectation above. This was previously done by the symmetrization argument. In the following we use stability property of the algorithm.

2. **Controlling the expectation via stability**: We denote our dataset with $\mathcal{D} = \{z_1,\dots,z_n\}$, and let $\mathcal{D}' = \{z'_1,\dots,z'_n\}$ be the iid copy of the $\mathcal{D}$, and the perturbation is given as $\mathcal{D}_j = \{z_1,..,z'_j,..,z_n\}$. We have

$$
\hat{R}(f;\mathcal{D}) = \frac{1}{n}\sum_{i=1}^{n}\ell(z_i,f) \quad \text{and} \quad \hat{R}(f;\mathcal{D}'_j) = \frac{1}{n}\sum_{i=1}^{n}\ell(z'_i,f).
$$

For a fixed f, the population risk will be identical for these datasets,

$$
R(f) = \mathbb{E}[\ell(z,f)] = \mathbb{E}[\hat{R}(f,\mathcal{D})] = \mathbb{E}[\hat{R}(f,\mathcal{D}'_j)].
$$

Let's investigate the quantity we would like to bound.

$$
\begin{aligned}
\mathbb{E}\left[R(\hat{f}_{\mathcal{D}}) - \hat{R}(\hat{f}_{\mathcal{D}};\mathcal{D})\right] &= \mathbb{E}_{\text{all}}\left[\mathbb{E}_z[\ell(z,\hat{f}_{\mathcal{D}})] - \frac{1}{n}\sum_{i=1}^{n}\ell(z_i,\hat{f}_{\mathcal{D}})\right] \\
&= \mathbb{E}\left[\mathbb{E}_{z'_i}\left[\frac{1}{n}\sum_{i=1}^{n}\ell(z'_i,\hat{f}_{\mathcal{D}})\right] - \frac{1}{n}\sum_{i=1}^{n}\ell(z'_i,\hat{f}_{\mathcal{D}_i})\right] \\
&= \mathbb{E}\left[\mathbb{E}_{z'_i}\left[\frac{1}{n}\sum_{i=1}^{n}\ell(z'_i,\hat{f}_{\mathcal{D}}) - \ell(z'_i,\hat{f}_{\mathcal{D}_i})\right]\right] \\
&\leq \beta \quad \text{by stability.}
\end{aligned}
$$

The second inequality is because z_i' is independent from $\mathcal{D}$, and $\mathcal{D}$ and $\mathcal{D}_j'$ are exchangable. Therefore we get, with probability at least $1 - \delta/2$

$$R(\hat{f}_{\mathcal{D}}) - \hat{R}(\hat{f}_{\mathcal{D}}) \leq \mathbb{E}\Big[R(\hat{f}_{\mathcal{D}}) - \hat{R}(\hat{f}_{\mathcal{D}}; \mathcal{D})\Big] + \frac{\epsilon}{2} \leq \beta + (\beta n + B)\sqrt{\frac{\log(2/\delta)}{2n}}.$$

3. **Uniform convergence $\implies$ generalization** (but almost): Combining this with (10.3), we write out generalization bound with probability at least $1 - \delta$,

$$\begin{aligned} R(\hat{f}_{\mathcal{D}}) - R(f_*) &\leq [R(\hat{f}_{\mathcal{D}}) - \hat{R}(\hat{f}_{\mathcal{D}}; \mathcal{D})] + 0 + [\hat{R}(f_*; \mathcal{D}) - R(f_*)] \\ &\leq \beta + (\beta n + B)\sqrt{\frac{\log(2/\delta)}{2n}} + B\sqrt{\frac{2\log(2/\delta)}{n}} \\ &\leq \beta + (\beta n + 3B)\sqrt{\frac{\log(2/\delta)}{2n}} \end{aligned}$$

which concludes the proof.

□

10.2 PAC-Bayes bounds

In this section, we scratch the surface of PAC-Bayesian bounds. The PAC-Bayes theory is originally developed as an attempt to explain Bayesian learning from a learning theory perspective. But these tools have to be proved very useful in various context. The main idea is to place a prior distribution π_0 over the function class $\mathcal{F}$, which encodes our prior knowledge over the set of hypotheses. After observing data $\mathcal{D}$, we update our view of the function class, which is referred to as the posterior distribution $\pi_{\mathcal{D}}$.

The bounds that rely on the concept "uniform convergence $\implies$ generalization" hold for all functions in the function class. Consider for example a finite function class. By a simple application of the union bound, we were able to derive a generalization error bound of (ignoring constants)

$$R(\hat{f}) - R(f_*) < \sqrt{\frac{\log(|\mathcal{F}|) + \log(1/\delta)}{n}},$$

which we proved as Theorem 18. However, the main building block of this theorem was to show the uniform convergence, which reads (again ignoring constants), with probability at least $1 - \delta$

$$\forall f \in \mathcal{F}, \ : \ R(f) \leq \hat{R}(f) + \sqrt{\frac{\log(|\mathcal{F}|) + \log(1/\delta)}{n}}. \tag{10.4}$$

This is equivalent to saying $\sup_{f \in \mathcal{F}} R(f) - \hat{R}(f) \leq$ the last term above. However, we notice that the above bound gives a worst case bound, in other words it gives a bound for all functions by treating them all the same. But we know some are more likely than the others!

If we had a prior distribution $\pi_0(f)$ over the class of functions $\mathcal{F}$ that are available to us, we can incorporate this to our bound. Intuitively, if there is a function $f \in \mathcal{F}$ that we are certain it is not going to be returned by our algorithm, it shouldn't count towards the size of the function class which appears in the numerator of (10.4).

Let's start with the simplest of PAC-Bayes style bounds, Occam's bound.

Theorem 44 (Occam's bound)**.** *For a countable function class $\mathcal{F}$, and a bounded loss function $0 \le \ell \le B$, if we have the prior distribution π_0 over the function class $\mathcal{F}$, then with probability at least $1-\delta$, we have*

$$\forall f \in \mathcal{F} \ : \ R(f) \le \hat{R}(f) + B\sqrt{\frac{\log(1/\pi_0(f)) + \log(1/\delta)}{2n}}. \tag{10.5}$$

Remark. We make the following immediate remarks.

1. The bound is not for the excess risk. The difference between training and the test error is small for a function f, if its prior is large.
2. If the prior distribution $\pi_0(f)$ is uniform over $\mathcal{F}$, i.e. each function is equally likely and $\pi_0(f) = \mathbb{P}(f = f_i) = 1/|\mathcal{F}|$, the above bound reduces to the bound in (10.4).
3. If the prior distribution is uniform over a subset $\mathcal{G}$ of $\mathcal{F}$, bound reduces to $\sqrt{\frac{\log(|\mathcal{G}|)+\log(1/\delta)}{2n}}$. This was exactly our intuition; the functions that are unlikely to come up shouldn't count towards the complexity of the function class.
4. If the prior puts all its mass on a single function f_0, i.e. $\pi_0(f_0) = 1$, then the bound reduces to just a concentration result, since we only have a single function that is available to us.
5. This bound allows $\mathcal{F}$ to have large size as long as the prior behaves nicely for a specific function $f \in \mathcal{F}$. For that particular function, above result will yield a good bound. However, if the prior is somewhat close to uniform distribution, then $\pi_0(f) \approx 1/|\mathcal{F}|$ will get worse with an increase in the size of the function class.

Proof. The main idea in this proof is to simply allocate the confidence parameter δ over different functions based on their prior.

For a fixed (non-random) function $f \in \mathcal{F}$, by the Hoeffding's inequality, we have

$$\mathbb{P}\Big(R(f) \ge \hat{R}(f) + \epsilon\Big) \le \exp\Big\{-\frac{2n\epsilon^2}{B^2}\Big\} := \delta_f = \pi_0(f)\delta.$$

Notice that $\sum_f \delta_f = \delta$ since π_0 is a probability distribution. The above bound reads,

$$\mathbb{P}\left(R(f) \ge \hat{R}(f) + B\sqrt{\frac{\log(1/\pi_0(f)) + \log(1/\delta_f)}{2n}}\right) \le \delta_f.$$

Note that the above bound holds for a fixed f. By applying the union bound over $f \in \mathcal{F}$, we obtain

$$\mathbb{P}\left(\forall f \in \mathcal{F} : R(f) \ge \hat{R}(f) + B\sqrt{\frac{\log(1/\pi_0(f)) + \log(1/\delta_f)}{2n}}\right) \le \sum_{f\in\mathcal{F}} \delta_f = \delta,$$

which completes the proof.

□

Let's recall our objective: We want to minimize the population risk (aka test error). The bound (10.6) upper bounds the quantity we would like to minimize. Therefore, we can minimize this

upper bound, and hope that we get close to minimizing itself! That is, the above theorem suggest to minimize the following objective

$$\hat{R}(f) + \underbrace{B\sqrt{\frac{\log(1/\pi_0(f)) + \log(1/\delta)}{2n}}}_{\text{regularizer}}. \tag{10.6}$$

The bound $B\sqrt{\frac{\log(1/\pi_0(f))+\log(1/\delta_f)}{2n}}$ will serve as a regularizer by penalizing functions that are less likely according to the prior π_0. We also observe that its impact decreases with increased sample size n.

There are two shortcomings of the Occam's bound.

- First, it relies on the union bound which requires the function class $\mathcal{F}$ to be countable.
- Second, it only allows an algorithm to return a single function rather than a posterior distribution. These are addressed in the following theorem.

Theorem 45 (McAllester's PAC-Bayes theorem)**.** *For any prior π_0 and any posterior $\pi_{\mathcal{D}}$, and a bounded loss function $0 \le \ell(z, f) \le 1$, with probability at least $1 - \delta$, we have*

$$\mathbb{E}_{f\sim\pi_{\mathcal{D}}}[R(f)] \le \mathbb{E}_{f\sim\pi_{\mathcal{D}}}[\hat{R}(f)] + \sqrt{\frac{KL(\pi_{\mathcal{D}}||\pi_0) + \log(4n/\delta)}{2n-1}}.$$

Remark.

- Compared to Occam's bound, instead of for all f, this one is for expectation under the posterior.
- If the posterior puts all its mass on one function f_0 in $\mathcal{F}$, the above bound recovers Occam's bound. Say for example, π_0 is uniform over a finite set $\mathcal{F}$. Then,

$$\begin{aligned}\mathrm{KL}(\pi_{\mathcal{D}}||\pi_0) &= \sum_f \log\left(\frac{\pi_{\mathcal{D}}(f)}{\pi_0(f)}\right)\pi_{\mathcal{D}}(f) \\ &= \log\left(\frac{\pi_{\mathcal{D}}(f_0)}{\pi_0(f_0)}\right)\pi_{\mathcal{D}}(f_0) = \log(|\mathcal{F}|).\end{aligned}$$

- Converting the above bound to a (kind of) bound on the excess risk requires characterizing the expected suboptimality,

$$\mathbb{E}_{f\sim\pi_{\mathcal{D}}}[R(f)] - R(f_*).$$

- In literature, the expectations are generally denoted with $\mathbb{E}_{f\sim\pi}[R(f)] = R(\pi)$.

Proof. Skipped in class. To be added. □

11 Kernel Methods: Basics

Up until now, we have considered the supervised learning framework where we have data points $(y, x) \in \mathcal{Y} \times \mathcal{X}$ and a loss function $\ell((y, x), f)$. Much of the focus was on generalization properties of the empirical risk minimizers, and different measures of complexity for the function class at hand. In the sequel, we focus on kernel methods which have a lot of connections to previous setup, but we will barely scratch the surface here, so it may seem like quite disconnected at first.

In classical machine learning, it is often the case to consider minimizing some loss function over a mapped feature space $\phi : \mathcal{X} \to \Phi$, with $\ell((y, \phi(x)), f)$. For example, we have been considering linear functions as a popular example $\langle \theta, x \rangle$ where x is the set of features. If we only have a single feature, instead of fitting a 1-dimensional linear regression, we can use a polynomial transformation as a feature map, e.g. $\phi(x) = [1, x, x^2]$, which allows us to fit a degree-3 polynomial by simply using a linear regression. This can be easily generalized to higher dimensions, and there are several reasons such as the ability to represent non-linear dependencies in the data.

One concern is that by increasing the dimension, how much additional computation do we need?

Example. Let's turn to our canonical example, linear regression where we have data points $(y, x) \in \mathcal{Y} \times \mathcal{X}$ and a linear hypothesis class $\mathcal{F} = \{f(\cdot) = \langle \cdot, \theta \rangle,\ \theta \in \mathbb{R}^d\}$, with squared loss $\ell((y, \phi(x)), f) = (y - \langle \phi(x), \theta \rangle)^2$. Notice that there is a feature map $\phi : \mathcal{X} \to \mathbb{R}^d$ applied to the features $x \in \mathcal{X}$.

Using the easily derived closed form solution for the least squares problem, we write

$$\begin{aligned}\hat{\theta} &= \arg\min_{\theta} \hat{R}(\theta) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \langle \theta, \phi(x_i) \rangle)^2 \\ &= (\Phi^\top \Phi)^{-1} \Phi^\top y \quad \text{where} \quad \Phi = \begin{bmatrix} —\phi(x_1)^\top— \\ \vdots \\ —\phi(x_n)^\top— \end{bmatrix}.\end{aligned}$$

Here, Φ takes the place of design matrix X. If we take the SVD of $\Phi = UDV^\top$, we can write $(\Phi^\top \Phi)^{-1} = (VDU^\top UDV^\top)^{-1} = VD^{-2}V^\top \triangleq Q^\top Q$ where $Q = D^{-1}V^\top$. Therefore,

$$\hat{\theta} = Q^\top Q \Phi^\top y.$$

For a new data point x the predicted value from the linear regression model will be

$$\begin{aligned}\hat{y} &= \langle \phi(x), Q^\top Q \Phi^\top y \rangle \\ &= \langle \underbrace{Q\phi(x)}_{\phi'(x)}, Q\Phi^\top y \rangle \quad \text{define} \quad \phi'(x) = Q\phi(x), \\ &= \sum_{i=1}^{n} \langle \phi'(x), \phi'(x_i) \rangle y_i,\end{aligned}$$

which shows that any new predicted value will be a weighted average of the response y_i's with n inner-product operations. The inner product $\langle \phi'(x), \phi'(x_i) \rangle$ should be understood as a similarity metric, i.e., if x is close to the data point x_i, the inner product will be large.

- Even though we possibly increased the dimension of the original features by applying a feature map, we observe that we only need to be able to efficiently compute the inner products $\langle \phi'(x), \phi'(x') \rangle$.

- We only need to know $k(x, x') = \langle \phi'(x), \phi'(x') \rangle$ which is termed as the "kernel" which allows us to efficiently work with high dimensional features (maps).
- There is no unique way of defining a kernel. For instance, if P is another orthogonal matrix, one can use $\psi'' = P\phi'$ which is also a valid kernel.

11.1 Basics of Hilbert Spaces

We will recall some of the basic definitions in this section.

Definition 46 (Hilbert Space)**.** *A Hilbert space $\mathcal{H}$ is a real (or complex) inner product space that is also a complete metric space with respect to the norm induced by its inner product.*

We have been using inner products throughout the lecture; thus, it is useful remind ourselves their formal definition. There are two important characteristics of an Hilbert space, its inner product and completeness. We will define inner products next, but very briefly, completeness of a space means if every Cauchy sequence of points in $\mathcal{H}$ has a limit that is also in $\mathcal{H}$. We will mostly focus on the inner product property.

Definition 47 (Inner product)**.** *An inner product is a function $\langle \cdot, \cdot \rangle : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ which has the following three properties.*

1. *Symmetry: If $f, g \in \mathcal{H}$, then $\langle f, g \rangle = \langle g, f \rangle$.*
2. *Linearity: If $f, g, h \in \mathcal{H}$ and $a, b \in \mathbb{R}$, then $\langle af + bg, h \rangle = a\langle f, h \rangle + b\langle g, h \rangle$.*
3. *Non-negativity:*
   - *For all $f \in \mathcal{H}$, we have $\langle f, f \rangle \geq 0$.*
   - *Further, $\langle f, f \rangle = 0$ if and only if $f = 0$.*

We finally note that the norm defined by this inner product is: $\|f\|_{\mathcal{H}} = \sqrt{\langle f, f \rangle}$.

Example. [Euclidean space] If we have vectors $u, v \in \mathbb{R}^d$, the standard inner product is given as $\langle u, v \rangle = \sum_i u_i v_i$ which defines the Euclidean norm as $\|u\| = \sqrt{\sum_i u_i^2}$.

Example. [Square integrable functions] Let's consider the square integrable functions on $[0, 1]$. That is,

$$L^2([0,1]) = \left\{ f : [0,1] \to \mathbb{R}, \text{ and } \int_0^1 f(x)^2 dx < \infty \right\}$$

with the inner product $\langle f, g \rangle = \int_0^1 f(x)g(x)dx$.

Definition 48 (Dual space)**.** *The dual space $\mathcal{H}^*$ of a Hilbert space $\mathcal{H}$ is the space of all continuous linear functions from the space $\mathcal{H}$ into $\mathbb{R}$. It carries a norm $\|\phi\|_* = \sup_{\|x\|_{\mathcal{H}}=1} |\phi(x)|$.*

This definition will be useful when in the main theorem, but before moving forward, we need to define what a linear function in this context means.

Definition 49 (Linear function)**.** *A function $f : \mathcal{X} \to \mathbb{R}$ is linear if for $x, x' \in \mathcal{X}$ and any $c \in \mathbb{R}$ it satisfies,*

$$f(x + y) = f(x) + f(y) \quad \textit{and} \quad f(cx) = cf(x).$$

It is important to highlight that the linear functions defined as above is different than what we normally refer to in machine learning. That is, $f(x) = ax + b$ is **not linear in** $\mathcal{X}$, but $f(x) = ax$ is linear. This can be easily verified by checking the conditions in the above definition.

Example. [Euclidean space] The dual space of Euclidean space $\mathbb{R}^d$ is given as

$$\mathcal{H}^* = \{\phi : \mathbb{R}^d \to \mathbb{R} \text{ where } \phi \text{ is linear and continuous}\}.$$

Intuitively, since $\phi(x)$ has to be linear and continuous, a linear function in $\mathbb{R}^d$ is given as

$$\phi(x) = \langle u, x \rangle,$$

for some u. Are there any other functions in $\mathcal{H}^*$?

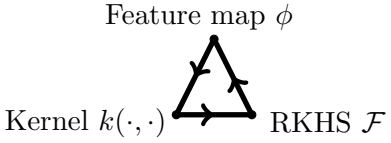
The answer to the above questions is given by the Riesz-Fréchet representation theorem, which is the main building block of what comes next.

Theorem 50 (Riesz-Fréchet representation theorem)**.** *For every element $f \in \mathcal{H}$, there is a unique element $\phi_f \in \mathcal{H}^*$ defined by $\phi_f(g) = \langle f, g \rangle$. Also, for every element $\phi \in \mathcal{H}^*$, there is a unique element $f_\phi \in \mathcal{H}$ such that $\phi(g) = \langle f_\phi, g \rangle$.*

Using this theorem, we can answer the question in the previous example. For every element in the Hilbert space $u \in \mathbb{R}^d$, there is a unique function $\phi \in \mathcal{H}^*$ defined as $\phi(x) = \langle u, x \rangle$. The converse is also true. Therefore the dual space is exactly those functions that can be written as $\phi(x) = \langle x, u \rangle$ where $u \in \mathbb{R}^d$.

11.2 Kernels: formal definitions

In this section, we will formally define kernels. Our objective is to complete the triangular relationship between the feature map ϕ, the kernel k, and the reproducing kernel Hilbert space (RKHS) to be denoted with $\mathcal{F}$ and defined later. We start with the feature map.



Definition 51 (Feature map)**.** *A feature map is a function from the input space $\mathcal{X}$ to a Hilbert space $\mathcal{H}$, i.e.,*

$$\phi : \mathcal{X} \to \mathcal{H}.$$

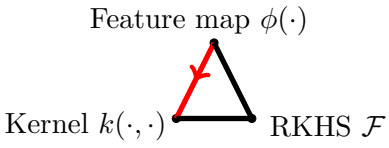
The following notation will be useful. When we define a function f from another function with two arguments $g(x, y)$, e.g. $f(x) = g(x, 3)\ \forall x$, we write this as $f(\cdot) = g(\cdot, 3)$.

Definition 52 (Kernel)**.** *A kernel is a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that for any n points $x_1, ..., x_n \in \mathcal{X}$, the matrix defined as $K_{ij} = k(x_i, x_j)$ is positive semidefinite, i.e. $K \succeq 0$.*

Example. Linear kernel $k(x, x') = \langle x, x' \rangle$ is a kernel since for any $x_1, ..., x_n$, the matrix $K_{ij} = k(x_i, x_j)$ can be written as $K = XX^\top$ where X is a matrix with rows $x_i^\top$. It is positive semidefinite (psd) since

$$K \text{ is psd if } \forall u, \quad \langle u, Ku \rangle \geq 0,$$
$$\langle u, Ku \rangle = \langle u, XX^T u \rangle = \langle X^T u, X^T u \rangle = \|X^T u\|^2 \geq 0.$$

We will see more examples of kernels later. The following result connects the feature map to kernel.



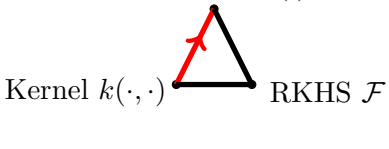
Theorem 53 (Feature map defines a kernel $[\phi(\cdot) \to k(\cdot,\cdot)]$). *A feature map $\phi : \mathcal{X} \to \mathcal{H}$ defines a kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$.*

Proof. Let $k(x, x') = \langle \phi(x), \phi(x') \rangle$, then $\forall n \in \mathbb{N}$, $x_1, \ldots, x_n \in \mathcal{X}$, the kernel matrix is given as $K_{ij} = k(x_i, x_j)$. We show that this matrix is positive semi-definite, $\forall u \in \mathbb{R}^n$,

$$\begin{aligned}
\langle u, Ku \rangle &= \sum_{ij} u_i u_j K_{ij} \\
&= \sum_{ij} u_i u_j \langle \phi(x_i), \phi(x_j) \rangle \\
&= \left\langle \sum_i u_i \phi(x_i), \sum_j u_j \phi(x_j) \right\rangle = \left\| \sum_i u_i \phi(x_i) \right\|_2^2 \geq 0.
\end{aligned}$$

□

The following result connects the kernel to the feature map when the input space $\mathcal{X}$ is finite.



Theorem 54 (Kernel defines a feature map $[k(\cdot,\cdot) \to \phi(\cdot)]$). *For every kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, there exists a Hilbert space $\mathcal{H}$ and a feature map $\phi : \mathcal{X} \to \mathcal{H}$ such that $k(x, x') = \langle \phi(x), \phi(x') \rangle$.*

We prove the above theorem after introducing some key concepts. However, it is quite straightforward to prove it when the input space $\mathcal{X}$ is finite.

Proof. [for finite $\mathcal{X}$] Let $\mathcal{X} = \{x_1, \ldots, x_n\}$, and define the kernel matrix $K_{ij} = k(x_i, x_j)$. Since K is positive semidefinite, its eigen decomposition can be written as $K = UDU^\top \triangleq \Phi\Phi^\top$, therefore $\phi(x_i)^\top = u_i^\top D^{1/2}$ defines a feature map. □

Notice that the choice of feature map is not unique. That is, $\phi'(x) = Q\phi(x)$ also defines a feature map when Q is an orthogonal matrix.

The next section introduces a key concept.

11.3 Hilbert Space defined by the Reproducing Kernel

For the dataset (y_i, x_i) for $i = 1, ..., n$, such that $y_i \in \mathbb{R}$ and the function $\mathcal{F}$ consists of functions that belongs to $f \in L^2([0, 1])$, we consider the canonical ℓ_2-regularized least squares problem

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \frac{\lambda}{2} \|f\|_{\mathcal{F}}^2$$

where $\mathcal{F}$ is the set of functions that are in consideration. If we choose $\mathcal{F}$ as the entire $f \in L^2([0,1])$, this is too complex and result in overfitting. In this case, the minimizer of the above problem would be simply the function $f(x_i) = y_i$, and $f(x) = 0$ otherwise. This function has $\|f\|_{\mathcal{F}}^2 = \int_0^1 f(x)^2 dx = 0$ and also 0 training loss. The main problem here is that the space covers indicator functions of the form $f(x) = y_i \mathbb{1}_{\{x=x_i\}}$. Clearly, Hilbert spaces are too complex of a search space, so we need some sort of restriction on the space we work with.

Definition 55 (**Lipschitz functional**). *For a Hilbert space $\mathcal{H}$, we say $L : \mathcal{H} \to \mathbb{R}$ is a Lipschitz functional if $\exists M < \infty$,*

$$|L(h) - L(h')| \leq M\|h - h'\|_{\mathcal{H}} \quad \textit{for all} \quad h, h' \in \mathcal{H}.$$

Example. If the Hilbert space is the Euclidean space $\mathbb{R}^d$ with standard inner product, define the functional $L(h) = \langle h, u \rangle$ for some $u \in \mathbb{R}^d$. Then

$$|L(h) - L(h')| = |\langle u, h \rangle - \langle u, h' \rangle| \leq \underbrace{\|u\|}_{M} \|h - h'\|_{\mathcal{H}}.$$

Definition 56 (**Evaluation functional**). *For an Hilbert space $\mathcal{H}$ consisting of functions $h : \mathcal{X} \to \mathbb{R}$, for each $x \in \mathcal{X}$, we define the evaluation functional $L_x : \mathcal{H} \to \mathbb{R}$*

$$L_x(h) = h(x).$$

A little inspection reveals that evaluation functionals are indeed linear! Notice that for $h, h' \in \mathcal{H}$ and $c \in \mathbb{R}$

$$\begin{aligned} L_x(h + h') =& (h + h')(x) = h(x) + h'(x) = L_x(h) + L_x(h'), \\ L_x(ch) =& (ch)(x) = ch(x) = cL_x(h). \end{aligned}$$

Linearity property will be crucial.

Example. Consider the Euclidean input space $\mathcal{X} = \mathbb{R}^d$, and class of linear functions $\mathcal{H} = \{h_\theta(x) = \langle x, \theta \rangle,\ \theta \in \mathbb{R}^d\}$, then $L_x(h_\theta) = \langle x, \theta \rangle$. Linearity can be verified as well.

We are ready to define RKHS.

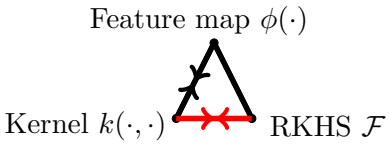
Definition 57 (Reproducing kernel Hilbert space (RKHS)). *An RKHS $\mathcal{F}$ is a Hilbert space over functions $f : \mathcal{X} \to \mathbb{R}$ such that $\forall x \in \mathcal{X}$, the evaluation functionals L_x are Lipschitz continuous.*

The constraint on the evaluation functional of a Hilbert space also restricts the function class. Notice that the reason for overfitting in the ℓ_2-regularized least squares example was the availability of indicator functions which allowed for interpolation. These indicators are not Lipschitz. For example, the problematic indicator function $f(x) = \mathbb{1}_{\{x=1\}} = L_x(f)$ violates the Lipschitz condition, hence doesn't belong to the RKHS.

One key observation was that evaluation functionals L_x are linear. Another one is that they are also continuous, which together imply that they belong to the dual space $\mathcal{F}^*$ (See Definition 48). Therefore, we can apply the second statement in the Riesz-Fréchet representation Theorem 50 and conclude that $\forall f \in \mathcal{F}$, $\exists R_x \in \mathcal{F}$ such that

$$f(x) = L_x(f) = \langle R_x, f \rangle.$$

This tells us that **function evaluations can be written as inner products**. We continue completing the triangular relationship between these key concepts. Next two results completes the following edge in the triangle.



Theorem 58 (Every RKHS defines a unique kernel $[\mathcal{F} \to k(\cdot,\cdot)]$)**.**

Proof.

- By the definition of RKHS, evaluation functionals L_x are Lipschitz (continuous) and linear, so
$$L_x \in \mathcal{F}^*.$$
- By Riesz-Fréchet representation Theorem 50, for L_x, there exists a unique $R_x \in \mathcal{F}$ such that
$$\forall f \in \mathcal{F}, \quad L_x(f) = \langle f, R_x \rangle = f(x). \tag{11.1}$$
The last equality is since L_x is an evaluation functional.
- R_x is called the *representer* and (11.1) is called the *reproducing property.*
- Since $\forall x$, the representer belongs to RKHS $R_x \in \mathcal{F}$, we can use the reproducing property on this functional as well. That is, $\forall x' \in \mathcal{X}$, and for the evaluation functional $L_{x'} \in \mathcal{F}^*$, there exists $R_{x'} \in \mathcal{F}$ such that
$$R_x(x') = L_{x'}(R_x) = \langle R_x, R_{x'} \rangle \triangleq k(x, x'),$$
where k is the kernel. This can be seen by noticing that the representer defines a feature map, i.e., $R_x = \phi(x)$; thus we can invoke Theorem 53, where we showed feature maps define a proper kernel function.

□

Remark. RKHS $\mathcal{F}$ defines a unique kernel $k(\cdot,\cdot)$ which is termed as the *reproducing kernel.* The reason for the name is
$$f(x) = L_x(f) = \langle f, R_x \rangle = \langle f, k(x,\cdot) \rangle,$$
which is where RKHSs get their name. The kernel can be transformed into being representer.

Theorem 59 (Moore-Aronszajn: Every kernel corresponds to a unique RKHS $[k(\cdot,\cdot) \to \mathcal{F}]$)**.** *For every kernel $k(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, there exists a unique RKHS $\mathcal{F}$ with the reproducing kernel k.*

Proof.

- The basic idea is to use the reproducing kernel $k(x,\cdot)$ as a basis for the RKHS $\mathcal{F}$.
- Let $\forall n \in \mathbb{N}$, $\forall x_1, \cdots, x_n \in \mathcal{X}$, $\forall \alpha_1, \cdots, \alpha_n \in \mathbb{R}$, $\forall \theta_1, \cdots, \theta_n \in \mathbb{R}$,
$$f(x) = \sum_i \alpha_i k(x, x_i), \quad \text{and} \quad g(x) = \sum_i \theta_i k(x, x_i).$$

- Let $\mathcal{F}$ be the space composed of functions of the above form. $\mathcal{F}$ is vector space, but not necessarily complete.
- Define the function $\langle \cdot, \cdot \rangle : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$ as

$$\langle f, g \rangle = \sum_{ij} \alpha_i \theta_j k(x_i, x_j).$$

We show it is an inner product. Let $f, g, h \in \mathcal{F}$ and $a \in \mathbb{R}$

  - Symmetry: holds.
  - Linearity: for $\langle af + g, h \rangle = a\langle f, h \rangle + \langle g, h \rangle$.
  - Non-negativity: It is easy to show that $\langle f, f \rangle = \sum_{ij} \alpha_i \alpha_j k(x_i, x_j) = \alpha^\top K \alpha \geq 0$ since k is a kernel. We also need to show $\langle f, f \rangle = 0$ if and only if $f = 0$. Here, $f = 0$ translates to $\alpha = 0$. It is clear that if $f = 0$, then $\langle f, f \rangle = 0$. For the other direction, we define $c(x)^\top = [k(x, x_1), \cdots, k(x, x_n)]^\top$, $\forall x \in \mathcal{X}$. The augmented kernel for a point $x \in \mathcal{X}$ is

$$K' = \begin{bmatrix} K & c(x) \\ c(x)^\top & k(x,x) \end{bmatrix}.$$

    We will prove this by contradiction. Assume that $\langle f, f \rangle = \alpha^\top K \alpha = 0$ but $f \neq 0$ (equivalently $\alpha \neq 0$). For a scalar $b \in \mathbb{R}$, let $u^\top = [\alpha, b]^\top$. Then

$$\begin{aligned} u^\top K' u =& \underbrace{\alpha^\top K \alpha}_{=0} + 2b\alpha^\top c(x) + b^2 k(x,x), \\ =& 2b\alpha^\top c(x) + b^2 k(x,x) \geq 0 \quad \text{since } K' \text{ is psd.} \end{aligned}$$

    But b can be any number, the only way to preserve the inequality for any b is when $\alpha = 0$. To see this, we investigate a function of the form $g(b) = b\xi_1 + b^2\xi_2$. We have $g(0) = 0$ and $g'(0) = \xi_1$. This means that unless $\xi_1 = 0$, the function g is either strictly increasing or decreasing at 0. Thus, one of $g(0 \pm \epsilon)$ for a small ϵ has to be negative.

- We showed that $\mathcal{F}$ is a Hilbert space. To show it is an RKHS, we need to prove that all its evaluation functionals are Lipschitz. We write $\forall f \in \mathcal{F}$

$$\begin{aligned} f(x) =& \sum_i \alpha_i k(x_i, x) \quad \text{by construction of } \mathcal{F} \\ =& \langle f, \underbrace{k(x, \cdot)}_{R_x} \rangle \quad \implies k(x, \cdot) \text{ is indeed the representer.} \end{aligned}$$

  This notation may seem confusing at first. Here, we have

$$k(x, \cdot) = \sum_{i=1}^{n} \alpha_i k(x_i, \cdot) = \underbrace{1}_{\alpha_1} \cdot k(\underbrace{x}_{x_1}, \cdot), \quad \alpha_i = 0, i > 1.$$

  For an evaluation functional L_x, for $f, g \in \mathcal{F}$, we have

$$\begin{aligned} |L_x(f-g)| = |\langle f - g, R_x \rangle| =& |\langle f - g, k(x, \cdot) \rangle| \\ \leq& \|f - g\|_{\mathcal{F}} \|k(x, \cdot)\|_{\mathcal{F}} \quad \text{by Cauchy-Schwartz} \\ =& \|f - g\|_{\mathcal{F}} \sqrt{k(x,x)} \quad \text{since } \|k(x, \cdot)\|_{\mathcal{F}}^2 = \langle k(x, \cdot), k(x, \cdot) \rangle = k(x,x). \end{aligned}$$

- To complete the proof, one needs to consider the completion of $\mathcal{F}$ by including all the limit points of $\mathcal{F}$. We skip this part.

□

Perhaps, the most important property we derived so far is that a function f in an RKHS $\mathcal{F}$ can be written as a linear combination of kernel evaluations

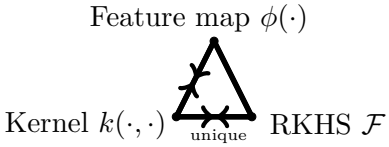
$$f(x) = \sum_i \alpha_i k(x, x_i) \quad \text{for some } x_i \in \mathcal{X}$$

where k is the unique kernel associated with the RKHS. This will help us reduce complex learning problems to least squares.

It is important to note that the above theorem also proves that if you have a kernel $k(x, x')$, you have a feature map $k(x, \cdot)$.

12 Kernel Methods: Properties & Applications

We have focused on showing that the three key concepts in kernel methods, 1- the feature map ϕ, 2- the kernel k, and 3- Reproducing Kernel Hilbert Space (RKHS) commute according to the following diagram.



Along the way, we derived a few key properties associated with the kernel and its RKHS that will be useful in the sequel. We recall them below.

- Reproducing property: Function evaluations can be written as inner products

$$f(x) = \langle R_x, f\rangle = \langle k(x,\cdot), f\rangle, \quad f \in \mathcal{F}, x \in \mathcal{X}.$$

- Moore-Aronszajn theorem: Given a kernel k, its RKHS is set of functions $f, g \in \mathcal{F}$ given as

$$f(x) = \sum_{i=1}^{n} \alpha_i k(x, x_i) \text{ for some } \alpha_i \text{ and } g(x) = \sum_{i=1}^{n} \beta_i k(x, x_i) \text{ for some } \beta_i$$

$$\text{Inner product in } \mathcal{F}: \quad \langle f, g\rangle = \left\langle \sum_{i=1}^{n} \alpha_i k(\cdot, x_i), \sum_{i=1}^{n} \beta_i k(\cdot, x_i) \right\rangle$$

$$= \sum_{ij} \alpha_i \beta_j \underbrace{\langle k(\cdot, x_i), k(\cdot, x_j)\rangle}_{=k(x_i,x_j) \text{ by reproducing prop}}$$

$$= \sum_{ij} \alpha_i \beta_j k(x_i, x_j).$$

Perhaps, the above properties of kernels are the most useful ones as far as machine learning is concerned. We first start with a few more basics related to kernels and continue with some applications in machine learning.

12.1 Basic properties and examples

We first look at a few simple examples of kernels and identify their associated RKHS.
Example. [Linear kernel] Consider the kernel $k(x, x') = \langle x, x'\rangle$ where $x, x' \in \mathcal{X} = \mathbb{R}^d$. The RKHS for this kernel can be written as

$$\begin{aligned}
\text{RKHS}(k) \ = \mathcal{F} &= \left\{ f(x) = \sum_{i=1}^{n} \alpha_i k(x, x_i), \forall n \in \mathbb{N}, \forall x_i \in \mathbb{R}^d, \forall \alpha_i \in \mathbb{R} \right\} \\
&= \left\{ f(x) = \sum_{i=1}^{n} \alpha_i \langle x, x_i\rangle, \forall n \in \mathbb{N}, \forall x_i \in \mathbb{R}^d, \forall \alpha_i \in \mathbb{R} \right\} \\
&= \left\{ f(x) = \left\langle x, \sum_{i=1}^{n} \alpha_i x_i \right\rangle, \forall n \in \mathbb{N}, \forall x_i \in \mathbb{R}^d, \forall \alpha_i \in \mathbb{R} \right\} \\
&= \{ f(x) = \langle x, \beta\rangle, \beta \in \mathbb{R}^d \} \quad \text{since } \mathbb{R}^d \text{ is a vector space.}
\end{aligned}$$

Accordingly, for $f(x) = \langle \beta, x \rangle = \sum_{i=1}^n \alpha_i \langle x_i, x \rangle$ and $f'(x) = \langle \beta', x \rangle = \sum_{i=1}^n \alpha'_i \langle x'_i, x \rangle$, the inner product is given as

$$\langle f, f' \rangle = \langle \beta, \beta' \rangle.$$

This can be seen by writing

$$\begin{aligned}
\langle f, f' \rangle &= \sum_{ij} \alpha_i \alpha'_j k(x_i, x'_j) \\
&= \sum_{ij} \alpha_i \alpha'_j \langle x_i, x'_j \rangle \\
&= \left\langle \sum_{i=1}^n \alpha_i x_i, \sum_{j=1}^n \alpha'_j x'_j \right\rangle \\
&= \langle \beta, \beta' \rangle.
\end{aligned}$$

Example. [Common kernels]

- **Identity kernel** is given as $k(x, x') = 1$. This is a kernel since for any $x_1, ..., x_n$, the kernel matrix defined as $K_{ij} = k(x_i, x_j) = 1$ is a matrix of 1's, and it is positive semidefinite.
- **Indicator function** $k(x, x') = \mathbb{1}_{\{\|x - x'\| \leq 0\}}$ is a kernel since the kernel matrix K is an identity matrix (when $x_i \neq x_j$).
- **Indicator function** $k(x, x') = \mathbb{1}_{\{\|x - x'\| \leq \epsilon\}}$ for $\epsilon > 0$ is **not** a kernel. This can be seen by choosing $x_1 = 0$, $x_2 = \epsilon e_1$ and $x_3 = 2\epsilon e_1$ where $e_1 = [1, 0, 0, ...]^\top$ is the first standard basis vector in $\mathbb{R}^d$. The kernel matrix $K = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}$ is not positive semi-definite (exercise).
- **Polynomial kernel** is given as $k(x, x') = (1 + \langle x, x' \rangle)^p$ for $p \in \mathbb{N}$. We will verify that $k(x, x')$ is a kernel shortly.
- **Gaussian kernel** is given as $k(x, x') = \exp\{-\frac{1}{2\sigma^2} \|x - x'\|^2\}$. Here, σ^2 determines the width of the kernel. Large σ^2 corresponds to smoother kernel. We will verify that $k(x, x')$ is a kernel shortly. What happens when $\sigma^2 \downarrow 0$?

In the sequel, we discuss some key properties of kernels.

1. **Inner product:** A function of the form $k(x, x') = \langle \phi(x), \phi(x') \rangle$ is a kernel (See Theorem 53).
2. **Summation:** Summation of two kernels is a kernel $k(x, x') = k_1(x, x') + k_2(x, x')$. This can be seen by considering the summation two PSD kernels K_1 and K_2 associated with the kernels k_1 and k_2, respectively, and showing that it is in fact PSD.

$$\forall u \in \mathbb{R}^d, \quad \langle u, Ku \rangle = \langle u, (K_1 + K_2)u \rangle = \langle u, K_1 u \rangle + \langle u, K_2 u \rangle \geq 0.$$

3. **Elementwise product:** (Hadamard) product of two kernels is a kernel $k(x, x') = k_1(x, x') \cdot k_2(x, x')$. Because the kernel matrices K_1 and K_2 are PSD, and we can write the following eigenvalue decomposition.

$$K_1 = UDU^T = \sum_k d_k u_k u_k^T \quad \text{and} \quad K_2 = VBV^T = \sum_k b_k v_k v_k^T$$

Here, U and V are orthogonal matrices, and D and B are diagonal matrices with nonnegative entries $d_i, b_i \geq 0$. We write

$$[K_1]_{ij} = \sum_k d_k u_{ki} u_{kj} \quad \text{and} \quad [K_2]_{ij} = \sum_k b_k v_{ki} v_{kj}$$

$$\begin{aligned}[K]_{ij} = [K_1]_{ij}[K_2]_{ij} &= \left(\sum_k d_k u_{ki} u_{kj}\right)\left(\sum_l b_l v_{li} v_{lj}\right) \\ &= \sum_{kl} d_k b_l (u_{ki} v_{li})(u_{kj} v_{lj}) \\ K &= \sum_{kl} d_k b_l (u_k \circ v_l)(u_k \circ v_l)^T \succeq 0.\end{aligned}$$

Now, we can go back and verify that polynomial and Gaussian kernels are valid kernels. We start with the polynomial kernel $k(x, x') = (1 + \langle x, x' \rangle)^p$.

1. $\langle x, x' \rangle$ is a kernel by the **inner product** property.
2. 1 is the identity kernel, so $1 + \langle x, x' \rangle$ is kernel by the **summation** property.
3. Since $1 + \langle x, x' \rangle$ is a kernel, $(1 + \langle x, x' \rangle)^p$ is kernel by the **product** property.

Using these properties, we can also verify that the Gaussian kernel is a valid kernel. The trick is to write a Taylor's series expansion.

$$\begin{aligned}k(x, x') &= \exp\left\{ -\frac{\|x - x'\|^2}{2\sigma^2} \right\} \\ &= \underbrace{\exp\left\{ -\frac{\|x\|^2}{2\sigma^2} \right\} \cdot \exp\left\{ -\frac{\|x'\|^2}{2\sigma^2} \right\}}_{k_1(x,x')} \cdot \underbrace{\exp\left\{ \frac{\langle x, x' \rangle}{\sigma^2} \right\}}_{k_2(x,x')}\end{aligned}$$

Clearly k_1 above is a valid kernel by the **inner product** property with $\phi(x) = \exp\left\{ -\frac{\|x\|^2}{2\sigma^2} \right\}$. If k_2 is a valid kernel, the **product** property will ensure that Gaussian kernel is a valid kernel. But we have

$$k_2(x, x') = \exp\left\{ \frac{\langle x, x' \rangle}{\sigma^2} \right\} = \sum_{i=0}^{\infty} \frac{1}{i!} \left(\frac{\langle x, x' \rangle}{\sigma^2} \right)^i \quad \text{by the Taylor series of } e^x.$$

Since k_2 is a sum of polynomial kernels, it is also a valid kernel.

12.2 Learning with kernels

Let's turn our attention to applications of kernels in machine learning. Suppose that we collected a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, and we would like to fit a function $y \approx f(x)$ using the function class $\mathcal{F}$ which is a RKHS. We consider the ℓ_2 regularized empirical risk minimization problem,

$$\hat{f} = \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \ell(y_i, f(x_i)) + \frac{\lambda}{2} \|f\|_{\mathcal{F}}^2. \tag{12.1}$$

Theorem 60 (Representer theorem). *For the dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, and a kernel $k(x, x')$, we define the set of functions*

$$\mathcal{V}_\mathcal{D} = \Big\{ f(x) = \sum_{i=1}^{n} \alpha_i k(x, x_i) : \alpha_i \in \mathbb{R} \text{ for } i = 1, ..., n \Big\}.$$

Then, $\hat{f}$ in (12.1) belongs to $\mathcal{V}_\mathcal{D}$, i.e. $\hat{f} \in \mathcal{V}_D \subset \mathcal{F}$.

Remark. Representer theorem has a remarkable algorithmic consequence. It tells us that minimizing over the entire RKHS $\mathcal{F}$ is equivalent to minimizing over $\mathcal{V}_\mathcal{D}$. This will reduce the empirical risk minimization problem (12.1) to a simple least squares problem over α_i's. We will revisit this after proving the theorem.

Proof.

- First, we note that in the definition of $\mathcal{V}_\mathcal{D}$, n is the number of samples and x_i's are input data, and both are fixed, whereas we recall from Moore-Aronszajn Theorem 59 that,

$$\mathcal{F} = \Big\{ f(x) = \sum_{j=1}^{m} \alpha'_j k(x, x'_j), \quad \forall m \in \mathbb{N}, \forall \alpha'_j \in \mathbb{R}, \forall x'_j \in \mathbb{R}^d \Big\}.$$

  Also, we notice that $\mathcal{V}_\mathcal{D}$ is a subspace in $\mathcal{F}$ (exercise).

- We define the orthogonal complements of the subspace $\mathcal{V}_\mathcal{D}$ as

$$\mathcal{V}_D^\perp = \{ f' \in \mathcal{F} : \ \langle f', f \rangle = 0 \ \forall f \in \mathcal{V}_D \}.$$

  A vector space is the summation of a subspace and its orthogonal complement. This enables us to write a function in RKHS $\mathcal{F}$ as the summation of a parallel and a orthogonal component (not union!). That is, $\forall f \in \mathcal{F}$, we can write

$$f(x) = f^{\parallel}(x) + f^{\perp}(x)$$

  where $f^{\parallel} \in \mathcal{V}_\mathcal{D}$ and $f^{\perp} \in \mathcal{V}_\mathcal{D}^\perp$. In other words, projection of f on $\mathcal{V}_\mathcal{D}$ is $f^{\parallel}$, and on $\mathcal{V}_\mathcal{D}^\perp$ is $f^{\perp}$.

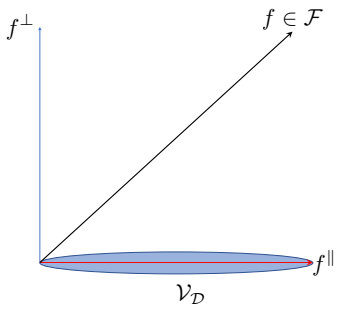


Figure 4: Decomposing RKHS into two subspaces.

- But notice that for $(x_i, y_i) \in \mathcal{D}$, by the reproducing property, we can write for $f \in \mathcal{F}$

$$f^\perp(x_i) = \langle \underbrace{f^\perp}_{\in \mathcal{V}_\mathcal{D}^\perp}, \underbrace{k(x_i, \cdot)}_{\in \mathcal{V}_\mathcal{D}} \rangle = 0$$

where the last step follows since $k(x_i, \cdot) \in \mathcal{V}_\mathcal{D}$ and $f^\perp \in \mathcal{V}_\mathcal{D}^\perp$.

Therefore for $(x_i, y_i) \in \mathcal{D}$, we have $f(x_i) = f^\parallel(x_i) + f^\perp(x_i) = f^\parallel(x_i)$. This implies that the loss over f only depends on its projection onto $\mathcal{V}_\mathcal{D}$, i.e.

$$\ell(y_i, f(x_i)) = \ell(y_i, f^\parallel(x_i)).$$

Consequently, the training error of f only depends on its projection $f^\parallel$.

- For the regularizer, we have for every $f \in \mathcal{F}$

$$\|f\|_\mathcal{F}^2 = \|f^\parallel\|_\mathcal{F}^2 + \|f^\perp\|_\mathcal{F}^2.$$

- Combining these, we obtain that the minimization problem over f can be written as

$$\begin{aligned}\hat{f} &= \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^n \ell(y_i, f(x_i)) + \frac{\lambda}{2}\|f\|_\mathcal{F}^2, \\ &= \operatorname*{argmin}_{\substack{f = f^\parallel + f^\perp: \\ f^\parallel \in \mathcal{V}_\mathcal{D},\ f^\perp \in \mathcal{V}_\mathcal{D}^\perp}} \frac{1}{n}\sum_{i=1}^n \ell(y_i, f^\parallel(x_i)) + \frac{\lambda}{2}\|f^\parallel\|_\mathcal{F}^2 + \frac{\lambda}{2}\|f^\perp\|_\mathcal{F}^2.\end{aligned}$$

Since $f^\perp$ doesn't affect the training error, we might as well choose it to be zero so that the regularizer becomes smaller. Thus, the minimizer $\hat{f}$ can be obtained by just minimizing over $f^\parallel \in \mathcal{V}_\mathcal{D}$.

□

Remark. The above proof also holds for any loss function $\ell(\{x, y, f(x)\})$, and regularizer $r(\|f\|_\mathcal{F})$ where r is monotone and strictly increasing.

Example. [Squared error loss] We choose $\ell(y, f(x)) = \frac{1}{2}(y - f(x))^2$ and the empirical risk minimization in (12.1) reduces to

$$\hat{f} = \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{2}\sum_{i=1}^n (y_i - f(x_i))^2 + \frac{\lambda}{2}\|f\|_\mathcal{F}^2.$$

Here, $\mathcal{F}$ is a RKHS. Applying the representer theorem, we obtain that the above minimizer has to satisfy

$$\hat{f}(x) = \sum_{j=1}^n \alpha_j k(x, x_j),$$

where $\alpha_i \in \mathbb{R}$. Note that the only thing that is not known to us is α_j's, which we will use our data to learn. Concatenating α_j's, we define $\alpha = [\alpha_1, ..., \alpha_n]^\top \in \mathbb{R}^n$, and we can convert the original problem to a minimization over α.

$$\hat{\alpha} = \operatorname*{argmin}_{\alpha \in \mathbb{R}^n} \frac{1}{2}\sum_{i=1}^n \Big(y_i - \sum_{j=1}^n \alpha_j k(x_i, x_j)\Big)^2 + \frac{\lambda}{2}\|f\|_\mathcal{F}^2$$

Recall the definition of kernel matrix $K_{ij} = k(x_i, x_j)$ and notice that

$$\begin{aligned}\|f\|_\mathcal{F}^2 = \langle f, f\rangle &= \Big\langle \sum_i \alpha_i k(\cdot, x_i), \sum_j \alpha_j k(\cdot, x_j)\Big\rangle \\ &= \sum_i \sum_j \alpha_i \alpha_j \langle k(\cdot, x_i), k(\cdot, x_j)\rangle \quad \text{(by prop of inner prod)} \\ &= \sum_{ij} \alpha_i \alpha_j k(x_i, x_j) = \alpha^T K \alpha \quad \text{(by def of inner prod in RKHS)}\end{aligned}$$

For the training error, we have

$$\frac{1}{2}\sum_{i=1}^{n}\Big(y_i - \sum_{j=1}^{n}\alpha_j k(x_i, x_j)\Big)^2 = \frac{1}{2}\sum_{i=1}^{n}\Big(y_i - \langle K_i, \alpha\rangle\Big)^2 \quad K_i \text{ is the } i\text{-th row of } K$$
$$= \frac{1}{2}\|y - K\alpha\|_2^2.$$

Therefore, the problem reduces to

$$\hat{\alpha} = \operatorname*{argmin}_{\alpha\in\mathbb{R}^n} \frac{1}{2}\|y - K\alpha\|_2^2 + \frac{\lambda}{2}\alpha^T K\alpha \coloneqq \hat{R}(\alpha)$$

which can be easily solved by taking derivatives and solving for α

$$\nabla_\alpha \hat{R}(\alpha) = K(y - K\alpha) + \lambda K\alpha = 0$$
$$\implies \hat{\alpha} = (K + \lambda I_n)^{-1} y.$$

Note that the solution is not unique unless $K \succ 0$. If $\hat{\alpha}$ is a solution, so is $\hat{\alpha} + u$ where u belongs to the null space of K.

12.3 Maximum mean discrepancy (MMD)

In this section, we discuss another example of kernel methods, that is RKHS embeddings of probability distributions. We start with a few definitions.

Definition 61 (∞**-norm**). *For a function $f : \mathcal{X} \to \mathbb{R}$, infinity norm is given as $\|f\|_\infty = \sup_{x\in\mathcal{X}} |f(x)|$. This defines the following metric,*

$$\|f - f'\|_\infty = \sup_{x\in\mathcal{X}} |f(x) - f'(x)|.$$

The above metric measures the worst case difference between two functions. It is often the case that we want to measure the difference between two probability distributions. For this, we have distance measures such as KL-divergence, total variation, Wasserstein distance etc. We should note that not all of these are proper metrics. The following is another way to measure distance between two probability distributions.

Definition 62 (**Maximum mean discrepancy (MMD)**). *Let p, q be probability distributions on $\mathcal{X}$. Define MMD as*

$$d_{\mathcal{F}}(p, q) = \sup_{f\in\mathcal{F}} \big|\mathbb{E}_p[f(x)] - \mathbb{E}_q[f(x)]\big|.$$

The above definition seems like something that can be useful in practice. But as in any learning algorithm, we need to choose a reference function class.

- *How complex the function class (set of test functions) $\mathcal{F}$ should be so that the above metric is good, i.e., $d_{\mathcal{F}}(p, q) = 0$ if and only if $p = q$?*

At least one side is obvious, if $p = q$ then $d_{\mathcal{F}}(p, q) = 0$. For the other side, as a starter, we have that if $\mathcal{F}$ is a 1-Lipschitz continuous functions on $\mathcal{X}$ denoted by L_1, Monge-Kantorovich duality says the above MMD metric reduces to Wasserstein-1 distance. That is,

$$d_{L_1}(p, q) = \mathcal{W}_1(p, q) \triangleq \inf_{\substack{\text{couplings } (x,y)\\ x\sim p,\ y\sim q}} \mathbb{E}[\|x - y\|].$$

Another function class that proves the above metric useful is the class of bounded continuous functions on $\mathcal{X}$, which we denote by C_0.

Theorem 63 (Dudley's result on MMD)**.** *If the function class is the set of all bounded continuous functions $\mathcal{F} = C_0$, then $d_{C_0}(p, q) = 0$ if and only if $p = q$.*

But taking supremum over L_1 or C_0 may be too much to ask since these function classes are too complex. If we can find a good representation of, say C_0, then we may be able to come up with something useful.

In lieu of Dudley's result on MMD, Theorem 63, we will require the RKHS defined by its unique kernel to be representative of the space of bounded continuous functions.

Definition 64 (**Universal Kernel**)**.** *For the set C_0 of all bounded continuous functions $f : \mathcal{X} \to \mathbb{R}$, we call a kernel k* a universal kernel *if its RKHS $\mathcal{F}$ is dense in C_0.*

Note that $\mathcal{F}$ is dense in C_0 if for every function $f \in C_0$, $\forall \epsilon > 0$, there exists $f' \in \mathcal{F}$ such that

$$\|f - f'\|_\infty \leq \epsilon.$$

The notion of universality is exactly what we need from an RKHS $\mathcal{F}$ to be a good representation of C_0. Indeed, this property translates the desired feature of C_0 to RKHS.

Theorem 65 (Steinwart's theorem on unit RKHS ball)**.** *Define the unit ball centered at origin*

$$\mathcal{G} = \{f \in \mathcal{F} : \|f\|_\mathcal{F} \leq 1\}$$

where $\mathcal{F}$ is a RKHS of a universal kernel k, then $d_\mathcal{G}(p, q) = 0 \Leftrightarrow p = q$.

Remark.

- The ball doesn't need to be unit, that is, the radius of the ball can be arbitrary as long as it is non-zero.

Proof. One side is obvious as before. For the other side, we assume that $d_\mathcal{G}(p, q) = 0$ but $p \neq q$ and hope to achieve contradiction.

- If $p \neq q$, this implies by Theorem 63, that $d_{C_0}(p, q) = \epsilon$ for some $\epsilon > 0$. Hence, there exists a function $h \in C_0$ such that

  $$|\mathbb{E}_p[h(x)] - \mathbb{E}_q[h(y)]| = \epsilon,$$

  since C_0 is compact. h may not belong to $\mathcal{F}$.

- But since k is a universal kernel, $\mathcal{F}$ is dense in C_0 which implies that there exists $f \in \mathcal{F}$ such that $\|f - h\|_\infty \leq \epsilon/3$, which in turn implies that

  $$|\mathbb{E}_p[f(x)] - \mathbb{E}_p[h(x)]| \leq \frac{\epsilon}{3} \quad \text{and} \quad |\mathbb{E}_q[f(x)] - \mathbb{E}_q[h(x)]| \leq \frac{\epsilon}{3}. \tag{12.2}$$

  To see this, we can write

  $$\begin{aligned} |\mathbb{E}_p[f(x)] - \mathbb{E}_p[h(x)]| &= \left| \int [f(x) - h(x)] dp(x) \right| \\ &\leq \int |f(x) - h(x)| dp(x) \\ &\leq \int \sup_{x \in \mathcal{X}} |f(x) - h(x)| dp(x) \\ &= \int \|f - h\|_\infty dp(x) = \|f - h\|_\infty \leq \epsilon/3. \end{aligned}$$

  The other term can be bounded by following the same steps.

- We should be careful about that $f \in \mathcal{F}$ may not belong to $\mathcal{G}$. This is okay since we can update $h \to h/\|f\|_{\mathcal{F}}$ and $\epsilon \to \epsilon/\|f\|_{\mathcal{F}}$ so that $f \to f/\|f\|_{\mathcal{F}} \in \mathcal{G}$.

- We proceed with the triangle inequality,

$$\begin{aligned}
\epsilon =& |\mathbb{E}_p[h(x)] - \mathbb{E}_q[h(y)]| \\
=& |\mathbb{E}_p[h(x)] \pm \mathbb{E}_p[f(x)] \pm \mathbb{E}_q[f(x)] - \mathbb{E}_q[h(y)]| \\
\leq& \underbrace{|\mathbb{E}_p[h(x)] - \mathbb{E}_p[f(x)]|}_{\leq \epsilon/3 \text{ by (12.2)}} + \underbrace{|\mathbb{E}_p[f(x)] - \mathbb{E}_q[f(x)]|}_{d_{\mathcal{G}}(p,q)} + \underbrace{|\mathbb{E}_q[f(x)] - \mathbb{E}_q[h(y)]|}_{\leq \epsilon/3 \text{ by (12.2)}}
\end{aligned}$$

  Notice that the first and last terms are smaller than $\epsilon/3$ and the summation of all three terms is equal to ϵ. This implies that the second term $d_{\mathcal{G}}(p, q)$ has to be larger than $\epsilon/3$ which contradicts with our initial assumption that $d_{\mathcal{G}}(p, q) = 0$.

□

Seems like the unit ball $\mathcal{G}$ is a representative function class to work with. We don't have access to the expectations, but we may be able to leverage some data and estimate the MMD metric. But how can we compute the MMD $d_{\mathcal{G}}(p, q)$ between two probability distributions p and q in practice? The answer is given by the reproducing property of the kernel k associated to its unique RKHS $\mathcal{F}$.

Recall the reproducing property of kernels that says the representer $R_x = k(x, \cdot)$ satisfies $\langle R_x, f\rangle = f(x)$. Thus, for $f \in \mathcal{F}$, we can write

$$\mathbb{E}_p[f(x)] = \mathbb{E}_p[\langle f, \underbrace{k(x, \cdot)}_{\text{random}}\rangle_{\mathcal{F}}] = \langle f, \underbrace{\mathbb{E}_p[k(x, \cdot)]}_{\triangleq \mu_p}\rangle_{\mathcal{F}} = \langle f, \mu_p\rangle_{\mathcal{F}}$$

where we defined the RKHS embedding of the probability distribution p as $\mu_p = \mathbb{E}_p[k(x, \cdot)]$. This tells us that expectations under the distribution p can be written as inner products. Therefore we can rewrite the MMD metric in its simple form

$$\begin{aligned}
d_{\mathcal{G}}(p, q) &= \sup_{f \in \mathcal{G}} |\mathbb{E}_p[f(x)] - \mathbb{E}_q[f(y)]| \\
&= \sup_{f \in \mathcal{G}} |\langle f, \mu_p - \mu_q\rangle_{\mathcal{F}}| \\
&= \|\mu_p - \mu_q\|_{\mathcal{F}},
\end{aligned}$$

where the last equality follows from $\sup_{f:\|f\|_{\mathcal{G}} \leq 1} \langle f, \mu\rangle = \|\mu\|_{\mathcal{F}}$. It turns out that MMD between p and q was just the distance between their corresponding RKHS embeddings.

Now that we converted MMD to a simple norm, we can do a lot. We write,

$$\begin{aligned}
d_{\mathcal{G}}(p, q)^2 &= \|\mu_p - \mu_q\|_{\mathcal{F}}^2 \\
&= \|\mu_p\|_{\mathcal{F}}^2 + \|\mu_q\|_{\mathcal{F}}^2 - \langle \mu_p, \mu_q\rangle_{\mathcal{F}} - \langle \mu_q, \mu_p\rangle_{\mathcal{F}}.
\end{aligned} \tag{12.3}$$

Note that

$$\begin{aligned}
\|\mu_p\|_{\mathcal{G}}^2 = \langle \mu_p, \mu_p\rangle &= \langle \mathbb{E}_p[k(x, \cdot)], \mathbb{E}_p[k(x, \cdot)]\rangle_{\mathcal{F}} \\
&= \mathbb{E}_p[\langle k(x, \cdot), k(x', \cdot)\rangle_{\mathcal{F}}] \quad \text{where} \quad x, x' \sim p \ \text{ are independent.} \\
&= \mathbb{E}_{p,p}[k(x, x')].
\end{aligned}$$

Similarly, we have

$$\begin{aligned}\langle \mu_p, \mu_q \rangle_{\mathcal{G}} =& \langle \mathbb{E}_p[k(x,\cdot)], \mathbb{E}_q[k(y,\cdot)] \rangle_{\mathcal{F}} \\ =& \mathbb{E}_{p,q}[k(x,y)] \quad \text{where } x \sim p, y \sim q \text{ are independent.}\end{aligned}$$

Plugging this back in (12.3), we obtain

$$d_{\mathcal{G}}(p,q)^2 = \mathbb{E}_{p,p}[k(x,x')] + \mathbb{E}_{q,q}[k(y,y')] - \mathbb{E}_{p,q}[k(x,y')] - \mathbb{E}_{q,p}[k(y,x')] \tag{12.4}$$

where $x, x' \sim p$, $y, y' \sim q$ and all random variables are mutually independent.

Now assume that we have iid samples from two distributions $x_1, x_2, ..., x_n \sim p$ and $y_1, y_2, ..., y_n \sim q$. We cannot calculate the MMD expression given in (12.4), but we can estimate this using the sample mean estimator. That is, we can write the following U-statistic.

$$\widehat{d_{\mathcal{G}}(p,q)^2} = \frac{1}{\binom{n}{2}} \sum_{i<j} k(x_i, x_j) + k(y_i, y_j) - k(x_i, y_j) - k(y_i, x_j).$$

This is clearly an unbiased estimator of the squared MMD $d^2_{\mathcal{G}}$. It is also consistent, i.e. it converges to $d^2_{\mathcal{G}}$ in probability,

$$\widehat{d_{\mathcal{G}}(p,q)^2} \xrightarrow{p} d_{\mathcal{G}}(p,q)^2.$$

Looking at this value can give us an idea about how close the distributions p and q are.

Acknowledgments

This course is partially designed from several courses at Stanford (Stat306A by Efron, 315A by Hastie, 300A by Siegmund, 300B by Johnstone, 300C by Candès, CS229T by Liang). Many students who took this course contributed to these lecture notes.

References