Statistical Machine Learning - Notes

Zhehao Li

April 15, 2022

Contents

This note **aggregates** the contents from the following books, articles and notes:

- Bach 2021
- Mohri, Rostamizadeh, and Talwalkar 2018
- Shalev-Shwartz and Ben-David 2014
- Tropp 2012
- Wainwright 2019
- the lecture notes of Patrick Rebeschini

1 Introduction to Supervised Learning

1.1 Decision Theory

Main Concern. What is the optimal perforamce, regardless of the finiteness of the training data? In other words, if we have a perfect knowledge of the underlying probability distribution of the data, what should be done?

We consider a fixed (testing) distribution $P_{X,Y}(x, y)$ on $\mathcal{X} \times \mathcal{Y}$, with marginal distribution $P_X(x)$ on $\mathcal{X}$. At this point we make no assumptions on the input space $\mathcal{X}$.

1.1.1 Loss Functions

We consdier a loss function $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, where $l(y, \hat{y})$ is the loss of predicting $\hat{y}$ while the true label is y. Hera are some examples:

- **Binary classification**. $\mathcal{Y} = \{0, 1\}$ or $\mathcal{Y} = \{-1, 1\}$, and we consider the 0-1 loss

$$l(y, \hat{y}) = \mathbb{1}(y \neq \hat{y})$$

- **Multi-category classification**. $\mathcal{Y} = \{1, \ldots, k\}$ and still the 0-1 loss
- **Regression**. $\mathcal{Y} = \mathbb{R}$ and consider the square loss

$$l(y, \hat{y}) = (y - \hat{y})^2$$

or the abosulte loss

$$l(y, \hat{y}) = |y - \hat{y}|$$

which is often used for "robust" estimamtion since the penaly for lager errors is smaller.

1.1.2 Risks

What should be the performance criterion for supervised learning? The answer is, given the loss function $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, we can define the expected risk or testing error of a function $f : \mathcal{X} \to \mathcal{Y}$ as the expectation of the loss function between the output y and the prediction $f(x)$.

Definition 1.1 (Expected Risk)**.** Given a loss function $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, and a distribution $P(x, y)$, the expected risk of a prediction function $f : \mathcal{X} \to \mathcal{Y}$ is defined as:

$$\mathcal{R}(f) = \mathbb{E}_{X,Y}\big[l(Y, f(X))\big] = \int_{\mathcal{X}\times\mathcal{Y}} l(y, f(x))dP(x, y) \tag{1.1}$$

Note that the risk depends on the joint distribution $P(x, y)$.

Definition 1.2 (Empirical Risk)**.** Given a loss function $l : \mathcal{Y}\times\mathcal{Y} \to \mathbb{R}$, and data $(x_i, y_i) \in \mathcal{X}\times\mathcal{Y}, i = 1, \ldots, n$, the empirical risk (training error) of a prediction function $f : \mathcal{X} \to \mathcal{Y}$ is defined as:

$$\hat{\mathcal{R}}(f) = \frac{1}{n}\sum_{i=1}^{n} l(y_i, f(x_i)) \tag{1.2}$$

where here $1/n$ is the empirical distribution function $\hat{P}(x, y)$.

- **Binary classification**. $\mathcal{Y} = \{0, 1\}$ and 0-1 loss $l(y, \hat{y}) = \mathbb{1}(y \neq \hat{y})$, we can express the risk as

  $$\mathcal{R}(f) = \mathbb{E}_{X,Y}\big[\mathbb{1}(Y \neq f(X))\big] = \mathbb{P}(Y \neq f(X))$$

  which is simply the probability of making a mistake on the testing data; while the empricial risk is the proportion of mistake on tranining data

  $$\hat{\mathcal{R}}(f) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(y_i \neq f(x_i))$$

- **Multi-category classification**. $\mathcal{Y} = \{1, \ldots, k\}$ and 0-1 loss, we can express the risk $\mathcal{R}(f)$ and empirical risk $\hat{\mathcal{R}}(f)$ in a similiar way.

- **Regression**. $\mathcal{Y} = \mathbb{R}$ and the square loss $l(y, \hat{y}) = (y - \hat{y})^2$, the risk is then

  $$\mathcal{R}(f) = \mathbb{E}_{X,Y}(Y - f(X))^2$$

  and the empirical risk

  $$\hat{\mathcal{R}}(f) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2$$

  Sometimes we use absolute loss $l(y, \hat{y}) = |y - \hat{y}|$ for robust optimization (since the penalty for large errors is smaller).

1.1.3 Bayes Risk

After the definition of performance criterion for supervised learning, what is the best prediction function f regardless of the data? The answer is, the Bayes risk and Bayes predictor.

Using the the law of iterated expecation, we have

$$\mathcal{R}(f) = \mathbb{E}\big[l(Y, f(X))\big] = \mathbb{E}\Big[\mathbb{E}\big[l(Y, f(X)) \mid X\big]\Big] = \int_{\mathcal{X}} \mathbb{E}\big[l(Y, f(X)) \mid X = x\big] dP_X(x)$$

where

$$\mathcal{R}(f; x) = \mathbb{E}[l(Y, f(X)) \mid X = x]$$

can be viewed as conditional risk, which is a deterministic function.

Proposition 1.1 (Bayes Predictor and Bayes Risk)**.** The expected risk is minimized at a Bayes predictor $f^* : \mathcal{X} \to \mathcal{Y}$ satisfying for all $x \in \mathcal{X}$,

$$f^*(x) \in \underset{f(X)\in\mathcal{Y}}{\arg\min}\, \mathbb{E}\big[l(Y, f(X)) \mid X = x\big] \tag{1.3}$$

The Bayes risk $\mathcal{R}^*$ is the risk of all Bayes predictors and we have

$$\mathcal{R}^* = \mathbb{E}_X\Big[\inf_{f(X)\in\mathcal{Y}} \mathbb{E}\big[l(Y, f(X)) \mid X\big]\Big] \tag{1.4}$$

Remark 1.1. At every point $x \in \mathcal{X}$ we have a minimizer $z^* \in \mathcal{Y}$. We unite all (x, z^*) pair, we can construct a maps from $\mathcal{X}$ to $\mathcal{Z}^*$, where $\mathcal{Z}$ is the set of all minimizer z^* at each point x. We denotes such map as $f^* : \mathcal{X} \to \mathcal{Z}^*$. The Bayes risk is then the probability weighted average ($P(x)$) of the conditional loss at each point x.

Proof. By definition, we have

$$\begin{aligned}
\mathcal{R}^* = \mathcal{R}(f^*) &= \mathbb{E}\big[l(Y, f^*(X))\big] \\
&= \mathbb{E}\Big[\mathbb{E}\big[l(Y, f^*(X)) \mid X\big]\Big] \\
&= \mathbb{E}\Big[\inf_{f(X)\in\mathcal{Y}} \mathbb{E}\big[l(Y, f(X)) \mid X\big]\Big]
\end{aligned}$$

□

Note that the Bayes predictor is not always unique, but that all lead to the same Bayes risk (e.g. in binary classification when $\mathbb{P}(Y = 1 \mid X) = 1/2$), and the Bayes risk is usually nonzero, unless the dependence between X and Y is deterministic. Given a supervised learning problem, the Bayes risk is the optimal performance; we define the excess risk as the deviation with respect to the optimal risk.

Proposition 1.2 (Excess Risk)**.** The excess risk of a function $f : \mathcal{X} \to \mathcal{Y}$ is equal to $\mathcal{R}(f) - \mathcal{R}^*$, and it is always non-negative.

Proof. We prove the propsition under the setting of binary classification with loss $l(y, \hat{y}) = \mathbb{1}(y \neq \hat{y})$. For

any fixed $X = x, x \in \mathcal{X}$, the conditional risk is

$$\begin{aligned}
\mathcal{R}(f; x) &= \mathbb{E}[\mathbb{1}(Y \neq f(X)) \mid X = x] \\
&= \mathbb{P}(Y \neq f(X) \mid X = x) \\
&= 1 - \mathbb{P}(Y = 1, f(X) = 1 \mid X = x) - \mathbb{P}(Y = -1, f(X) = -1 \mid X = x) \\
&= 1 - \mathbb{P}(Y = 1 \mid X = x)\mathbb{P}(f(X) = 1 \mid X = x) - \mathbb{P}(Y = -1 \mid X = x)\mathbb{P}(f(X) = -1 \mid X = x) \\
&= 1 - \eta(x)\mathbb{E}\big[\mathbb{1}(f(x) = 1)\big] - (1 - \eta(x))\mathbb{E}\big[\mathbb{1}(f(x) = -1)\big]
\end{aligned}$$

Hence, for $X = x, x \in \mathcal{X}$, the difference of excess conditional risk is

$$\begin{aligned}
\mathcal{R}(f; x) - \mathcal{R}(f^*; x) &= \mathbb{E}\big[\mathbb{1}(Y \neq f(X)) \mid X = x\big] - \mathbb{E}\big[\mathbb{1}(Y \neq f^*(X)) \mid X = x\big] \\
&= \mathbb{P}(Y \neq f(X) \mid X = x) - \mathbb{P}(Y \neq f^*(X) \mid X = x) \\
&= \eta(x)\Big(\mathbb{E}\big[\mathbb{1}(f^*(x) = 1)\big] - \mathbb{E}\big[\mathbb{1}(f(x) = 1)\big]\Big) + (1 - \eta(x))\Big(\mathbb{E}\big[\mathbb{1}(f^*(x) = 0)\big] - \mathbb{E}\big[\mathbb{1}(f(x) = 0)\big]\Big) \\
&= (2\eta(x) - 1)\Big(\mathbb{E}\big[\mathbb{1}(f^*(x) = 1)\big] - \mathbb{E}\big[\mathbb{1}(f(x) = 1)\big]\Big) \\
&\geq 0
\end{aligned}$$

The last inequality is due to the fact that

- if $\eta(x) \geq 1/2$, $2\eta(x) - 1 \geq 0$ and $\mathbb{E}[\mathbb{1}(f^*(x) = 1)] = 1 \geq \mathbb{E}[\mathbb{1}(f(x) = 1)]$; the product of two nonnegatives are nonnegative
- if $\eta(x) \leq 1/2$, $2\eta(x) - 1 \leq 0$ and $\mathbb{E}[\mathbb{1}(f^*(x) = 1)] = 0 \leq \mathbb{E}[\mathbb{1}(f(x) = 1)]$; the product of two negatives are nonnegative

Therefore, we have shown that $\mathcal{R}(f) \geq \mathcal{R}(f^*)$ for any function f. □

Therefore, machine learning is "trivial": given the distribution $Y \mid X$ for any $X = x$, the optimal predictor is known. The difficulty will be that this distribution is unknown.

- **Binary classification**. the Bayes predictor for $\mathcal{Y} = \{-1, 1\}$ and $l(y, \hat{y}) = \mathbb{1}(y \neq \hat{y})$ is

$$f^* = \operatorname{sgn}(2\eta(x) - 1) = \begin{cases} +1, & \text{if } \eta(x) \geq 1/2 \\ -1, & \text{otherwises} \end{cases} \tag{1.5}$$

  where $\eta(x) = \mathbb{P}(Y = 1 \mid X = x)$. Note that $\eta(x) \geq 1/2$ is equivalent to

$$\frac{\mathbb{P}(Y = +1 \mid X = x)}{\mathbb{P}(Y = -1 \mid X = x)} \geq 1$$

  The corresponding Bayes risk is

$$\mathcal{R}^* = \mathbb{P}(Y \neq f^*(X)) = \mathbb{E}\big[\min\{\eta(x), 1 - \eta(x)\}\big]$$

Proof. The conditional risk is

$$\begin{aligned}\mathcal{R}(f;x) &= \mathbb{E}\big[\mathbb{1}(Y \neq f(X)) \mid X = x\big] \\ &= \mathbb{P}(Y \neq f(X) \mid X = x) \\ &= \mathbb{P}(Y = 1, f(X) = -1 \mid X = x) + \mathbb{P}(Y = -1, f(X) = 1 \mid X = x) \\ &= \mathbb{P}(Y = 1 \mid X = x)\mathbb{P}(f(x) = -1) + \mathbb{P}(Y = -1 \mid X = x)\mathbb{P}(f(x) = 1) \\ &= \eta(x)\mathbb{P}(f(x) = -1) + (1 - \eta(x))\mathbb{P}(f(x) = 1)\end{aligned}$$

To minimize $\mathcal{R}(f;x)$, we wish the function $f : \mathcal{X} \to \{-1, 1\}$ satisfying

$$\mathcal{R}(f;x) = \begin{cases}\eta(x) & \text{if } \eta(x) \leq 1/2 \\ 1 - \eta(x) & \text{if } \eta(x) > 1/2\end{cases}$$

Therefore, we can simply let

$$f^*(x) = \underset{f(x)\in\{-1,1\}}{\arg\min}\ \mathcal{R}(f;x) = \operatorname{sgn}(2\eta(x) - 1) = \begin{cases} -1 & \text{if } \eta(x) \leq 1/2 \\ +1 & \text{if } \eta(x) > 1/2\end{cases}$$

as the Bayes predictor. The Bayes risk is therefore

$$\begin{aligned}\mathcal{R}^* &= \mathbb{E}_X\big[\mathcal{R}(f^*;X)\big] \\ &= \mathbb{E}_X\Big[\eta(X)\mathbb{P}\big(\eta(X) \leq 1/2\big) + (1 - \eta(X))\mathbb{P}\big(\eta(X) > 1/2\big)\Big] \\ &= \mathbb{E}_X\Big[\min\{\eta(X), 1 - \eta(X)\}\Big]\end{aligned}$$

□

- **Regression**. the Bayes predictor for $\mathcal{Y} = \mathbb{R}$ and $l(y, \hat{y}) = (y - \hat{y})^2$ is such that

$$f^*(x) = \mathbb{E}[Y \mid X = x] \tag{1.6}$$

The corresponding Bayes risk is

$$\mathcal{R}^* = \mathcal{R}(f^*) = \mathbb{E}\Big[\big(Y - \mathbb{E}[Y \mid X]\big)^2\Big]$$

Proof. The conditional risk of given function $f : \mathcal{X} \to \mathbb{R}$ is

$$\begin{aligned}\mathcal{R}(f;x) &= \mathbb{E}[(Y - f(X))^2 \mid X = x] \\ &= \mathbb{E}\Big[\big(Y - \mathbb{E}[Y \mid X = x] + \mathbb{E}[Y \mid X = x] - f(X)\big)^2 \mid X = x\Big] \\ &= \mathbb{E}\Big[\big(Y - \mathbb{E}[Y \mid X = x]\big)^2 \mid X = x\Big] + \mathbb{E}\Big[\big(f(X) - \mathbb{E}[Y \mid X = x]\big)^2 \mid X = x\Big] \\ &\quad + 2\mathbb{E}\Big[\big(Y - \mathbb{E}[Y \mid X = x]\big)\big(\mathbb{E}[Y \mid X = x] - f(X)\big) \mid X = x\Big]\end{aligned}$$

Notice that

$$\begin{aligned}
&\mathbb{E}\Big[\big(Y - \mathbb{E}[Y \mid X = x]\big)\big(\mathbb{E}[Y \mid X = x] - f(X)\big) \mid X = x\Big] \\
=&\big(\mathbb{E}[Y \mid X = x] - f(x)\big)\mathbb{E}\Big[Y - \mathbb{E}[Y \mid X = x] \mid X = x\Big] \\
=&\big(\mathbb{E}[Y \mid X = x] - f(x)\big)\big(\mathbb{E}[Y \mid X = x] - \mathbb{E}[Y \mid X = x]\big) \\
=&0
\end{aligned}$$

We have

$$\begin{aligned}
f^*(x) &= \operatorname*{arg\,min}_{f(X)\in\mathbb{R}} \mathbb{E}[(Y - f(X))^2 \mid X = x] \\
&= \operatorname*{arg\,min}_{f(X)\in\mathbb{R}} \mathbb{E}\Big[\big(Y - \mathbb{E}[Y \mid X = x]\big)^2 \mid X = x\Big] + \mathbb{E}\Big[\big(f(X) - \mathbb{E}[Y \mid X = x]\big)^2 \mid X = x\Big] \\
&= \mathbb{E}[Y \mid X = x]
\end{aligned}$$

and the Bayes risk is

$$\mathcal{R}^* = \mathbb{E}_X[\mathcal{R}(f^*; X)] = \mathbb{E}\big[(Y - f^*(X))^2\big] = \mathbb{E}\Big[\big(Y - \mathbb{E}[Y \mid X]\big)^2\Big]$$

□

1.2 Learning from Data

The decision theory framework outlined in the previous section gives a test performance criterion and optimal predictors, but it depends on the full knowledge of the *test distribution* $P(x, y)$. We now briefly review how we can obtain good prediction functions from *training data*, that is data sampled i.i.d. from the same distribution $P(x, y)$. There are two main classes prediction algorihms:

- Local Averaging Methods
  - Nearest-neighbors
- Empirical Risk Minimization
  - Linear least-squares regression
  - Kernel methods
  - Sparse methods
  - Neural networks

1.2.1 Local Averaging Methods

Given n observations $(x_1, y_1), \ldots, (x_n, y_n)$ where $\mathcal{X}$ is a metric space and $\mathcal{Y} \in \{0, 1\}$, a new point x^{test} is classified by a majority vote among the k-nearest neighbor of x^{test}.

- Pros:
  (a) no optimization or training
  (b) often easy to implement
  (c) can get very good performance in low dimensions (in particular for non-linear dependences between x and y)

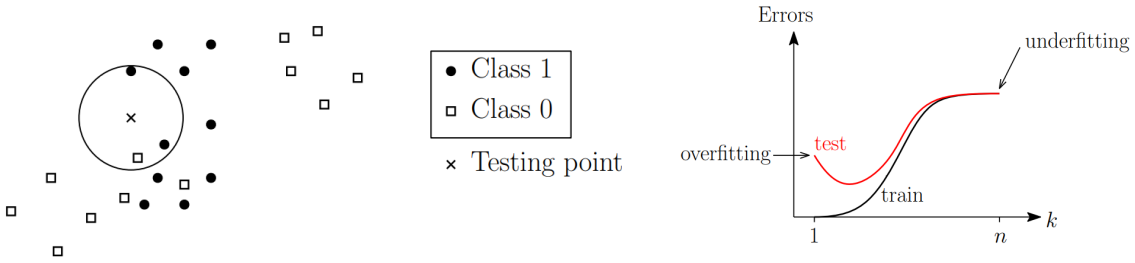


Figure 1: k-nearest-neighbor

- Cons:
  (a) slow at query time: must pass through all training data at each testing point (there are ways to reduce complexity)
  (b) bad for high-dimensional data (curse of dimensionality)
  (c) the choice of local *distance function* is crucial
  (d) the choice of "width" parameter or k has to be performed

1.2.2 Empirical Risk Minimiation

Consider a parameterized family of prediction functions $f_\theta : \mathcal{X} \to \mathcal{Y}$ for $\theta \in \Theta$ and minimize the empirical risk with respect to $\theta \in \Theta$:

$$\hat{\mathcal{R}}(f_\theta) = \frac{1}{n}\sum_{i=1}^{n} l(y_i, f_\theta(x_i))$$

this defines the estimator

$$\hat{\theta} \in \arg\min_{\theta\in\Theta} \hat{\mathcal{R}}(f_\theta)$$

and a function $f_{\hat{\theta}} : \mathcal{X} \to \mathcal{Y}$.

- Pros:
  (a) can be relatively easy to optimize (simple derivation and numerical algebra), many algorithms avaliable (mostly based on gradient descent)
  (b) can be applied in any dimension (if a reasonable feature vector is available)
- Cons:
  (a) can be relatively hard to optimize (e.g. neural networks)
  (b) need a good feature vector for linear methods
  (c) dependence on parameters can be complex (e.g. neural networks)
  (d) need some capacity control to avoid overfitting
  (e) how to parameterize functions with values in $\{0, 1\}$ (convex surrogates)
- Example: linear least-squares regression

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \theta^\top \phi(x_i))^2$$

here $f_\theta = \theta^\top \phi(x_i)$ is linear in some feature vector $\phi(x) \in \mathbb{R}^d$ (no need for $\mathcal{X}$ to be a vector space). The vector $\phi(x)$ can be quite large.

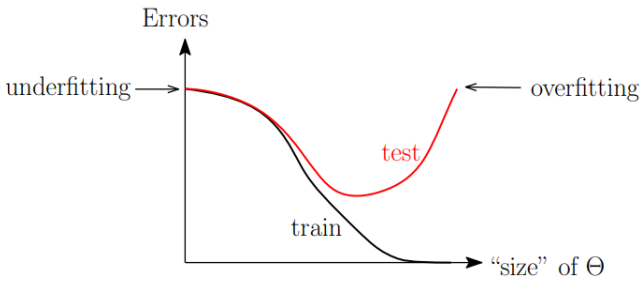


Figure 2: empirical risk

- **Risk decompmosition**

  Given any $\hat{\theta} \in \Theta$, we can weite the excess risk of $f_{\hat{\theta}}$ as:

$$\begin{aligned} \mathcal{R}(f_{\hat{\theta}}) - \mathcal{R}^* &= \left\{ \mathcal{R}(f_{\hat{\theta}}) - \inf_{\theta \in \Theta} \mathcal{R}(f_\theta) \right\} + \left\{ \inf_{\theta \in \Theta} \mathcal{R}(f_\theta) - \mathcal{R}^* \right\} \\ &= \text{estimation error} \quad + \quad \text{approximation error} \end{aligned}$$

  - The *estimation error* is typically random, becasue the function $f_{\hat{\theta}}$ is random (depend on random training data). It is typically decreasing in n (more data, less uncertainty), and usually goes up when Θ grows.
  - The *approximation error* does not depend on the chosen of $f_{\hat{\theta}}$, and is also independent of traning size n. It depends only on the class of functions parameterized by $\theta \in \Theta$, and hence it is always a determinsitic function. When Θ grows, the approximation error goes down, and to zero if arbitrary functions can be approximated arbitrary well by the function f_θ.

  Typically, for any $\hat{\theta} \in \Theta$, the *estimation error* is often decomposed as

$$\begin{aligned} \left\{ \mathcal{R}(f_{\hat{\theta}}) - \mathcal{R}(f_{\theta^*}) \right\} &= \left\{ \mathcal{R}(f_{\hat{\theta}}) - \hat{\mathcal{R}}(f_{\hat{\theta}}) \right\} + \left\{ \hat{\mathcal{R}}(f_{\hat{\theta}}) - \hat{\mathcal{R}}(f_{\theta^*}) \right\} + \left\{ \hat{\mathcal{R}}(f_{\theta^*}) - \mathcal{R}(f_{\theta^*}) \right\} \\ &\leq 2 \sup_{\theta \in \Theta} \left| \hat{\mathcal{R}}(f_\theta) - \mathcal{R}(f_\theta) \right| \quad + \quad \text{empirical optimizati error} \end{aligned}$$

  where θ^* is a minimizer on Θ. The uniform deviation grows with the "size" of Θ, and usually decays with n.

- **Capacity control**

  In order to avoid overfitting, we need to make sure that the set of allowed functions is not too large, by typically reducing the number of parameters, or by restricting the norm of predictors (thus by reducing the "size" of Θ): this typically leads to constrained optimization, and allows for risk decompositions as done above.

This can be done by regularization, that is, minimizing the follow:

$$\hat{\mathcal{R}}(f_\theta) + \lambda\Omega(\theta) = \frac{1}{n}\sum_{i=1}^{n} l(y_i, f_\theta(x_i)) + \lambda\Omega(\theta)$$

where $\Omega(\theta)$ controls the complexity of f_θ. The main example is ridge regression:

$$\min_{\theta\in\mathbb{R}^d} \quad \frac{1}{n}\sum_{i=1}^{n}\big(y_i - \theta^\top\phi(x_i)\big)^2 + \lambda||\theta||_2^2$$

this is often easier to optimization, but harder to analyze.

Note: There is a difference between parameters (e.g., θ) learned on the training data and hyperparameters (e.g., λ) learned on the validation data

1.3 Statistical Learning Theory

The goal of learning theory is to provide some guarantess of performance on unseen data. A common assumption is that the data $\mathcal{D}_P(n) = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ is obtained as i.i.d. observations from some unknown distribution P from a family $\mathcal{P}$.

An algorithm $\mathcal{A}$ is a mapping from $\mathcal{D}_P(n)$ (for any size n) to a function from $\mathcal{X}$ to $\mathcal{Y}$. The risk depends on the probability $P \in \mathcal{P}$, as $\mathcal{R}_P(f)$. The goal is to find $\mathcal{A}$ such that the risk

$$\mathcal{R}_P\big(\mathcal{A}(\mathcal{D}_P(n))\big) - \mathcal{R}_P^*$$

is small enough, where $\mathcal{R}_P^*$ is the Bayes risk. Here we assume that $\mathcal{D}_P(n)$ is sampled from P, where $P \in \mathcal{P}$ is unknown. Moreover, the risk is random because $\mathcal{D}(n)$ is random.

1.3.1 Measures of Performance

There are several ways of dealing with the randomness of the risk in order to obtain a criterion.

- **Expected Error**: we measure the performance of algorithm as

  $$\mathbb{E}\Big[\mathcal{R}_P\big(\mathcal{A}(\mathcal{D}_n(P))\big)\Big]$$

  where the expectation is with respect to the training data.

  - Consistency: an algorithm $\mathcal{A}$ is called *consistent in expectation* for the distribution P, if

    $$\mathbb{E}\Big[\mathcal{R}_P\big(\mathcal{A}(\mathcal{D}_n(P))\big)\Big] - \mathcal{R}_P^*$$

    goes to zero when n tends to infinity.

- **Probably Approximately Correct (PAC) Learning**: for a given $\delta \in (0, 1)$ and $\varepsilon > 0$

  $$\mathbb{P}\Big(\mathcal{R}_P\big(\mathcal{A}(\mathcal{D}_n(P))\big) - \mathcal{R}_P^* \leq \varepsilon\Big) \geq 1 - \delta$$

  The crux is to find ε which is as small as possible (typically as a function of δ).

– Consistency: an algorithm $\mathcal{A}$ is called consistent in PAC for the distribution P, if for any $\varepsilon > 0$, there exists $\delta_n \in (0,1)$, such that

$$\mathbb{P}\Big(\mathcal{R}_P\big(\mathcal{A}(\mathcal{D}_n(P))\big) - \mathcal{R}_P^* \leq \varepsilon\Big) \geq 1 - \delta_n$$

and the sequence δ_n goes to zero as $n \to \infty$.

1.3.2 Some Notions in Learning Problems

Definition 1.3 (Uniform Consistency)**.** An algorithm is called universally consistent (in expectation) if for all distributions P on (X, Y), the algorithm $\mathcal{A}$ is consistent in expectation with respect to distribution P.

Most often, we want to study uniform consistency within a class $\mathcal{P}$ of distributions satisfying some regularities. We thus aim at finding an algorithm $\mathcal{A}$ such that

$$\sup_{P \in \mathcal{P}} \mathbb{E}\Big[\mathcal{R}_P\big(\mathcal{A}(\mathcal{D}_P(n))\big)\Big] - \mathcal{R}_P^*$$

is as small as possible.

Definition 1.4 (Minimax Risk)**.** The minimax risk is defined as

$$\inf_{\mathcal{A}} \sup_{P \in \mathcal{P}} \mathbb{E}\Big[\mathcal{R}_P\big(\mathcal{A}(\mathcal{D}_P(n))\big)\Big] - \mathcal{R}_P^* \tag{1.7}$$

The minimax risk is typlically a function of the sample size n, the properties of $\mathcal{X}, \mathcal{Y}$ and the distribution space $\mathcal{P}$. In order to compute the estimates of minimax risk, several techniques exist:

- **Upper-bounding**

  One given algorithm with a convergence proof provides an upper-bound on the optimal performance

- **Lower-bounding** In some setups, it is possible to show that the infimum over all algorithms is greater than a certian quantity.

The ML researcher are happy when upper-bounds and lower-bounds match (up to constant factors).

- **Non-asymptotic Analysis**

  The analysis can be "non-asymptotic", with an upper-bound with explicit dependence on all quantities; the bound is then valid for all n, even if sometimes vacuous.

- **Asymptotic Analysis**

  The analysis can also be "asymptotic", where for examples n goes to infinity and limits are taken (alternatively, several quantities can be made to grow simultaneously).

What (arguably) matters most here is the dependence of these **rates** on the problem, not the choice of "in expectation" vs. in "high probability", or "asymptotic" vs. "non-asymptotic", as long as the problem parameters explicitly appear.

1.3.3 No Free Lunch Theorems

Although it may be tempting to define the optimal learning algorithm that works optimally for all distributions, this is impossible. In other words, learning is not possible without assumptions.

The following theorems shows that for any algorithm, for a fixed n, there is a data distribution P that makes the algorithm useless

Theorem 1.1 (No Free Lunch - Fixed n)**.** Consider the binary classification with 0-1 loss, with $\mathcal{X}$ infinite. Let $\mathcal{P}$ denote the set of all probability distributions on $\mathcal{X} \times \{0, 1\}$. For any $n > 0$and learning algorithm $\mathcal{A}$,

$$\sup_{P \in \mathcal{P}} \mathbb{E}\Big[\mathcal{R}_P\big(\mathcal{A}(\mathcal{D}_P(n))\big)\Big] - \mathcal{R}_P^* \geq 1/2$$

The proof see Bach 2021. The folliwng theroem is much stronger, as it more convincingly shows that learning can be arbitrarily slow without assumption.

Theorem 1.2 (No Free Lunch - Sequence of Errors)**.** Consider the binary classification with 0-1 loss, with $\mathcal{X}$ infinite. Let $\mathcal{P}$ denote the set of all probability distributions on $\mathcal{X} \times \{0, 1\}$. For any decreasing sequence a_n tending to zero and such that $a_1 \leq 1/16$, for any learning algorithm $\mathcal{A}$, there exists $P \in \mathcal{P}$, such that for all $n \geq 1$,

$$\mathbb{E}\Big[\mathcal{R}_P\big(\mathcal{A}(\mathcal{D}_P(n))\big)\Big] - \mathcal{R}_P^* \geq a_n$$

2 Empirical Risk Minimization

Main Concern. Given a joint distribution $P(X, Y)$, and n independent and identically distributed observations from $P(X, Y)$, our goal is to learn a function $f : \mathcal{X} \to \mathcal{Y}$ with minimum risk:

$$\mathcal{R}(f) = \mathbb{E}[l(y, f(x))]$$

or equivalently minimum excss risk:

$$\mathcal{R}(f) - \mathcal{R}^* = \mathcal{R}(f) - \inf_g \mathcal{R}(g)$$

where g is a measurable function. In this section, we consdier the methods based on empirical risk minimization.

2.1 Convexification of the Risk

Before looking at the necessary probabilistic tools, we will first show how problems where the output space is not a vector space, such as binary classification with $y = \{-1, 1\}$, can be reformulated with so-called convex surrogates of loss functions.

Remark 2.1 (Motivation of Risk Convexificatoin)**.** As our goal is to estimate a binary-valued function, the first idea that comes into mind is to minimize the empirical risk over a hypothesis space of binary-valued functions. However, this approach leads to a combinatorial problem which can be computationally intractable and moreover, it is not clear how to control the capacity for these type of hypothesis spaces. Learning a real-valued function instead through the problem the framework of convex surrogates simplifies and overcomes this problem as it convexifies the problem and classical penalty-based regularization techniques can be used for theoretical analysis and for algorithms.

Instead of learning $f : \mathcal{X} \to \{-1, 1\}$, we will thus learn a function $g : \mathcal{X} \to \mathbb{R}$ and define $f(x) = \text{sgn}(g(x))$ where

$$\text{sgn}(a) = \begin{cases} +1 & \text{if} \quad a \geq 0 \\ -1 & \text{if} \quad a < 0 \end{cases}$$

The risk of the function $f = \text{sgn} \circ g$, still denoted $\mathcal{R}(g)$, is then equal to

$$\mathcal{R}(g) = \mathbb{P}\big(\text{sgn}(g(x) \neq y)\big) = \mathbb{E}\big[\mathbb{1}(g(x) \neq y)\big] = \mathbb{E}\big[\mathbb{1}(yg(x) < 0)\big] = \mathbb{E}\big[\phi_{0-1}(yg(x))\big]$$

where $\phi_{0-1} : \mathbb{R} \to \mathbb{R}$, with $\phi_{0-1}(u) = \mathbb{1}(u < 0)$ is called the "margin-based" $0-1$ loss function. For empirical risk minimization, we then miniize the empirical risk

$$\hat{\mathcal{R}}(g) = \frac{1}{n} \sum_{i=1}^{n} \phi_{0-1}(y_i g(x_i))$$

with respect to $g : \mathcal{X} \to \mathbb{R}$. However, the function ϕ_{0-1} is not continuous (and thus also non-convex) and leads to difficult optimization problems.

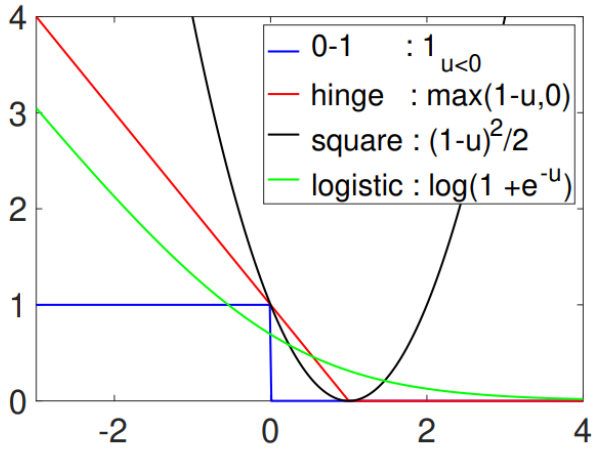
2.1.1 Convex Surrogates

A key concept in machine learning is the use of convex surrogates, where we replace ϕ_{0-1} by another function ϕ with better numerical properties (all will be convex). Instead of minimizing the classical risk $\mathcal{R}(g)$ or its empirical version $\hat{\mathcal{R}}(g)$, one then minimizes the ϕ-risk (and its empirical version) defiend as

$$\mathcal{R}_\phi(g) = \mathbb{E}\big[\phi(yg(x))\big]$$

and

$$\hat{\mathcal{R}}_\phi(g) = \frac{1}{n} \sum_{i=1}^{n} \phi(yg(x))$$



- **Quadratic loss**. $\phi(u) = (u-1)^2$ leading to

$$\begin{aligned}\phi(yg(x)) &= (yg(x)-1)^2 = (yg(x)-y^2)^2 = y^2(g(x)-y)^2 \\ &= (y-g(x))^2 = (g(x)-y)^2\end{aligned}$$

  with the notice that $y^2 = 1$. Hence, we get leas-squares.

- **Logistic loss**. $\phi(u) = \log(1+e^{-u})$, leading to

$$\phi(yg(x)) = \log(1+e^{-yg(x)}) = -\log\left(\frac{1}{1+e^{-yg(x)}}\right) = -\log\sigma(yg(x))$$

  where $\sigma(\nu) = 1/(1+e^{-\nu})$ is the sigmoid function. Note the link with maximum likelihood estimation, where we define the model through

$$\mathbb{P}(y=1 \mid x) = \sigma(f(x)) \quad \text{and} \quad \mathbb{P}(y=-1 \mid x) = \sigma(-f(x)) = 1-\sigma(f(x))$$

- **Hinge loss**. $\phi(u) = \max(1-u, 0)$, with linaer predictors, this leads to the support vector machine, and $yf(x)$ is often called the "margin" in this context. This loss has a geometric interpretation.

- **Squared Hinge loss**. $\phi(u) = \max(1-u,0)^2$, this is a smooth counterpart to the regular hinge loss.

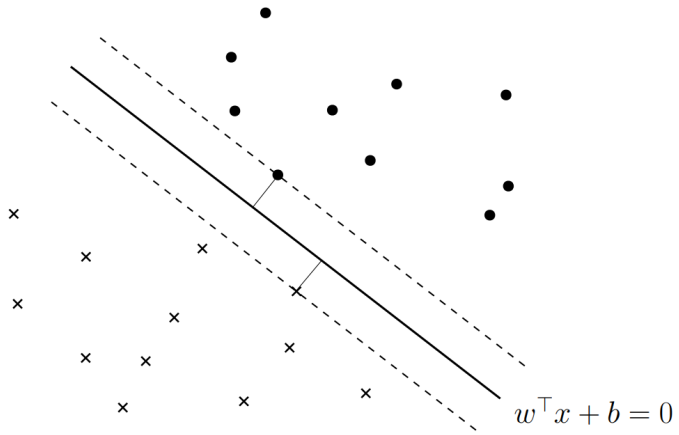
2.1.2 Geometric Interpretation of the Support Vector Machine

We consider n observations $(x_i, y_i) \in \mathbb{R}^d \times \{-1, 1\}$ for $i = 1, \dots, n$.

Separable data. We first assume that the data are separable by an affine hyperpalne, that is, there exist $\omega \in \mathbb{R}^d$ and $b \in \mathbb{R}$ such that for all $i \in \{1, \dots, n\}$,

$$y_i(\omega^\top x_i + b) > 0$$

Among the infinitely many separating hyperplane, we aim at selecting the one for which closet point from the dataset is farthest.



The distance from x_i to the hyperplane $\{x \in \mathbb{R}^d, \omega^\top x + b = 0\}$ is equal to $\frac{|\omega^\top x_i + b|}{||\omega||_2}$, and thus, this minimal distance is

$$\underset{x_i}{\text{minimize}} \quad \frac{y_i(\omega^\top x_i + b)}{||\omega||_2}$$

and we thus aim at maximizing this quantity over ω and b. Because of the invariance by rescaling ω and b, that is, we can rescale (ω, b) pair such that

$$\underset{x_i}{\text{minimize}} \quad \frac{y_i(\omega^\top x_i + b)}{||\omega||_2} = \frac{1}{||\omega||_2}$$

Then this problem is equivalent to

$$\begin{aligned} \underset{\omega \in \mathbb{R}^d,\ b \in \mathbb{R}}{\text{minimize}} \quad & \frac{1}{2}||\omega||_2^2 \\ \text{subject to} \quad & y_i(\omega^\top x_i + b) \geq 1, \quad \forall\ i \in \{1, \dots, n\} \end{aligned} \tag{2.1}$$

General data. When data may not separated by an hyperplane, then we can introduce so-called "slack variables" $\xi_i \geq 0, i = 1, \dots, n$, allowing the constraint $y_i(\omega^\top x_i + b) \geq 1$ to be not satisfied, by introducing instead the constraint

$$y_i(\omega^\top x_i + b) \geq 1 - \xi_i$$

The overall amount of slack is then minimized, leading tot he following problem (with $C > 0$)

$$\begin{aligned} \underset{\omega \in \mathbb{R}^d, b \in \mathbb{R}, \xi \in \mathbb{R}^n}{\text{minimize}} \quad & \frac{1}{2}||\omega||_2^2 + C\sum_{i=1}^n \xi_i \\ \text{subject to} \quad & y_i(\omega^\top x_i + b) \geq 1 - \xi_i \\ & \xi_i \geq 0 \quad \forall\ i \in \{1, \dots, n\} \end{aligned} \tag{2.2}$$

With $\lambda = \frac{1}{nC}$, the problem above is equivalent to

$$\underset{\omega \in \mathbb{R}^d, b \in \mathbb{R}}{\text{minimize}} \quad \frac{1}{n}\sum_{i=1}^n \big(1 - y_i(\omega^\top x_i + b)\big)^+ + \frac{\lambda}{2}||\omega||_2^2$$

which is exactly an l_2-regularized empirical risk minimization with the hinge loss, for the prediction function $f(x) = \omega^\top x + b$.

Lagrange dual. The problem in Eq.2.2 is a linearly constrained convex optimmization problem, and can be analyzed using Lagrangian duality. We consdier non-negative Lagrange multipliers α_i and β_i, $i \in \{1, \dots, n\}$ and the following Lagrangian

$$\mathcal{L}(\omega, b, \xi, \alpha, \beta) = \frac{1}{2}||\omega||_2^2 + C\sum_{i=1}^n \xi_i - \sum_{i=1}^n \alpha_i\big(y_i(\omega^\top x_i + b) - 1 + \xi_i\big) - \sum_{i=1}^n \beta_i \xi_i$$

Minimizing with respect to $\xi \in \mathbb{R}^n$ leads to the constraints

$$\alpha_i + \beta_i = C \quad \forall\ i \in \{1, \dots, n\}$$

while minimizing with respect to b leads to the constraints

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

and finally minimizing with respect to ω tells us

$$\omega = \sum_{i=1}^{n} \alpha_i y_i x_i$$

Substitute these constraints back to Lagrangian we get the dual function and dual optimizaton problem

$$\begin{aligned} \underset{\alpha \in \mathbb{R}^n}{\text{maximize}} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^\top x_j \\ \text{subject to} \quad & \sum_{i=1}^{n} \alpha_i y_i = 0 \\ & \alpha_i \in [0, C] \quad \forall\, i \in \{1, \dots, n\} \end{aligned} \tag{2.3}$$

As we will show in the future that all l_2-regularized learning problems with linear predictors, the optimization problem only depends on the dot-products

$$x_i^\top x_j \quad \forall\, i, j = 1, \dots, n$$

and the optimal predictor can be wriiten as a linear combination of input data points $x_i, i = 1, \dots, n$.

Support vector. For optimal primal and dual variables, the "complementary slackness" conditions for linear inequality constraints lead to

$$\alpha_i \big(y_i(\omega^\top x_i + b) - 1 + \xi_i \big) = 0$$

and

$$\beta_i \xi_i = (C - \alpha_i)\xi_i = 0$$

This implies that $\alpha_i = 0$ as soon as $y_i(\omega^\top x_i + b) < 1$, and thus many of the α_i are equal to zero, and the optimal predictor is a linear combination of only a few of the data point x_i's which are then called "support vectors".

2.1.3 Conditional Surrogate Risk and Classification Calibration

Most of the convex surrogates Φ are upper-bounds on the $0-1$ loss, and all can be made so with rescaling. Using this as the sole justification of the good performance of a convex surrogate is misleading justification, with the exception of problems with almost surely zero loss for the Bayes predictor (which is only possible when the Bayes risk is zero).

If we denote $\eta(X) = \mathbb{P}(Y = 1 \mid X) \in [0, 1]$, then we have, $\mathbb{E}[Y \mid X] = 2\eta(x) - 1$ and as in Section 1.1,

$$\mathcal{R}(g) = \mathbb{E}\Big[\Phi_{0-1}(Y g(X))\Big] = \mathbb{E}\Big[\mathbb{E}\big[\mathbb{1}(Y \neq g(X)) \mid X\big]\Big] \geq \mathbb{E}\Big[\min\big\{\eta(X), 1 - \eta(X)\big\}\Big] = \mathcal{R}^*$$

and one of the best classifier is

$$f^*(X) = \text{sgn}(2\eta(X) - 1)$$

Note that there are many potential other functions $g(x)$ than $2\eta(X) - 1$ so that $f^*(X) = \text{sgn}(g(X))$ is optimal. The first (minor) reason is the arbitrary choice of prediction for the tie $\eta(X) = 1/2$. The other reason is that $g(X)$ simply has to have the same sign as $2\eta(X) - 1$, which leads to many possibilities beyond $2\eta(X) - 1$.

In order to study the impact of using the Φ-risk, we first look at the conditional risk for a given X (as for the 0-1 loss, the function that g that will minimize the Φ-risk can be determined by looking at each x separately).

Definition 2.1 (Conditional Φ-risk)**.** Let $g : \mathcal{X} \to \mathbb{R}$, we define the conditional Φ-risk as

$$\mathbb{E}\big[\Phi(Yg(X)) \mid X\big] = \eta(X)\Phi(g(X)) + \big(1 - \eta(X)\big)\Phi(-g(X)) := C_{\eta(X)}\big(g(X)\big) \tag{2.4}$$

with

$$C_\eta(\alpha) = \eta\Phi(\alpha) + (1 - \eta)\Phi(-\alpha)$$

The least we can expect from a convex surrogate is that in the population case, where all X's decouple, the optimal $g(X)$ obtained by minimizing the conditional Φ-risk exactly leads to the same prediction as the Bayes predictor (at least when this prediction is unique). In other words, since the prediction is $\text{sgn}(g(X))$, we want that for any $\eta \in [0, 1]$:

$$\begin{aligned} &\text{(positive optimal prediction)}, \quad \eta > 1/2 \quad \Leftrightarrow \quad \arg\min_{\alpha\in\mathbb{R}} \; C_\eta(\alpha) \subset \mathbb{R}_+ \\ &\text{(negative optimal prediction)}, \quad \eta < 1/2 \quad \Leftrightarrow \quad \arg\min_{\alpha\in\mathbb{R}} \; C_\eta(\alpha) \subset \mathbb{R}_- \end{aligned} \tag{2.5}$$

A function Φ that satisfies these two statement is said *classification-calibrated*, or simply *calibrated*. It turns out that when Φ is convex, a simple sufficient and necessary condition is avaiable:

Proposition 2.1 (Bartlett, Jordan, and McAuliffe 2006)**.** Let $\Phi : \mathbb{R} \to \mathbb{R}$ convex. Φ calibrated $\Leftrightarrow$ Φ is differentiable at 0 and $\Phi'(0) < 0$.

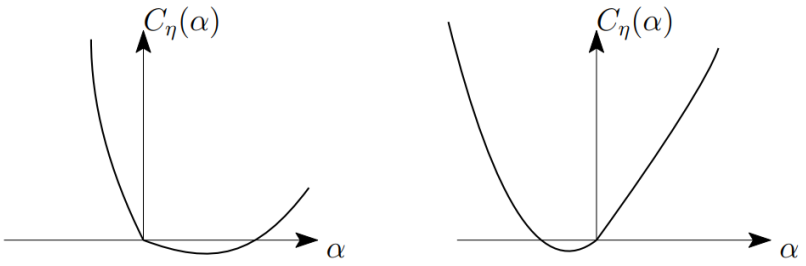


Figure 3: Classification calibration

Proof. Since Φ is convex, so is C_η for any $\eta \in [0, 1]$, and thus we simply consider left and right derivatives at zero too obatin conditions acbout location of minimizers, with the two possibilities below (minimizer in

$\mathbb{R}_+$ if and only if the right derivative at zero strictly negative, and minimizre in $\mathbb{R}_-$ if and only if the left derivative at zero is strictly positive):

$$\begin{aligned} \operatorname*{arg\,min}_{\alpha\in\mathbb{R}} \; C_\eta(\alpha) \subset \mathbb{R}_+ \quad &\Leftrightarrow \quad (C_\eta)_+(0)' = \eta\Phi'_+(0) - (1-\eta)\Phi'_-(0) < 0 \\ \operatorname*{arg\,min}_{\alpha\in\mathbb{R}} \; C_\eta(\alpha) \subset \mathbb{R}_- \quad &\Leftrightarrow \quad (C_\eta)_-(0)' = \eta\Phi'_-(0) - (1-\eta)\Phi'_+(0) < 0 \end{aligned} \tag{2.6}$$

(a) Assume Φ is calibrated.

By letting η tend to $(1/2)+$ in Eq.(2.6), we have

$$(C_{1/2})_+(0)' = \frac{1}{2}[\Phi'_+(0) - \Phi'_-(0)] \leq 0$$

Since Φ is convex, we always have $\Phi'_+(0) - \Phi'_-(0) \geq 0$. Thus the left and right derivatives are equal, which implies that Φ is differentiable at 0. Then $C'_\eta(0) = (2\eta-1)\Phi'(0)$ and from the first rows of Eq.(2.5) and Eq.(2.6), we need to have $\Phi'(0) < 0$.

(b) Assume Φ is differentiable at 0 and $\Phi'(0) < 0$, then $C'_\eta(0) = (2\eta-1)\Phi'(0)$; Eq.(2.5) are then direct consequnces of Eq.(2.6) by noticing the Fig. 3.

□

Note that the proposition above excludes the convex surrogate $u \mapsto (-u)^+ = \max\{-u, 0\}$, which is not differentiable at zero. From now on, we assume that Φ is calibrated and convex, that is, Φ convex, Φ differentiable in 0, and $\Phi(0) < 0$. We should also notice that if $\Phi(\alpha)$ is symmetric with respect to origin point 0, we have

$$C_{\eta(X)} = (2\eta - 1)\Phi(g(X)) \tag{2.7}$$

2.1.4 Relationship between Risk and Surrogate Risk

Now that we know that for any $x \in \mathcal{X}$, minimizing $C_{\eta(X)}(g(X))$ with respect to $g(X)$ leads to the optimal prediction through $\operatorname{sgn}(g(X))$, we would like to make sure that an explicit control of the excess Φ-risk leads to an explicit control of the original excess risk. In otherwords, we are looking for a monotonic function $H : \mathbb{R}_+ \to \mathbb{R}_+$ such that

$$\mathcal{R}(g) - \mathcal{R}^* \leq H[\mathcal{R}_\Phi(g) - \mathcal{R}^*_\Phi]$$

where $\mathcal{R}^*_\Phi$ is the minimum possible Φ-risk. The function H is often called the **calibration function**.

We first start with a simple lemma expressing the excess risk, as well as an upper bound, that we need for comparison inequalities below.

Lemma 2.1. For any function $g : \mathcal{X} \to \mathbb{R}$, and for a Bayes predictor g^*:

$$\mathcal{R}(g) - \mathcal{R}^* = \mathbb{E}[|2\eta(X) - 1| \cdot \mathbb{1}_{g(X)\cdot g^*(X)<0}]$$

Moreover, we have $\mathcal{R}(g) - \mathcal{R}(g^*) \leq \mathbb{E}|2\eta(X) - 1 - g(X)|$, and as a matter of fact, for any function $b : \mathbb{R} \to \mathbb{R}$ that preserves the sign (that is $b(\mathbb{R}_+) \subset \mathbb{R}_+$ and $b(\mathbb{R}_-) \subset \mathbb{R}_-$), we have

$$\mathcal{R}(g) - \mathcal{R}(g^*) \leq \mathbb{E}[|2\eta(X) - 1 - b(g(X))|]$$

Proof. Recall that $\eta(X) = \mathbb{P}(Y = 1 \mid X)$. We express the excess risk as:

$$\begin{aligned}\mathcal{R}(g) - \mathcal{R}(g^*) &= \mathbb{E}\Big[\mathbb{E}\big[\mathbb{1}_{\operatorname{sgn}(g(X))\neq Y} - \mathbb{1}_{\operatorname{sgn}(g^*(X))\neq Y} \mid X\big]\Big] \\ &= \mathbb{E}\Big[\mathbb{E}\big[\mathbb{1}_{\operatorname{sgn}(g(X))\neq 1} - \mathbb{1}_{\operatorname{sgn}(g^*(X))\neq 1} \mid X, Y = 1\big]\eta(X) + \big[\mathbb{1}_{\operatorname{sgn}(g(X))\neq 0} - \mathbb{1}_{\operatorname{sgn}(g^*(X))\neq 0} \mid X, Y = 0\big](1 - \eta(X))\Big]\end{aligned}$$

by definition of the $0-1$ loss. For any given $X \in \mathcal{X}$, we can look at the two possible case for the signs of $\eta(X) - 1/2$ and $g(X)$ that lead to different predictions for g and g^*, namely

(a) for $\eta(X) > 1/2$ and $g(X) < 0$, the expectation is $\eta(X) - (1 - \eta(X)) = 2\eta(X) - 1 > 0$; and

(b) for $\eta(X) < 1/2$ and $g(X) > 0$, we get $1 - 2\eta(X) > 0$

By combining these two cases into the condition $g(X) \cdot g^*(X) < 0$ and the condition expectation $|2\eta X - 1|$, we get

$$\mathcal{R}(g) - \mathcal{R}(g^*) = \mathbb{E}\Big[|2\eta(X) - 1| \cdot \mathbb{1}_{\operatorname{sgn}(g(X))\cdot\operatorname{sgn}(g^*(X))<0}\Big]$$

which is just the first result.

For the seoncd result, we simply use the fact that if $g(X) \cdot g^*(X) < 0$, then, by splitting the cases in two (the first one being $\eta(X) > 1/2$ and $g(X) < 0$, the second one being $\eta(X) < 1/2$ and $g(X) > 0$), we get

$$|2\eta(X) - 1| \leq |2\eta(X) - 1 - g(X)|$$

As long as the funcition b presever the sign of $g(X)$, we obtain the last result. □

We see that the excess risk is the expectation of a quantity $|2\eta(X) - 1| \cdot \mathbb{1}_{g(X)\cdot g^*(X)<0}$, which is equal to 0 if the classification is the same as the Bayes predictor and equal to $|2\eta(X) - 1|$ otherwise. On the other hand, the excess conditional Φ-risk is the quantity

$$\mathcal{R}_\Phi(g) - \mathcal{R}_\Phi^* = \eta(X)\Phi(g(X)) + (1 - \eta(X))\Phi(-g(X)) - \inf_\alpha\{\eta(X)\Phi(\alpha) + (1 - \eta(X))\Phi(-\alpha)\}$$

which, as a function of $g(X)$, is the deviation between a convex function of $g(X)$ and its minimum value. We simply need to relate it to the quantity $|2\eta(X)-1|\cdot\mathbb{1}_{g(X)\cdot g^*(X)<0}$ above for any $x \in \mathcal{X}$ and take expectations.

Bartlett, Jordan, and McAuliffe 2006 proposes a general framework. Here we will only consider the hinge loss and smooth losses for simplicity.

- **Hinge Loss**. For the hinge loss $\Phi(\alpha) = (1-\alpha)^+ = \max\{1-\alpha, 0\}$, we can easily compute the minimizer of the conditional Φ-risk (which leads to the minimizer of the Φ-risk). Indeed, we need to minimizer $\eta(X)(1-\alpha)^+ + (1-\eta(X))(1+\alpha)^+$, which is a piecewise affine function with kinks at -1 and 1, with a minimizer attained at $u = 1$ for $\eta(X) > 1/2$, and symmetrically at $u = -1$ for $\eta(X) < 1/2$, with a minimum conditional Φ-risk equal to $2\min\{1 - \eta(X), \eta(X)\}$.

  The two excess risks are plotted below for the hinge loss and the 0-1 loss, for $\eta(X) > 1/2$, showing pictorially that the conditional excess Φ-risk is greater than the excess risk.

  This leads to the calibration function $H(\sigma) = \sigma$ for the hinge loss.

  Note that when the Bayes risk is zero, that is, $\eta(X) \in \{0, 1\}$ almost surely, then using the fact that the hinge loss is an upper-bound on the 0-1 loss is enough to show that the excess risk is less than the excess Φ-risk (indeed, the two optimal risk $\mathcal{R}^*$ and $\mathcal{R}_\Phi^*$ are equal to zero).

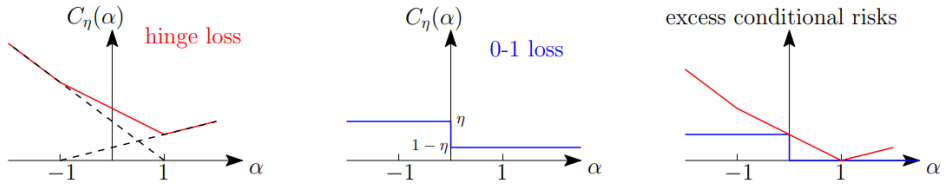


Figure 4: The excess risk of hinge loss and 0-1 loss

- **Smooth Loss**. We consider smooth losses of the form (up to additive and multiplicative constants) $\Phi(v) - a(v) - v$, where $a(v) = \frac{1}{2}v^2$ for the quadratic loss, $a(v) = 2\log(e^{v/2} + e^{-v/2})$ for the logistic loss. We assume that a is even ($a(-v) = a(v)$), $a(0) = 0$, a is β-smooth (that is, $a''(v) \leq \beta$ for all $v \in \mathbb{R}$). This implies that for all $v \in \mathbb{R}$,

$$a(v) - \alpha v - \inf_{w \in \mathbb{R}} \{a(w) - \alpha w\} \geq \frac{1}{2\beta}|\alpha - a'(v)|^2$$

leading to

$$\begin{aligned}
\mathcal{R}_\Phi(g) - \mathcal{R}_\Phi^* &= \mathbb{E}[a(g(X)) - (2\eta(X) - 1)g(X) - \inf_{w \in \mathbb{R}} a(w) - (2\eta(X) - 1)w] \\
&\geq \frac{1}{2\beta}\mathbb{E}|2\eta(X) - 1 - a'(g(X))|^2 \quad \text{by property above} \\
&\geq \frac{1}{2\beta}(\mathbb{E}|2\eta(X - 1 - a'(g(X))|)^2 \quad \text{by Jensens's inequality} \\
&\geq \frac{1}{2\beta} \quad \text{by Lemma 2.1}
\end{aligned}$$

This leads to the calibration function $H(\sigma) = \sqrt{\sigma}$ for the square loss and $H(\sigma) = \sqrt{2\sigma}$ for the logistic loss.

Remark 2.2. Show that the function a^* satisfies $a^*\big(\mathcal{R}(g) - \mathcal{R}^*\big) \leq \mathcal{R}_\Phi(g) - \mathcal{R}_\Phi^*$ for any function $g : \mathcal{X} \to \mathbb{R}$.

We can make the following observations:

- For the (non-smooth) hinge loss, the calibration function is identity, so if the excess Φ-risk goes to zero at a certain rate, the excess risk is goes to zero at the same rate; whereas for smooth losses, the upper-bound only ensures a (worse) rate with a square root. Therefore, when going from the excess Φ-risk to the excess risk, that is, after thresholding the function g at zero, the observed rates may be worse.
- Note that the noiseless case when $\eta(X) \in \{0, 1\}$ (zero Bayes risk) leads to stronger calibration function, as well as a series of intermediate "low-noise" conditions.

2.1.5 Impact on Approximation Errors

For the same classification problem, several convex surrogates can be used. While the Bayes classifier is always the same, that is,

$$f^*(X) = \operatorname{sgn}(2\eta(X) - 1)$$

the minimmizer of the testing Φ-risk will be different. For example, for the hinge loss, the minimizer $g(X)$ is exactly $\text{sgn}(2\eta(X)-1)$, while for losses of the form like above $\Phi(v) = a(v) - v$, we have $a'(g(X)) = 2\eta(X) - 1$, and thus for the square loss $g(X) = 2\eta(X) - 1$, while for the logistic loss, one can check that $g(X) = \text{atanh}(2\eta(X) - 1)$ (hyperbolic arc tangent). See example below, with $\mathcal{X} = \mathbb{R}$ and Gaussian class conditional densities.

2.2 Risk Minimization Decomposition

We consider a family $\mathcal{F}$ of prediction functions $f : \mathcal{X} \to \mathcal{Y}$. Empirical risk mnimization aims at finding

$$\hat{f} \in \arg\min_{f \in \mathcal{F}} \hat{\mathcal{R}}(f) = \frac{1}{n} \sum_{i=1}^{n} l(Y_i, f(X_i))$$

We can decompose the risk as follows into two terms:

$$\begin{aligned} \mathcal{R}(\hat{f}) - \mathcal{R}^* &= \Big\{ \mathcal{R}(\hat{f}) - \inf_{f' \in \mathcal{F}} \mathcal{R}(f') \Big\} + \Big\{ \inf_{f' \in \mathcal{F}} \mathcal{R}(f') - \mathcal{R}^* \Big\} \\ &= \text{estimation error} \quad + \quad \text{approximation error} \end{aligned} \tag{2.8}$$

A classical example is the situation where the family of functions is parameterized by a subset of $\mathbb{R}^d$, that is, $\mathcal{F} = \{f_\theta, \theta \in \Theta\}$ for $\Theta \subset \mathbb{R}^d$. This includes neural networks and the simplest case of linear model of the form $f_\theta(x) = \theta^T \varphi(x)$, for a certain feature vector $\varphi(x)$.

2.3 Approximation Error

Bounding the approximation error corresponds to bounding $\inf_{f \in \mathcal{F}} \mathcal{R}(f) - \mathcal{R}^*$ and requires assumptions on the Bayes predictor f^* to achieve non-trival learning rates.

Here we will focus on $\mathcal{F} = \{f_\theta, \theta \in \Theta\}$ for $\Theta \subset \mathbb{R}^d$ and convex Lipschitz-continuous losses, assuming that θ_* is the minimizer of $\mathcal{R}(f_\theta)$ over $\theta \in \mathbb{R}^d$ (typically, it does not belong to Θ). This implies that the approximation error decomposes into

$$\inf_{\theta \in \Theta} \mathcal{R}(f_\theta) - \mathcal{R}^* = \left(\inf_{\theta \in \Theta} \mathcal{R}(f_\theta) - \inf_{\theta \in \mathbb{R}^d} \mathcal{R}(f_\theta) \right) + \left(\inf_{\theta \in \mathbb{R}^d} \mathcal{R}(f_\theta) - \mathcal{R}^* \right) \tag{2.9}$$

- the second term $\inf_{\theta \in \mathbb{R}^d} \mathcal{R}(f_\theta) - \mathcal{R}^*$ is the incompressible error coming from the chosen of models f_θ
- the first term $\theta \mapsto \mathcal{R}(f_\theta) - \inf_{\theta \in \mathbb{R}^d} \mathcal{R}(f_\theta)$ is positive on $\mathbb{R}^d$, which can be typically upperbounded by a certain norm $\Omega(\theta - \theta_*)$. Hence it represents a "distance" between minimizer θ_* and set Θ on $\mathbb{R}$

For example, if the loss $l(y, \hat{y})$ which is considered as G-Lipschitz-continuous with respect to the second variable $\hat{y}$ (possible for regression or convex surrogate for binary classification), we have

$$\mathcal{R}(f_\theta) - \mathcal{R}(f_{\theta'}) = \mathbb{E}\Big[l\big(Y, f_\theta(X)\big) - l\big(Y, f_{\theta'}(X)\big) \Big] \leq G \cdot \mathbb{E}\big[|f_\theta(X) - f_{\theta'}(X)| \big]$$

and hence the first term is upper bounded by G times the smallest distance between f_{θ_*} and $\mathcal{F} = \{f_\theta, \theta \in \Theta\}$. A classical example will be $f_\theta(x) = \theta^T \varphi(x)$, and $\Theta = \{\theta \in \mathbb{R}^d, ||\theta||_2 \leq D\}$, leading to the upper bound

$$\inf_{\theta \in \Theta} \mathcal{R}(f_\theta) - \inf_{\theta \in \mathbb{R}^d} \mathcal{R}(f_\theta) \leq G \cdot \mathbb{E}\big[||\varphi(x)||_2 \big] \big(||\theta_*||_2 - D \big)^+$$

which is equalt zero if $||\theta_*||_2 \leq D$.

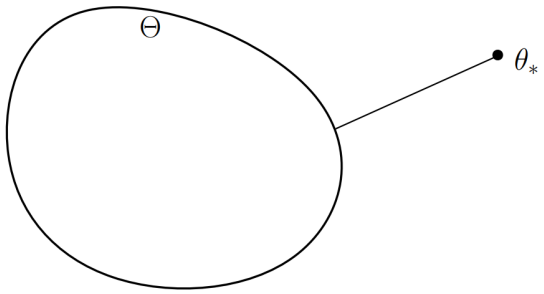


Figure 5: The distance between minimizer θ_* and set Θ on $\mathbb{R}$

2.4 Estimation Error

The estimation error is often decomposed using the minimizer of the expected risk for our class of models $\mathcal{F}$, $g \in \arg\min_{g\in\mathcal{F}} \mathcal{R}(g)$; and the minimizer of the empirical risk, $\hat{f} \in \arg\min_{f\in\mathcal{F}} \hat{\mathcal{R}}(f)$. That is

$$\begin{aligned}
\mathcal{R}(\hat{f}) - \inf_{f\in\mathcal{F}} \mathcal{R}(f) = \mathcal{R}(\hat{f}) - \mathcal{R}(g) &= \left\{\mathcal{R}(\hat{f}) - \hat{\mathcal{R}}(\hat{f})\right\} + \left\{\hat{\mathcal{R}}(\hat{f}) - \hat{\mathcal{R}}(g)\right\} + \left\{\hat{\mathcal{R}}(g) - \mathcal{R}(g)\right\} \\
&\leq \sup_{f\in\mathcal{F}} \left\{\mathcal{R}(f) - \hat{\mathcal{R}}(f)\right\} + \left\{\hat{\mathcal{R}}(\hat{f}) - \hat{\mathcal{R}}(g)\right\} + \sup_{f\in\mathcal{F}} \left\{\hat{\mathcal{R}}(f) - \mathcal{R}(f)\right\} \\
&\leq \sup_{f\in\mathcal{F}} \left\{\mathcal{R}(f) - \hat{\mathcal{R}}(f)\right\} + 0 + \sup_{f\in\mathcal{F}} \left\{\hat{\mathcal{R}}(f) - \mathcal{R}(f)\right\} \\
&\leq 2 \sup_{f\in\mathcal{F}} \left|\hat{\mathcal{R}}(f) - \mathcal{R}(f)\right|
\end{aligned} \tag{2.10}$$

The third inequality holds because, $\hat{f}$ is the minimizer of empirical risk, so we have $\hat{\mathcal{R}}(\hat{f}) - \hat{\mathcal{R}} \leq 0$.

Remark 2.3. Now, we can make the following observations:

- When $\hat{f}$ is not the global minimizer of $\hat{\mathcal{R}}$ but simply satisfies $\hat{\mathcal{R}}(\hat{f}) \leq \inf_{f\in\mathcal{F}} \hat{\mathcal{R}}(f) + \varepsilon$, then the *optimization error* ε has to be added to the bound above
- The uniform deviation grows with the size of $\mathcal{F}$, and decays with n.
- A key issue is that we need a **uniform control** for all $f \in \mathcal{F}$: with a single f, we could apply any concentration inequality to the random variable $l(Y, f(X))$ to obtain a bound in $\mathcal{O}(1/\sqrt{n})$; however, when controlling the maximal deviations over many value of f, there is always a small chance that one of these deviations get large. We thus need an explicit control of this phenomenon, which we can focus on the expectation alone, see section 2.4.1.

In general, there are two ways to bound the supremum of empirical process $\sup_{f\in\mathcal{F}} \left|\hat{\mathcal{R}}(f) - \mathcal{R}(f)\right|$.

- Directly bound $\sup_{f\in\mathcal{F}} |\hat{\mathcal{R}}(f) - \mathcal{R}(f)|$ with high probability (note that $\hat{\mathcal{R}}(f)$ here is a random variable, so we can bound it with high probability)
- Bound the uniform deviaiton of $\sup_{f\in\mathcal{F}} |\hat{\mathcal{R}}(f) - \mathcal{R}(f)|$ from its expectation; and then bound the expectation $\mathbb{E}[\sup_{f\in\mathcal{F}} |\hat{\mathcal{R}}(f) - \mathcal{R}(f)|]$

Before we go deep into the theory of uniform convergence, we see some simple examples.

2.4.1 Uniform Deviation from Expectation

Let

$$H(Z_1, \ldots, Z_n) = \sup_{f \in \mathcal{F}} \Big\{ \hat{\mathcal{R}}(f) - \mathcal{R}(f) \Big\}$$

where the random variables $z_i = (x_i, y_i)$ are independent and identically distributed, and $\hat{\mathcal{R}}(f) = \frac{1}{n} \sum_{i=1}^{n} l(Y_i, f(X_i))$. We let l_∞ be the maximal absolute value of the loss functions for all (X, Y) in the support of the data generating distribution and $f \in \mathcal{F}$, that is $l_\infty = \max_i |l(Y_i, f(X_i))|$.

When changing a single $Z_i \in \mathcal{X} \times \mathcal{Y}$ into $Z_i' \in \mathcal{X} \times \mathcal{Y}$, the bounded difference of H is almost surely at most $\frac{2}{n} l_\infty$, that is because

$$\begin{aligned}
\Big| H(Z_1, \ldots, Z_i, \ldots, Z_n) - H(Z_1, \ldots, Z_i', \ldots Z_n) \Big| &= \left| \sup_{f \in \mathcal{F}} \Big\{ \hat{\mathcal{R}}(f) - \mathcal{R}(f) \Big\} - \sup_{f \in \mathcal{F}} \Big\{ \hat{\mathcal{R}}'(f) - \mathcal{R}'(f) \Big\} \right| \\
&\leq \left| \sup_{f \in \mathcal{F}} \hat{\mathcal{R}}(f) - \sup_{f \in \mathcal{F}} \hat{\mathcal{R}}'(f) \right| \\
&\leq \left| \sup_{f \in \mathcal{F}} \Big\{ \hat{\mathcal{R}}(f) - \hat{\mathcal{R}}'(f) \Big\} \right| \\
&= \left| \sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} (l(Z_i) - l(Z_i')) \right\} \right| \leq \frac{2}{n} l_\infty
\end{aligned}$$

where the third inequality holds beacue in general, $\sup_h A(f) - \sup_f B(f) \leq \sup_f [A(f) - B(f)]$. Now, we can apply the MacDiarmid inequality,

$$\mathbb{P}\Big(H(Z_1, \ldots, Z_n) - \mathbb{E}[H(Z_1, \ldots, Z_n)] \geq t \Big) \leq \exp\left(- \frac{2t^2}{\sum_{i=1}^{n} (\frac{2}{n} l_\infty)^2} \right) = \exp\left(- \frac{nt^2}{2 l_\infty^2} \right)$$

By setting $\delta = \exp\left(-nt^2 / 2 l_\infty^2 \right)$, which leads to $t = l_\infty \sqrt{\frac{2 \log(1/\delta)}{n}}$, with probability greater than $1 - \delta$, we have

$$H(Z_1, \ldots, Z_n) - \mathbb{E}[H(Z_1, \ldots, Z_n)] \leq l_\infty \sqrt{\frac{2 \log(1/\delta)}{n}}$$

Therefore, recall that $H(Z_1, \ldots, Z_n) = \sup_{f \in \mathcal{F}} \Big\{ \hat{\mathcal{R}}(f) - \mathcal{R}(f) \Big\}$, we have

$$\sup_{f \in \mathcal{F}} \Big\{ \hat{\mathcal{R}}(f) - \mathcal{R}(f) \Big\} \leq l_\infty \sqrt{\frac{2 \log(1/\delta)}{n}} + \mathbb{E}\left[\sup_{f \in \mathcal{F}} \Big\{ \hat{\mathcal{R}}(f) - \mathcal{R}(f) \Big\} \right] \tag{2.11}$$

We thus only need to bound the expectation of $\sup_{f \in \mathcal{F}} \Big\{ \hat{\mathcal{R}}(f) - \mathcal{R}(f) \Big\}$, and add on top of the above result.

2.4.2 Linear Hypothesis Space

In this case, we consider the case when the hypothesis function space $\mathcal{F} = \{ \theta^\top \varphi(x) \mid ||\theta||_2 \leq D \}$ is linear with l_2-ball constraint (l_2-norm bounded by D), and the loss function is quadratic, that is

$$l(Y, f(X)) = (Y - \theta^\top \varphi(X))^2$$

From these we get

$$\hat{\mathcal{R}}(f) - \mathcal{R}(f) = \theta^\top \left(\frac{1}{n} \sum_{i=1}^{n} \varphi(X_i)\varphi(X_i)^\top - \mathbb{E}[\varphi(X)\varphi(X)^\top] \right) \theta$$
$$- 2\theta^\top \left(\frac{1}{n} \sum_{i=1}^{n} Y_i \varphi(X_i) - \mathbb{E}[Y\varphi(X)] \right) + \left(\frac{1}{n} \sum_{i=1}^{n} Y_i^2 - \mathbb{E}[Y^2] \right)$$

Hence, the supremum can be upper bounded in closed from as

$$\sup_{||\theta||_2 \le D} |\hat{\mathcal{R}}(f) - \mathcal{R}(f)| \le D^2 \left|\left| \frac{1}{n} \sum_{i=1}^{n} \varphi(X_i)\varphi(X_i)^\top - \mathbb{E}[\varphi(X)\varphi(X)^\top] \right|\right|_{op}$$
$$+ 2D \left|\left| \frac{1}{n} \sum_{i=1}^{n} Y_i \varphi(X_i) - \mathbb{E}[Y\varphi(X)] \right|\right|_2 + \left| \frac{1}{n} \sum_{i=1}^{n} Y_i^2 - \mathbb{E}[Y^2] \right|$$

where $||M||_{op}$ is the operator norm of the matrix M defined as $||M||_{op} = \sup_{||u||_2=1} ||Mu||_2$.

- Bounding the Matrix

  Suppose $\varphi(\cdot)$ is a d-dimensional function of X. Let

  $$M_i = \varphi(X_i)\varphi(X_i)^\top - \mathbb{E}[\varphi(X)\varphi(X)^\top]$$

  Then M_i is a $d \times d$ symmetric matrix with $\mathbb{E}[M_i = 0]$. Given a sequence of n i.i.d symmetric matricies $\{M_i,\ i = 1, \ldots, n\}$, we can apply matrix Hoeffding's inequality and get

  $$\mathbb{P}\left(\lambda_{\max} \left(\frac{1}{n} \sum_{i=1}^{n} M_i \right) \ge t \right) \le d \cdot \exp\left(-\frac{nt^2}{8\sigma^2} \right)$$

  where $\sigma^2 = \lambda_{\max}(\bar{M})$. With probability $1 - \delta$, we have $t = \sigma \sqrt{\frac{8 \log(d/\delta)}{n}}$ and

  $$\lambda_{\max} \left(\frac{1}{n} \sum_{i=1}^{n} M_i \right) \le \sigma \sqrt{\frac{8 \log(d/\delta)}{n}}$$

  Notice that $\bar{M} = (\frac{1}{n} \sum_{i=1}^{n} M_i)$ is also a symmetric matrix, for any vector θ, we have

  $$\theta^T \left(\frac{1}{n} \sum_{i=1}^{n} M_i \right) \theta \le \lambda_{\max} \left(\frac{1}{n} \sum_{i=1}^{n} M_i \right) \theta^T \theta \le D^2 \sigma \sqrt{\frac{8 \log(d/\delta)}{n}}$$

- Bounding the Vector

Suppose $\varphi(X)$ is a d-dimensional vector, then we're going to find a uniform bound for its l_2-norm.

$$\begin{aligned}
\mathbb{P}\left(\left|\left|\frac{1}{n}\sum_{i=1}^{n} Y_i\varphi(X_i) - \mathbb{E}[Y_i\varphi(X_i)]\right|\right|_2 \geq t\right) &= \mathbb{P}\left(\left[\sum_{j=1}^{d}\left|\frac{1}{n}\sum_{i=1}^{n} Y_i\varphi_j(X_i) - \mathbb{E}[Y_i\varphi_j(X_i)]\right|^2\right]^{1/2} \geq t\right) \\
&= \mathbb{P}\left(\sum_{j=1}^{d}\left|\frac{1}{n}\sum_{i=1}^{n} Y_i\varphi_j(X_i) - \mathbb{E}[Y_i\varphi_j(X_i)]\right|^2 \geq t^2\right) \\
&\leq \sum_{j=1}^{d}\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Y_i\varphi_j(X_i) - \mathbb{E}[Y_i\varphi_j(X_i)]\right|^2 \geq \frac{t^2}{d}\right) \quad \text{(union bound)} \\
&= \sum_{j=1}^{d}\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Y_i\varphi_j(X_i) - \mathbb{E}[Y_i\varphi_j(X_i)]\right| \geq \frac{t}{\sqrt{d}}\right)
\end{aligned}$$

Now, if we assume $|Y\varphi_j(X)|$ are uniformly bounded by constant c for any $j \in \{1, \ldots, d\}$, we can apply Hoeffding's inequality and get

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Y_i\varphi_j(X_i) - \mathbb{E}[Y_i\varphi_j(X_i)]\right| \geq t\right) \leq 2\exp\left(-\frac{2nt^2}{dc^2}\right)$$

which leads to the fact that

$$\mathbb{P}\left(\left|\left|\frac{1}{n}\sum_{i=1}^{n} Y_i\varphi(X_i) - \mathbb{E}[Y_i\varphi(X_i)]\right|\right|_2 \geq t\right) \leq \sum_{j=1}^{d} 2\exp\left(-\frac{2nt^2}{dc^2}\right) = 2d\exp\left(-\frac{2nt^2}{dc^2}\right) \tag{2.12}$$

Finally, with probability $1-\delta$, we have

$$\left|\left|\frac{1}{n}\sum_{i=1}^{n} Y_i\varphi(X_i) - \mathbb{E}[Y_i\varphi(X_i)]\right|\right|_2 \leq c\sqrt{\frac{d\log(2d/\delta)}{2n}}$$

- Bouding the Scalar

  Similarily, suppose $Z = Y^2$ is a bounded variable with support $[a, b]$, then applying the Hoeffding's bound, we have with probability $1-\delta$

$$\left|\frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \mathbb{E}[Y^2]\right| \leq (b-a)\sqrt{\frac{\log(2/\delta)}{2n}}$$

Finally, by letting $\delta' = \delta/3$ in each of the three bounds above and applying union bound again, we can upper-bond the empirical process with probability $1-\delta$,

$$\begin{aligned}
\sup_{||\theta||_2 \leq D} |\hat{\mathcal{R}}(f) - \mathcal{R}(f)| &\leq D^2\sigma\sqrt{\frac{8\log(3d/\delta)}{n}} + 2Dc\sqrt{\frac{\log(6d/\delta)}{2n}} + (b-a)\sqrt{\frac{\log(6/\delta)}{2n}} \\
&\approx (4D^2\sigma + 2Dc + b - a)\sqrt{\frac{\log(6/\delta)}{2n}} = \mathcal{O}\left(\frac{1}{n}\right)
\end{aligned}$$

2.4.3 Finite Hypothesis Space

We assume in this section that the loss functions $l(Y, f(X))$ are bounded between $-l_\infty$ and l_∞.

Direct Bounding Approach. Using the upper-bound $2\sup_{f\in\mathcal{F}}|\hat{\mathcal{R}}(f)-\mathcal{R}(f)|$ on the estimation error, we have the union bound:

$$\mathbb{P}\left(\mathcal{R}(\hat{f})-\inf_{f\in\mathcal{F}}\mathcal{R}(f)\geq t\right)\leq\mathbb{P}\left(2\sup_{f\in\mathcal{F}}|\hat{\mathcal{R}}(f)-\mathcal{R}(f)|\geq t\right)\leq\sum_{f\in\mathcal{F}}\mathbb{P}\left(2\left|\hat{\mathcal{R}}(f)-\mathcal{R}(f)\right|\geq t\right)$$

We have, for $f\in\mathcal{F}$ fixed, $\hat{\mathcal{R}}(f)=\frac{1}{n}\sum_{i=1}^{n}l(Y_i,f(X_i))$ and we can apply Hoeffding's inequality to bound each $\mathbb{P}\left(2|\hat{\mathcal{R}}(f)-\mathcal{R}(f)|\geq t\right)$, leading to

$$\mathbb{P}\left(\mathcal{R}(\hat{f})-\inf_{f\in\mathcal{F}}\mathcal{R}(f)\geq t\right)\leq\sum_{f\in\mathcal{F}}2\exp\left(-\frac{nt^2}{2l_\infty^2}\right)=2|\mathcal{F}|\exp\left(-\frac{nt^2}{2l_\infty^2}\right)$$

Thus, by setting $\delta=2|\mathcal{F}|\exp\left(-nt^2/2l_\infty^2\right)$, and finding the corresponding t, with probability greater than $1-\delta$,

$$\mathcal{R}(\hat{f})-\inf_{f\in\mathcal{F}}\mathcal{R}(f)\leq 2l_\infty\sqrt{\frac{\log(2|\mathcal{F}|/\delta)}{n}}\tag{2.13}$$

Bounding the Expectation. In terms of expectation, we get (using the proof of the expectation of the maximum, which apply both bounded and sub-Gaussian random variables)

$$\mathbb{E}\left[\mathcal{R}(\hat{f})-\inf_{f\in\mathcal{F}}\mathcal{R}(f)\right]\leq 2\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\hat{R}(f)-\mathcal{R}(f)\right|\right]\leq 2l_\infty\sqrt{\frac{2\log|\mathcal{F}|}{n}}\tag{2.14}$$

Here is the proof, when function family $\mathcal{F}$ is finite, we have

$$\begin{aligned}
\mathbb{E}\left[\sup_{f\in\mathcal{F}}|\hat{\mathcal{R}}(f)-\mathcal{R}(f)|\right]&=\mathbb{E}\left[\max\left\{\hat{R}(f_1)-\mathcal{R}(f),\ldots,\hat{\mathcal{R}}(f_{|\mathcal{F}|})-\hat{\mathcal{R}}(f_{|\mathcal{F}|})\right\}\right]\\
&=\mathbb{E}\left[\frac{1}{n}\log e^{t\max\left\{\hat{R}(f_1)-\mathcal{R}(f),\ldots,\hat{\mathcal{R}}(f_{|\mathcal{F}|})-\hat{\mathcal{R}}(f_{|\mathcal{F}|})\right\}}\right]\\
&\leq\frac{1}{t}\log\mathbb{E}\left[e^{t\max\left\{\hat{R}(f_1)-\mathcal{R}(f),\ldots,\hat{\mathcal{R}}(f_{|\mathcal{F}|})-\hat{\mathcal{R}}(f_{|\mathcal{F}|})\right\}}\right]\quad\text{(Jensen's Inequality)}\\
&=\frac{1}{t}\log\mathbb{E}\left[\max\left\{e^{t(\hat{\mathcal{R}}(f_1)-\mathcal{R}(f_1))}+\cdots+e^{t(\hat{\mathcal{R}}(f_{|\mathcal{F}|})-\mathcal{R}(f_{|\mathcal{F}|}))}\right\}\right]\\
&\leq\frac{1}{t}\log\mathbb{E}\left[e^{t(\hat{\mathcal{R}}(f_1)-\mathcal{R}(f_1))}+\cdots+e^{t(\hat{\mathcal{R}}(f_{|\mathcal{F}|})-\mathcal{R}(f_{|\mathcal{F}|}))}\right]\quad\text{(bounding the max by the sum)}
\end{aligned}$$

Since the Chernoff bound of bounded loss $l(Y,f(X))$ is

$$\begin{aligned}
\mathbb{E}\left[e^{t(\hat{\mathcal{R}}(f_k)-\mathcal{R}(f_k))}\right]&=\prod_{i=1}^{n}\mathbb{E}\left[e^{\frac{t}{n}(l(Y_i,f_k(X_i))-\mathbb{E}[l(Y_i,f_k(X_i))])}\right]\\
&\leq\prod_{i=1}^{n}\exp\left(\frac{l_\infty^2t^2}{2n^2}\right)=\exp\left(\frac{l_\infty^2t^2}{2n}\right)
\end{aligned}$$

Substitute the result back to the expectation of estimation error, we get

$$\begin{aligned}\mathbb{E}\left[\sup_{f\in\mathcal{F}}|\hat{\mathcal{R}}(f)-\mathcal{R}(f)|\right] &\leq \frac{1}{t}\log\mathbb{E}\left[e^{t(\hat{\mathcal{R}}(f_1)-\mathcal{R}(f_1))}+\cdots+e^{t(\hat{\mathcal{R}}(f_{|\mathcal{F}|})-\mathcal{R}(f_{|\mathcal{F}|}))}\right]\\ &\leq \frac{1}{t}\log\left(|\mathcal{F}|\exp\left(\frac{l_\infty^2 t^2}{2n}\right)\right)\\ &= \frac{\log|\mathcal{F}|}{t}+l_\infty^2\frac{t}{2n}\end{aligned}$$

Minimizer over t, we get $t=\frac{\sqrt{2n\log|\mathcal{F}|}}{l_\infty}$, and therefore

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}|\hat{\mathcal{R}}(f)-\mathcal{R}(f)|\right]\leq l_\infty\sqrt{\frac{2\log|\mathcal{F}|}{n}}$$

Finaly, plugging the above result into the equation 2.11, we have with probabability $1-\delta$

$$\mathcal{R}(\hat{f})-\inf_{f\in\mathcal{F}}\mathcal{R}(f)\leq 2\sup_{f\in\mathcal{F}}\left\{\hat{\mathcal{R}}(f)-\mathcal{R}(f)\right\}\leq 2l_\infty\left(\sqrt{\frac{2\log(1/\delta)}{n}}+\sqrt{\frac{2\log|\mathcal{F}|}{n}}\right) \tag{2.15}$$

2.4.4 Beyond the Finite Hypothesis Space

The simple idea behind covering numbers is to deal with function spaces (with infinitely many elements by approximating them through a finite numner of elements. This is often refered to as an "ε-net argument".

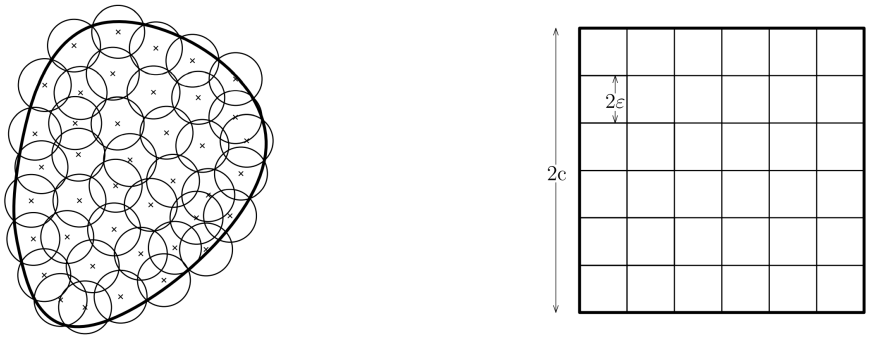


Figure 6: The left picture is an example in two dimensions of a covering with Euclidean balls; The right is an example of l_∞-balls

Definition 2.2 (Covering Numbers)**.** We assume there exists $m=m(\varepsilon)$ elements $f_1,\ldots,f_m$ such that for any $f\in\mathcal{F}$, there exists $i\in\{1,\ldots,n\}$ such that $d(f,f_i)\leq\varepsilon$. The minimal possible number $m(\varepsilon)$ is the covering number of $\mathcal{F}$ at precision ε.

We first need to assume that the risks $\mathcal{R}$ and $\hat{\mathcal{R}}$ are regular, for example, they are G-Lipschitz-continuous with respect to some distance d on $\mathcal{F}$. Now, given a cover of $\mathcal{F}$, for all $f\in\mathcal{F}$, and with $(f_i)_{i\in\{1,\ldots,m_\varepsilon\}}$ the associated cover elements

$$\begin{aligned}|\hat{\mathcal{R}}(f)-\mathcal{R}(f)| &\leq |\hat{\mathcal{R}}(f)-\hat{\mathcal{R}}(f_i)|+|\hat{\mathcal{R}}(f_i)-\mathcal{R}(f_i)|+|\mathcal{R}(f_i)-\mathcal{R}(f)|\\ &\leq 2G\varepsilon+\sup_{i\in\{1,\ldots,m(\varepsilon)\}}|\hat{\mathcal{R}}(f_i)-\mathcal{R}(f_i)|\end{aligned}$$

Using bounds 2.14 on the expectation of the maximum (bounded random variables are sub-Gaussian), we have

$$\begin{aligned}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\hat{\mathcal{R}}(f)-\mathcal{R}(f)\right|\right] &\leq 2G\varepsilon+\mathbb{E}\left[\sup_{i\in\{1,\dots,m(\varepsilon)\}}\left|\hat{\mathcal{R}}(f_i)-\mathcal{R}(f_i)\right|\right] \\ &\leq 2G\varepsilon+l_\infty\sqrt{\frac{2\log m(\varepsilon)}{n}}\end{aligned} \tag{2.16}$$

The first term of the bound capture the estmation biased controlled by ε, while the second tem characterize the complexity of the covering.

- Therefore, if $m(\varepsilon)\sim\varepsilon^{-d}$, ignoring constants, we need to balance $\varepsilon+\sqrt{\frac{d\log(1/\varepsilon)}{n}}$, which leads to, with a choice of ε proportional to $1/\sqrt{n}$, to a rate proportional $\sqrt{\frac{d\log n}{n}}$
- Unfortunately, this often leads to a non-optimal dependence on sample size n (because of the existence of $\log n$), as the rate is essentially proportional to $\sqrt{d/n}$
- One very powerful tool that avoids these undesired dependences on dimension is Rademacher complexities or Gaussian complexities

3 PAC Learning and Uniform Convergence

3.1 PAC Learning

In the previous section, we have shown that for a finite hypothesis space, if the ERM rule with respect to that class is applied on a sufficiently large training sample (whose size is independent of the underlying distribution or labeling function) then the output hypothesis will be probably approximately correct. More generally, we now define *Probably Approximately Correct* (PAC) learning.

Definition 3.1 (PAC Learnability)**.** A hypothesis class $\mathcal{F}$ is PAC learnable if there exist a function $m_{\mathcal{F}}:(0,1)^2\to\mathbb{N}$ and a learning algorithm with the following property: for every $\varepsilon,\delta\in(0,1)$ and for every probability distribution P over $\mathcal{X}$, and for every labeling function $g:\mathcal{X}\to\{0,1\}$, if the realizable assumption holds with respect to $\mathcal{F},P,g$, then when running the learning algorithm on $m\geq m_{\mathcal{F}}(\varepsilon,\delta)$ i.i.d examples generated by P and labeled by g, the algorithms returns a hypothesis f, such that, with probability of at least $1-\delta$,

$$\mathcal{R}_{P,g}(f)\leq\varepsilon$$

where $\mathcal{R}_{P,g}(f)=\mathbb{E}_P[l(g(X),f(X))]$ and $l(y,\hat{y})$ is a loss function.

The definition of Probably Approximately Correct learnability contains two approximation parameters:

- the accuracy parameter ε determines how far the output classifier can be from the optimal one (approximately correct)
- and the confidence parameter δ indicating how likely the classifier is to meet the accuracy requriement (probably)

Sample Complexity. The function $m_{\mathcal{F}}:(0,1)^2\to\mathbb{N}$ determines the *sample complexity* of learning $\mathcal{F}$: that is, how many examples are required to guarantee a probably approximately correct solution. The sample complexity $m_{\mathcal{F}}$ is a function of accuracy (ε) and confidence (δ) parameters. It also depends on properties

of the hypothesis class $\mathcal{F}$ – for example, for a finite class we showed that the sample comolexity depends on log the size of $\mathcal{F}$ (see Section 2.4.3).

Note that if $\mathcal{F}$ is PAC learnable, there are many functions $m_{\mathcal{F}}$ that satisfy the requirement given in the definition of PAC learnability. Therefore, to be precise, we will define the sample complexity of learning $\mathcal{F}$ to be the "minimal function", in the sense that for any (ε, δ), $m_{\mathcal{F}}(\varepsilon, \delta)$ is the minimal integer that satisfies the requirements of PAC learning.

Corollary 3.1. Every finite hypothesis class is PAC learnable with sample complexity

$$m_{\mathcal{F}}(\varepsilon, \delta) \leq \left\lceil \frac{\log(|\mathcal{F}|/\delta)}{\varepsilon} \right\rceil$$

There are infinite classes that are learnable as well. Later on we will show that what determines the PAC learnability of a class is not its finiteness but rather a combinatorial measure called the *VC dimension*.

3.2 Agnostic PAC Learning

The model we have just described can be readily generalized, so that it can be made relevant to a wide scope of learning tasks. We consider generalizations in two aspects:

- Relaxing the realizability assumption
- Learning problems beyond binary classification

Definition 3.2 (Agnostic PAC Learnability)**.** A hypothesis class $\mathcal{F}$ is agnostic PAC learnable if there exist a function $m_{\mathcal{F}} : (0,1)^3 \to \mathbb{N}$ and a learning algorithm with the following property: for every $\varepsilon, \delta \in (0,1)$ and for every distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, when running the learning algorithm on $m \geq m_{\mathcal{F}}(\varepsilon, \delta)$ i.i.d. examples generated by probability distribution P, the algorithm returns a hypothesis h such that, with probability of at least $1 - \delta$ (over the choice of the m training examples),

$$\mathcal{R}(f) - \inf_{f' \in \mathcal{F}} \mathcal{R}(f') \leq \varepsilon$$

where here we simply denote $\mathcal{R} = \mathcal{R}_P$ as the expected risk.

3.3 Uniform Convergence

In this section, we will show that uniform convergence is sufficient for learnability. The idea behind the learning condition discussed here is very simple. Recall that, given a hypothesis class, $\mathcal{F}$, the empirical risk minimization (ERM) learning a paradigm works as follows: Uppon recieving a training sample S, the learner evaluates the risk of each f in $\mathcal{F}$ on the given sample and outputs a member of $\mathcal{F}$ that minimizes this empirical risk.

The hope is that an f minimizes the empirical risk with respect to S is a risk minimizer (or has risk close to the minimum) with respect to the true data probability distribution P as well. Recall that we have shown previously that

$$\mathcal{R}(\hat{f}) - \inf_{f \in \mathcal{F}} \mathcal{R}(f) \leq 2 \sup_{f \in \mathcal{F}} \left| \hat{\mathcal{R}}(f) - \mathcal{R}(f) \right| \tag{3.1}$$

Hence, for that, it suffices to ensure that the empirical risks of all members of $\mathcal{F}$ are good approximations of their true risk. Put another way, we need that uniformly over all hypotheses in the hypothesis class, the empirical risk will be close to the true risk, as formalized in the following.

Definition 3.3 (ε-representative sample)**.** A training set S is called ε-representative (w.r.t. domain $\mathcal{X}$, distribution P, hypothesis class $\mathcal{F}$ and loss function l) if

$$\forall\ f \in \mathcal{F},\ \left|\hat{\mathcal{R}}(f) - \mathcal{R}(f)\right| \leq \varepsilon$$

where here we dentoe $\hat{\mathcal{R}} = \mathcal{R}_S$ as the empirical risk with respect to sample S and $\mathcal{R} = \mathcal{R}_P$ the expected risk.

The next simple lemma states that whenever the sample is $\varepsilon/2$-representative, the ERM leanring rule is guaranteed to return a good hypothesis.

Lemma 3.2. Assume that a training set S is $\varepsilon/2$-representative (w.r.t. domain $\mathcal{X}$, distribution P, hypothesis class $\mathcal{F}$ and loss function l). Then, any output of $\mathrm{ERM}_{\mathcal{R}}(S)$, namely, any $\hat{f} \in \arg\min_{f\in\mathcal{F}} \hat{R}(f)$, satisifies

$$\mathcal{R}(\hat{f}) - \inf_{f\in\mathcal{F}} \mathcal{R}(f) \leq \varepsilon$$

Proof. From Eq.(3.1), we know that

$$\mathcal{R}(\hat{f}) - \inf_{f\in\mathcal{F}} \mathcal{R}(f) \leq 2 \sup_{f\in\mathcal{F}} \left|\hat{\mathcal{R}}(f) - \mathcal{R}(f)\right|$$

Since the training sample S is $\varepsilon/2$-representative, namely, $|\hat{\mathcal{R}}(f) - \mathcal{R}(f)| \leq \varepsilon/2$, we have

$$\mathcal{R}(\hat{f}) - \inf_{f\in\mathcal{F}} \mathcal{R}(f) \leq \varepsilon$$

□

The preceding lemma implies that to ensure that the **ERM is an agnostic PAC learner**, it suffices to show that with probability of at least $1-\delta$ over the random choice of a training set, it will be an ε-representative training set. The uniform convergence condition formalizes this requirement.

Definition 3.4 (Uniform Convergence)**.** We say that a hypothesis class $\mathcal{F}$ has the uniform convergence property (w.r.t. a domain $\mathcal{X}$ and a loss function l) if there exists a function

$$m_{\mathcal{F}}^{\mathrm{UC}} : (0,1)^2 \to \mathbb{N}$$

such that for every $\varepsilon, \delta \in (0,1)$ and for every probability distribution P over $\mathcal{X}$, if $\mathcal{S}$ is a sample of $m \geq m_{\mathcal{F}}^{\mathrm{UC}}(\varepsilon, \delta)$ examples drawn i.i.d from P, then, with probability of at least $1-\delta$, sample $\mathcal{S}$ is ε-represnetative.

Similar to the definition of sample complexity for PAC learning, the function $m_{\mathcal{F}}^{\mathrm{UC}}$ measures the (minimal) sample complexity of obtaining the uniform convergence property, namely, how many examples we need to ensure that with probability of at least $1-\delta$ the sample would be ε-representative.

Remark 3.1. The term *uniform* here refers to having a fixed sample size that works for all members of $\mathcal{F}$ and for all possible probability distributions P over the domain and some loss function.

Corollary 3.3. If a class $\mathcal{F}$ has the uniform convergence property with a function $m_{\mathcal{F}}^{UC}$ then the class is agnostically PAC learnable with the sample complexity $m_{\mathcal{F}}(\varepsilon, \delta) \leq m_{\mathcal{F}}^{UC}(\varepsilon/2, \delta)$. Furthermore, in that case, the $\mathrm{ERM}_{\mathcal{F}}$ paradigm is a successful agnostic PAC learner for $\mathcal{F}$.

4 Rademacher Complexity

In Section 3 we have shown that uniform convergence is a sufficient condition for learnability. In this section, we study the Rademacher complexity, which measures the rate of uniform convergence. We will provide generailization bounds based on this measure. To begin with, let's recall the definition of an ε-representative sample.

4.1 Motivation for Rademacher Complexity

Definition 4.1 (ε-representative sample)**.** A training set S is called ε-representative (w.r.t. domain $\mathcal{X}$, distribution P, hypothesis class $\mathcal{F}$ and loss function l) if

$$\sup_{f\in\mathcal{F}}\left|\hat{\mathcal{R}}(f)-\mathcal{R}(f)\right|\leq\varepsilon$$

We have shown that if S is an $\varepsilon/2$-representative sample, then the ERM rule is ε-consistent, namely, $\mathcal{R}(\hat{f})-\inf_{f\in\mathcal{F}}\mathcal{R}(f)\leq\varepsilon$, where $\hat{f}=\mathrm{ERM}_{\mathcal{F}}(S)$.

Now, for simplicity, we define a new variable Z over domain $\mathcal{Z}=\mathcal{X}\times\mathcal{Y}$, and a function h from hypothesis space $\mathcal{H}=\{h:(Y,X)\mapsto l(Y,f(X))\mid f\in\mathcal{F}\}$. The expected risk (w.r.t. probability distribution P) and the empirical risk (w.r.t. some sample set) are defined as follows:

$$\mathcal{R}(h)=\mathbb{E}[h(Z)]\quad\text{and}\quad\hat{\mathcal{R}}(h)=\frac{1}{n}\sum_{S}h(Z)$$

We define the *representativeness* of S with respect to $\mathcal{H}$ as the largest gap between the expected risk of a function h and its empirical risk, that is,

$$\mathrm{Rep}_P(\mathcal{H},S):=\sup_{h\in\mathcal{H}}\hat{R}(h)-\mathcal{R}(h)\tag{4.1}$$

Now suppose we would like to estimate the representativeness of S using the sample S only. One simple idea is to split S into two disjionit sets, $S=S_1\cup S_2$; refer S_1 as a training set and to S_2 as a validation set. We can then estimate the representativeness of S by

$$\sup_{h\in\mathcal{H}}\ \hat{\mathcal{R}}_{S_1}(h)-\hat{\mathcal{R}}_{S_2}(h)\tag{4.2}$$

This can be written more compactly by defining $\sigma=(\sigma_1,\ldots,\sigma)\in\{\pm1\}^n$ to be a vector such that $S_1=\{Z_{:}\sigma_i=1\}$ and $S_2=\{Z_i:\sigma_i=-1\}$. Then, if we further assume that $|S_1|=|S_2|$, then the equation above can be rewritten as

$$\sup_{h\in\mathcal{H}}\ \frac{2}{n}\sum_{i=1}^{n}\sigma_i h(Z_i)\tag{4.3}$$

The Rademacher complexity measure this idea by considering the expectation of the above with respect to a random choice of Rademacher variable σ. We will see the formal definition in next section.

4.2 Rademacher Complexity

Rememeber that our goal here is to provide an upper-bound on the empirical process $\sup_{f\in\mathcal{F}}\{\hat{\mathcal{R}}(f)-\mathcal{R}(f)\}$ introduced in Section 2.4, which happens to be equal to

$$\sup_{h\in\mathcal{H}}\left\{\frac{1}{n}\sum_{i=1}^{n}h(Z_i)-\mathbb{E}[h(Z)]\right\}$$

Later we will show that the empirical process is upper-bounded by two times the Rademacher complexity. Now, let's see the definition of empirical Rademacher complexity and then the expected Rademacher complexity.

Definition 4.2 (Empirical Rademacher Complexity)**.** Let $\mathcal{H}$ be a family of functions mapping from $\mathcal{Z}$ to $\mathbb{R}$, and $S_n=(Z_1,\ldots,Z_n)$ a fixed sample of size n with elements in $\mathcal{Z}$. Then the empirical Rademacher compelxity of $\mathcal{H}$ with respect to the sample S_n is defined as

$$\hat{R}_{S_n}(\mathcal{H})=\mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i h(Z_i)\right] \tag{4.4}$$

where $\sigma=(\sigma_1,\ldots,\sigma_n)^\top$, with σ_i's independent uniform random variables taking values in $\{\pm1\}$. The random variables σ_i are called Rademacher variables.

Let h_S denote the vector of values taken by function h over the sample S, namely, $h_S=(h(Z_1),\ldots,h(Z_n))^\top$. Then the empirical Rademacher complexity can be rewritten as

$$\hat{R}_S(\mathcal{H})=\mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}}\frac{\langle\sigma,h_S\rangle}{n}\right]$$

Thus, the empirical Rademacher complexity measures on average how well the function class $\mathcal{H}$ correlates with random noise on S. This describes the richness of the family $\mathcal{H}$: richer or more complex families $\mathcal{H}$ can generate more vectors h_S and thus better correlate with random noise, on average.

Definition 4.3 (Rademacher Complexity)**.** Let P denote the probability distribution according to which samples are drawn. For any integer $n\geq1$, the Rademacher complexity is the expectation of the empirical Rademahcer complexity over all samples of size n drawn i.i.d. with respect to P.

$$R_n(\mathcal{H})=\mathbb{E}_P\left[\hat{R}_{S_n}(\mathcal{H})\right]=\mathbb{E}_{\sigma,P}\left[\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i h(Z_i)\right] \tag{4.5}$$

Now, we show that, through a general "symmetrization" property, the Rademacher complexity $R_n(\mathcal{H})$ directly controls the expectation of empirical process, that is $\mathbb{E}[\sup_{f\in\mathcal{F}}(\hat{\mathcal{R}}(f)-\mathcal{R}(f))]$.

Theorem 4.1 (Symmetrization)**.** Given the Rademacher complexity of $\mathcal{H}$ defined in equaiton 4.5, we have

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{n}\sum_{i=1}^{n}h(Z_i)-\mathbb{E}[h(Z)]\right)\right]\leq2R_n(\mathcal{H}) \tag{4.6}$$

and

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left(\mathbb{E}[h(Z)]-\frac{1}{n}\sum_{i=1}^{n}h(Z_i)\right)\right]\leq2R_n(\mathcal{H}) \tag{4.7}$$

Proof. Let $\mathcal{D}' = \{Z_1', \ldots, Z_n'\}$ be an independent copy of the data $\mathcal{D} = \{Z_1, \ldots, Z_n\}$. Let $(\sigma_i)_{i\in\{1,\ldots,n\}}$ be i.i.d. Rademacher random variables, which are also independent of $\mathcal{D}$ and $\mathcal{D}'$. Using that for all $i \in \{1, \ldots, n\}$, $\mathbb{E}[h(Z_i') \mid \mathcal{D}] = \mathbb{E}[h(Z)]$, we have

$$\begin{aligned}\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left(\mathbb{E}[h(Z)] - \frac{1}{n}\sum_{i=1}^n h(Z_i)\right)\right] &= \mathbb{E}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}[h(Z_i') \mid \mathcal{D}] - \frac{1}{n}\sum_{i=1}^n h(Z_i)\right)\right] \\ &= \mathbb{E}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}[h(Z_i') - h(Z_i) \mid \mathcal{D}]\right)\right]\end{aligned}$$

by definition of the independent copy $\mathcal{D}'$. Then using that the supremum of the expectation is less than expectation of the supremum,

$$\begin{aligned}\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left(\mathbb{E}[h(Z)] - \frac{1}{n}\sum_{i=1}^n h(Z_i)\right)\right] &\le \mathbb{E}\left[\mathbb{E}\left(\sup_{h\in\mathcal{H}}\left(\frac{1}{n}\sum_{i=1}^n [h(Z_i') - h(Z_i)]\right) \mid \mathcal{D}\right)\right] \\ &= \mathbb{E}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{n}\sum_{i=1}^n [h(Z_i') - h(Z_i)]\right)\right] \\ &= \mathbb{E}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{n}\sum_{i=1}^n \sigma_i[h(Z_i') - h(Z_i)]\right)\right] \quad \text{(symmetrization)} \\ &\le \mathbb{E}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{n}\sum_{i=1}^n \sigma_i h(Z_i)\right)\right] + \mathbb{E}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{n}\sum_{i=1}^n -\sigma_i h(Z_i)\right)\right] \\ &= 2\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{n}\sum_{i=1}^n \sigma_i h(Z_i)\right)\right] = 2R_n(\mathcal{H})\end{aligned}$$

The reasoning is essentially identical for $\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left(\mathbb{E}[h(Z)] - \frac{1}{n}\sum_{i=1}^n h(Z_i)\right)\right] \le 2R_n(\mathcal{H})$.

□

Theorem 4.2 (Generalization bonud via Rademacher Complexity)**.** Suppose for all $h \in \mathcal{H}$, $0 \le h(Z) \le 1$. Then, with probability at least $1 - \delta$,

$$\begin{aligned}\mathbb{E}[h(Z)] &\le \frac{1}{n}\sum_{i=1}^n h(Z_i) + 2R_n(\mathcal{H}) + \sqrt{\frac{\log(w/\delta)}{2n}} \\ \mathbb{E}[h(Z)] &\le \frac{1}{n}\sum_{i=1}^n h(Z_i) + 2\hat{R}_S(\mathcal{H}) + 3\sqrt{\frac{\log(2/\delta)}{2n}}\end{aligned} \tag{4.8}$$

Proof. For conciseness, define

$$H(Z_1, \ldots, Z_n) := \sup_{h\in\mathcal{H}}\left[\frac{1}{n}\sum_{i=1}^n h(Z_i) - \mathbb{E}[h(Z_i)]\right]$$

and we prove the theorem for four steps.

- Step 1. We bound H using McDiarmid's inequality. To use McDiarmid's inequality, we firstly check

that the bounded difference condition holds:

$$\begin{aligned}
H(Z_1,\dots,Z_i,\dots,Z_n) - H(Z_1,\dots,Z_i',\dots,Z_n) &\le \sup_{h\in\mathcal{H}}\left\{\frac{1}{n}\sum_{j=1}^{n} h(Z_j)\right\} - \sup_{h\in\mathcal{H}}\left\{\frac{1}{n}\sum_{j\neq i}^{n} h(Z_j) + \frac{1}{n}h(Z_i')\right\} \\
&\le \sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{j=1}^{n} h(Z_j) - \frac{1}{n}\sum_{j\neq i}^{n} h(Z_j) - \frac{h(Z_i')}{n}\right\} \\
&= \sup_{h\in\mathcal{H}}\left\{\frac{1}{n}(h(Z_i) - H(Z_i'))\right\} \\
&\le \frac{1}{n} \qquad (\text{given } h(Z)\le 1)
\end{aligned}$$

where the second inequality holds beacue in general, $\sup_h A(f) - \sup_f B(f) \le \sup_f[A(f) - B(f)]$. We can thus apply McDiarmid's inequality with parameters $c_1 = \cdots = c_n = 1/n$,

$$\mathbb{P}\Big(H(Z_1,\dots,Z_n) - \mathbb{E}[H(Z_1,\dots,Z_n)] \ge t\Big) \le \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right) = \exp\left(-2nt^2\right)$$

that is, with probability $1-\delta$

$$H(Z_1,\dots,Z_n) \le \mathbb{E}[H(Z_1,\dots,Z_n)] + \sqrt{\frac{\log(2/\delta)}{2n}}$$

where here we set $\exp(-2nt^2) = \delta/2$.

- Step 2. We apply the symmetrization theorem **??** to get the upper bound of the expectation of the empirical process

$$\mathbb{E}[H(Z_1,\dots,Z_n)] \le 2R_n(\mathcal{H})$$

- Step 3. Bound expected Rademacher complexity through empirical Rademacher complexity and McDiarmid inequality. To begin with, define

$$\tilde{H} = \hat{R}_S(\mathcal{H}) := \mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i h(Z_i)\right]$$

Using a similar argument in Step 1, we find that $\tilde{H}$ also satisfies the bounded difference condition:

$$\begin{aligned}
\tilde{H}(Z_1,\dots,Z_i,\dots,Z_n) - \tilde{H}(Z_1,\dots,Z_i',\dots,Z_n) &\le \mathbb{E}\left[\sup_{h\in\mathcal{H}}\left\{\frac{1}{n}\sum_{j=1}^{n} h(Z_j)\right\} - \sup_{h\in\mathcal{H}}\left\{\frac{1}{n}\sum_{j\neq i}^{n} h(Z_j) + \frac{1}{n}h(Z_i')\right\}\right] \\
&\le \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{j=1}^{n} h(Z_j) - \frac{1}{n}\sum_{j\neq i}^{n} h(Z_j) - \frac{h(Z_i')}{n}\right\}\right] \\
&= \mathbb{E}\left[\sup_{h\in\mathcal{H}}\left\{\frac{1}{n}(h(Z_i) - h(Z_i'))\right\}\right] \\
&\le \frac{1}{n}
\end{aligned}$$

because the term inside the sup is always upper-bounded by 1. We can therefore apply McDiarmid's

inequality again with parameter $c_1 = \cdots c_n = 1/n$ and get

$$\mathbb{P}\Big(\tilde{H}(Z_1,\ldots,Z_n) - \mathbb{E}[\tilde{H}(Z_1,\ldots,Z_n)] \geq t\Big) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right) = \exp\left(-2nt^2\right)$$

and

$$\mathbb{P}\Big(\tilde{H}(Z_1,\ldots,Z_n) - \mathbb{E}[\tilde{H}(Z_1,\ldots,Z_n)] \leq -t\Big) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right) = \exp\left(-2nt^2\right)$$

that is, with probability $1-\delta$

$$\mathbb{E}[\tilde{H}(Z_1,\ldots,Z_n)] \leq \tilde{H}(Z_1,\ldots,Z_n) + \sqrt{\frac{\log(2/\delta)}{2n}}$$

where here we set $\exp(-2nt^2) = \delta/2$.

- Step 4. Putting all things together by noticing that

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

we have with probability $1-\delta$,

$$\begin{aligned}
\sup_{h\in\mathcal{H}}\left[\frac{1}{n}\sum_{i=1}^n h(Z_i) - \mathbb{E}[h(Z)]\right] = H(Z_1,\ldots,Z_n) &\leq \mathbb{E}[H(Z_1,\ldots,Z_n)] + \sqrt{\frac{\log(2/\delta)}{2n}} && \text{(Step 1.)}\\
&\leq 2R_n(\mathcal{H}) + \sqrt{\frac{\log(2/\delta)}{2n}} && \text{(Step 2.)}\\
&\leq 2\left(\hat{R}_S(\mathcal{H}) + \sqrt{\frac{\log(2/\delta)}{2n}}\right) + \sqrt{\frac{\log(2/\delta)}{2n}} && \text{(Step 3.)}\\
&= 2\hat{R}_S(\mathcal{H}) + 3\sqrt{\frac{\log(2/\delta)}{2n}}
\end{aligned}$$

Therefore, we have shown that

$$\begin{aligned}
\mathbb{E}[h(Z)] &\leq \frac{1}{n}\sum_{i=1}^n h(Z_i) + 2R_n(\mathcal{H}) + \sqrt{\frac{\log(1/\delta)}{2n}}\\
\mathbb{E}[h(Z)] &\leq \frac{1}{n}\sum_{i=1}^n h(Z_i) + 2\hat{R}_S(\mathcal{H}) + 3\sqrt{\frac{\log(2/\delta)}{2n}}
\end{aligned}$$

□

A useful fact is that both empirical Rademacher complexity and expected Rademacher complexity are translation invariant.

Proposition 4.1. Let $\mathcal{H}$ be a family of functions mapping $\mathcal{Z} \to \mathbb{R}$ and define $\mathcal{H}' = \{h'(Z) = h(Z) + c_0 \mid h \in \mathcal{H}\}$ for some $c_0 \in \mathbb{R}$. Then we have

$$\hat{R}_S(\mathcal{H}) = \hat{R}_S(\mathcal{H}') \quad \text{and} \quad R_n(\mathcal{H}) = R_n(\mathcal{H}')$$

Hint. The property of Rademacher random variables. □

4.3 Lipschitz-continuous Losses

A particularly appealing property in our context is the following property, sometimes called the "contraction principle".

Remark 4.1. For a compact intevral, continuously differentiable $\subseteq$ Lipschitz continuous $\subseteq$ absolutely continuous $\subseteq$ bounded variation $\subseteq$ differentiable almost everywhere

Proposition 4.2 (Contraction Principle - Lipschitz-continuous Functions)**.** Given any functions b, $a_i : \Theta \to \mathbb{R}$ (no assumption on Θ) and $\varphi_i : \mathbb{R} \to \mathbb{R}$ any 1-Lipschitz-functions, for $i = 1, \dots, n$, we have, for $\sigma \in \mathbb{R}^n$ a vector of independent Rademacher random variables,

$$\mathbb{E}_\sigma \left[\sup_{\theta \in \Theta} \; b(\theta) + \sum_{i=1}^n \sigma_i \varphi_i(a_i(\theta)) \right] \leq \mathbb{E}_\sigma \left[\sup_{\theta \in \Theta} \; b(\theta) + \sum_{i=1}^n \sigma_i a_i(\theta) \right] \tag{4.9}$$

Proof. We consider a proof by induction on n. The case $n = 0$ is trivial, and we show how to go from $n \geq 0$ to $n + 1$. We thus consider $\mathbb{E}_{\sigma_1,\dots,\sigma_{n+1}} \left[\sup_{\theta \in \Theta} b(\theta) + \sum_{i=1}^{n+1} \sigma_i \varphi_i(a_i(\theta)) \right]$ and compute the expectation with respect to σ_{n+1} explicitly, by considering the two potential values with probability $1/2$,

$$\begin{aligned}
&\mathbb{E}_{\sigma_1,\dots,\sigma_{n+1}} \left[\sup_{\theta \in \Theta} b(\theta) + \sum_{i=1}^{n+1} \sigma_i \varphi_i(a_i(\theta)) \right] \\
=&\frac{1}{2} \mathbb{E}_{\sigma_1,\dots,\sigma_n} \left[\sup_{\theta \in \Theta} b(\theta) + \sum_{i=1}^{n} \sigma_i \varphi_i(a_i(\theta)) + \varphi_{n+1}(a_{n+1}(\theta)) \right] + \frac{1}{2} \mathbb{E}_{\sigma_1,\dots,\sigma_n} \left[\sup_{\theta \in \Theta} b(\theta) + \sum_{i=1}^{n} \sigma_i \varphi_i(a_i(\theta)) - \varphi_{n+1}(a_{n+1}(\theta)) \right] \\
=&\mathbb{E}_{\sigma_1,\dots,\sigma_n} \left[\sup_{\theta, \theta' \in \Theta} \frac{b(\theta) + b(\theta')}{2} + \sum_{i=1}^{n} \sigma_i \frac{\varphi_i(a_i(\theta)) + \varphi_i(a_i(\theta'))}{2} + \frac{\varphi_{n+1}(a_{n+1}(\theta)) - \varphi_{n+1}(a_{n+1}(\theta'))}{2} \right]
\end{aligned}$$

By taking the supremum over (θ, θ') and (θ', θ) and using Lipschitz-continuity, we get

$$\begin{aligned}
&\mathbb{E}_{\sigma_1,\dots,\sigma_n} \left[\sup_{\theta, \theta' \in \Theta} \frac{b(\theta) + b(\theta')}{2} + \sum_{i=1}^{n} \sigma_i \frac{\varphi_i(a_i(\theta)) + \varphi_i(a_i(\theta'))}{2} + \frac{\varphi_{n+1}(a_{n+1}(\theta)) - \varphi_{n+1}(a_{n+1}(\theta'))}{2} \right] \\
=&\mathbb{E}_{\sigma_1,\dots,\sigma_n} \left[\sup_{\theta, \theta' \in \Theta} \frac{b(\theta) + b(\theta')}{2} + \sum_{i=1}^{n} \sigma_i \frac{\varphi_i(a_i(\theta)) + \varphi_i(a_i(\theta'))}{2} + \frac{\left|\varphi_{n+1}(a_{n+1}(\theta)) - \varphi_{n+1}(a_{n+1}(\theta'))\right|}{2} \right] \\
\leq&\mathbb{E}_{\sigma_1,\dots,\sigma_n} \left[\sup_{\theta, \theta' \in \Theta} \frac{b(\theta) + b(\theta')}{2} + \sum_{i=1}^{n} \sigma_i \frac{\varphi_i(a_i(\theta)) + \varphi_i(a_i(\theta'))}{2} + \frac{\left|a_{n+1}(\theta) - a_{n+1}(\theta')\right|}{2} \right] \\
=&\mathbb{E}_{\sigma_1,\dots,\sigma_n} \left[\sup_{\theta, \theta' \in \Theta} \frac{b(\theta) + b(\theta')}{2} + \sum_{i=1}^{n} \sigma_i \frac{\varphi_i(a_i(\theta)) + \varphi_i(a_i(\theta'))}{2} + \frac{a_{n+1}(\theta) - a_{n+1}(\theta')}{2} \right]
\end{aligned}$$

The first and last equalities hold becasue of the fact that $\sup_{\theta, \theta' \in \Theta} a_{n+1}(\theta) - a_{n+1}(\theta')$ at least equal to zero. Now, we can redo the exact same sequence of equalities with φ_{n+1} being the identity, to obtain that the last expression above is equal to

$$\begin{aligned}
&\mathbb{E}_{\sigma_1,\dots,\sigma_n} \mathbb{E}_{\sigma_{n+1}} \left[\sup_{\theta \in \Theta} b(\theta) + \sum_{i=1}^{n} \sigma_i \varphi_i(a_i(\theta)) + \sigma_{n+1} a_{n+1}(\theta) \right] \\
\leq&\mathbb{E}_{\sigma_1,\dots,\sigma_n} \mathbb{E}_{\sigma_{n+1}} \left[\sup_{\theta \in \Theta} b(\theta) + \sum_{i=1}^{n} \sigma_i a_i(\theta) + \sigma_{n+1} a_{n+1}(\theta) \right] \quad \text{by the induction hypothesis}
\end{aligned}$$

which leads to the desired result.

□

We can apply the contraction principle above to supervised learning situations where $u_i \mapsto l(y_i, u_i)$ is G-Lipschitz-continuous for all i almost surely (possible for regression or when using a convex surrogate for binary classification), leading to

$$\mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \sigma_i l(Y_i, f(X_i)) \mid \mathcal{D} \right] \leq G \cdot \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \sigma_i f(X_i) \mid \mathcal{D} \right] \tag{4.10}$$

by the contraction principle, which leads to

$$\frac{1}{2} \mathbb{E} \left[\sup_{h \in \mathcal{H}} \left(\frac{1}{n} \sum_{i=1}^n \sum_{i=1}^n h(Z_i) - \mathbb{E}[h(Z)] \right) \right] \leq R_n(\mathcal{H}) \leq G \cdot R_n(\mathcal{F}) \tag{4.11}$$

Thus, the Rademacher complexity of the class of prediction functions $R_n(\mathcal{F})$ controls the expectation of empirical process.

4.4 Ball-constrained Linear Predictions

We now assume that $\mathcal{F} = \{f_\theta(X) = \theta^\top \varphi(X) \mid \Omega(\theta) \leq D\}$ where Ω is norm on $\mathbb{R}^d$. We denote by $\Phi \in \mathbb{R}^{n \times d}$ the design matrix, that is,

$$\Phi = \begin{bmatrix} \varphi_1(X_1) & \varphi_2(X_1) & \cdots & \varphi_d(X_1) \\ \varphi_1(X_2) & \varphi_2(X_2) & \cdots & \varphi_d(X_2) \\ \vdots & \vdots & & \vdots \\ \varphi_1(X_n) & \varphi_2(X_n) & \cdots & \varphi_d(X_n) \end{bmatrix}$$

We have

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E} \left[\sup_{\Omega(\theta) \leq D} \left(\frac{1}{n} \sum_{i=1}^n \sigma_i \theta^\top \varphi(X_i) \right) \right] = \mathbb{E} \left[\sup_{\Omega(\theta) \leq D} \left(\frac{1}{n} \sigma^\top \Phi \theta \right) \right] = \frac{D}{n} \mathbb{E} \Big[\Omega^*(\Phi^\top \sigma) \Big] \tag{4.12}$$

where $\Omega^*(u) = \sup\{u^\top \theta \mid \Omega(\theta) \leq 1\}$ is the dual norm of Ω.

- when Ω is the l_p-norm with $p \in [1, \infty]$, then Ω^* is the l_q-norm, with conjugate realtion $\frac{1}{p} + \frac{1}{q} = 1$
- $||\cdot||_1^* = ||\cdot||_\infty$ and $||\cdot||_\infty^* = ||\cdot||_1$ and $||\cdot||_2^* = ||\cdot||_2$.

Thus, computing Rademacher complexities is equivalent to computing expectation of norms. When $\Omega = ||\cdot||_2$, we get:

$$\begin{aligned} \mathcal{R}_n(\mathcal{F}) &= \frac{D}{n} \mathbb{E} \big[||\Phi^\top \sigma||_2 \big] \\ &\leq \frac{D}{n} \sqrt{\mathbb{E}[||\Phi^\top \sigma||_2^2]} \quad \text{(Jensens' inequality apply on } f(x) = x^2) \\ &= \frac{D}{n} \sqrt{\mathbb{E}[\mathrm{tr}(\Phi^\top \sigma \sigma^\top \Phi)]} \quad \text{(holds for any vector)} \\ &= \frac{D}{n} \sqrt{\mathbb{E}[\mathrm{tr}(\Phi^\top \Phi)]} \quad \text{(using that IID such that } \mathbb{E}[\sigma \sigma^\top] = I) \\ &= \frac{D}{n} \sqrt{\sum_{i=1}^n \mathbb{E}[\varphi(X_i)^\top \varphi(X_i)]} = \frac{D}{n} \sqrt{\sum_{i=1}^n \mathbb{E}||\varphi(X_i)||_2^2} = \frac{D}{\sqrt{n}} \sqrt{\mathbb{E}||\varphi(X)||_2^2} \end{aligned} \tag{4.13}$$

We thus obtain a *dimension-independent* Rademacher complexity that we can use in the summary below.

Example 4.1. Upper-bound the Rademacher complexity for $\Omega = ||\cdot||_1$.

4.5 Putting Things Together (Linear Predictions)

With all the elements above (section 4.3 and 4.4), we can now propose the following general result (where no convexity of the loss function is assumed).

Proposition 4.3 (Estimation Error)**.** Assume a G-Lipschitz-continuous loss function, linear prediction functions with $\mathcal{F} = \{f_\theta(X) = \theta^\top \varphi(X) \mid ||\theta||_2 \leq D\}$, where $\mathbb{E}||\varphi(X)||_2^2 \leq R^2$. Let $\hat{f} = f_{\hat{\theta}} \in \mathcal{F}$ be the minimizer of the empirical risk, then:

$$\mathbb{E}\big[\mathcal{R}(f_{\hat{\theta}})\big] - \inf_{||\theta||_2 \leq D} \mathcal{R}(f_\theta) \leq \frac{2GRD}{\sqrt{n}}$$

It is essential to know that $f_{\hat{\theta}}$ here is a random variable.

Proof. Using Proposition **??**, equation 4.11 and 4.13, we get the desire result. □

Proposition 4.4 (Approximation Error)**.** If we assume that there exists a minimizer θ_* of $\mathcal{R}(f_\theta)$ over $\mathbb{R}^d$, the approximation error is upper-bounded by

$$\begin{aligned}
\inf_{||\theta||_2 \leq D} \mathcal{R}(f_\theta) - \mathcal{R}(f_{\theta_*}) &\leq G \inf_{||\theta||_2 \leq D} \mathbb{E}[|f_\theta(X) - f_{\theta_*}(X)] \quad (G\text{-Lipschitz-continuous loss function}) \\
&= G \inf_{||\theta||_2 \leq D} \mathbb{E}[|\varphi(X)^\top(\theta - \theta_*)] \\
&\leq G \inf_{||\theta||_2 \leq D} ||\theta - \theta_*||_2 \cdot \mathbb{E}[||\varphi(X)||_2^2] \leq GR \inf_{||\theta||_2 \leq D} ||\theta - \theta_*||_2
\end{aligned}$$

This leads to empirical risk minimization error is uppber-bounded by

$$\mathbb{E}[\mathcal{R}(f_{\hat{\theta}})] - \mathcal{R}(f_{\theta_*}) \leq GR \inf_{||\theta||_2 \leq D} ||\theta - \theta_*||_2 + \frac{2GRD}{\sqrt{n}} = GR(||\theta_*||_2 - D)^+ + \frac{2GRD}{\sqrt{n}} \tag{4.14}$$

We see that for $D = ||\theta_*||_2$, we obtain the bound $\frac{2GR||\theta_*||_2}{\sqrt{n}}$, but this setting requires to know $||\theta_*||_2$ which is not possible in practice.

- if D is too large, the estimation error gets larger, leading to overfitting;
- while if D is too small, the approximation error can quickly kick in (with a value that does not go to zero when n tends to infinity), leading to underfitting.

4.6 From Constrained to Regularized Estimation

In practice, it is preferable to penalize by the norm $\Omega(\theta) = ||\theta||_2$ instead of constraining (the main reasons being that the hyperparameter is easier to find and the optimization is easier).

For simplicity, we only consider the l_2-norm here. We now denote $\hat{\theta}_\lambda$ the minimizer of

$$\hat{\mathcal{R}}(f_\theta) + \frac{\lambda}{2}||\theta||_2^2 \tag{4.15}$$

If the loss is always positive, then

$$\frac{\lambda}{2}||\hat{\theta}_\lambda||_2^2 \leq \hat{\mathcal{R}}(f_{\hat{\theta}_\lambda}) + \frac{\lambda}{2}||\hat{\theta}_\lambda||_2^2 \leq \hat{\mathcal{R}}(f_0)$$

leading to a bound $||\hat{\theta}_\lambda||_2 = \mathcal{O}(1/\sqrt{\lambda})$. Thus, with $D = \mathcal{O}(1/\sqrt{\lambda})$ in the bound above, this lead to a deviation of $\mathcal{O}(1/\sqrt{\lambda n})$, which is not optimal.

We now cite Sridharan, Shalev-Shwartz, and Srebro 2008 without proof an interesting stronger result using the strong convexity of the squared l_2-norm.

Proposition 4.5 (Fast Rates for Regularized Objectives)**.** Assume a G-Lipschitz-continuous **convex** loss function, linear prediction functions with $\mathcal{F} = \{f_\theta(X) = \theta^\top \varphi(X) \mid ||\theta||_2 \leq D\}$, where $\mathbb{E}||\varphi(X)||_2^2 \leq R^2$. Let $\hat{\theta}_\lambda \in \mathbb{R}^d$ be the minimizer of the regularized empirical risk in equation 4.15, then

$$\mathbb{E}[\mathcal{R}(f_{\hat{\theta}_\lambda})] \leq \inf_{\theta \in \mathbb{R}^d} \left\{ \mathcal{R}(f_\theta) + \frac{\lambda}{2}||\theta||_2^2 \right\} + \frac{32G^2R^2}{\lambda n}$$

Note that we obtain a "fast rate" in $\mathcal{O}(R^2/(\lambda n))$, which has a better dependence in n, but depends on λ, which can be very small in practice. One classical choice of λ if $\lambda \propto GR/(\sqrt{n}||\theta_*||)$, leading to the slow rate

$$\mathbb{E}[\mathcal{R}(f_{\theta_*})] \leq \mathcal{R}(f_{\theta_*}) + \mathcal{O}\left(\frac{GR}{\sqrt{n}}||\theta_*||_2\right)$$

5 Growth Function and VC-Dimension

5.1 Growth Function

Here we will show how the Rademacher complexity can be bounded in terms of the growth function in binary classification problem. To begin with, recall that the empirical Rademacher complexity with respect to sample S with size n is defined as

$$\hat{R}_S(\mathcal{H}) = \mathbb{E}_\sigma \left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \sigma_i h(X_i) \right]$$

Definition 5.1 (Growth Function)**.** The growth function $\Pi_\mathcal{H} : \mathbb{N} \to \mathbb{N}$ for a hypothesis set $\mathcal{H}$ is defined by

$$\forall\ n \in \mathbb{N}_+,\ \Pi_\mathcal{H}(n) = \max \left| \left\{ (h(X_1), \ldots, h(X_n)) \mid h \in \mathcal{H},\ X_1, \ldots, X_n \in \mathcal{X} \right\} \right| \tag{5.1}$$

where $|\cdot|$ compute the cardinality of a set.

Thus, $\Pi_\mathcal{H}(n)$ is the maximum number of distinct ways in which n points can be classified using hypotheses in $\mathcal{H}$. This provides another measure of the richness of the hypothesis set $\mathcal{H}$. However, unlike the Rademacher complexity, this measure does not depend on the distribution, it is purely **combinatorial**. To relate the Rademacher complexity to the growth function, we will use Massart's lemma.

Theorem 5.1 (Massart's Lemma)**.** Let $A \subset \mathbb{R}^n$ be a finite set, with $r = \sup_{X \in A} ||X||_2$, then the following holds:

$$\mathbb{E}_\sigma \left[\sup_{X \in \mathcal{A}} \frac{1}{n} \sum_{i=1}^n \sigma_i X_i \right] \leq \frac{r\sqrt{2 \log |A|}}{n} \tag{5.2}$$

where σ_i's are independent Rademacher variables taking values in $\{-1, +1\}$ and $X_1, \ldots, X_n$ are the components of vector X.

Proof. For any $t > 0$, using Jensen's inequality, rearranging terms, and bounding the supremum by a sum, we obtain:

$$\begin{aligned}\exp\left(t \cdot \mathbb{E}_\sigma\left[\sup_{X\in A}\sum_{i=1}^n \sigma_i X_i\right]\right) &\leq \mathbb{E}_\sigma\left[\exp\left(t\sup_{X\in A}\sum_{i=1}^n \sigma_i X_i\right)\right] \\ &= \mathbb{E}_\sigma\left[\sup_{X\in A}\exp\left(t\sum_{i=1}^n \sigma_i X_i\right)\right] \\ &\leq \sum_{X\in A}\mathbb{E}_\sigma\left[\exp\left(t\sum_{i=1}^n \sigma_i X_i\right)\right]\end{aligned}$$

We next use the independence of the σ_i's, then apply the bound in Eq.(C.14), and the definition of r to write:

$$\begin{aligned}\exp\left(t \cdot \mathbb{E}_\sigma\left[\sup_{X\in A}\sum_{i=1}^n \sigma_i X_i\right]\right) &\leq \sum_{X\in A}\prod_{i=1}^n \mathbb{E}[\exp(t\sigma_i X_i)] \\ &\leq \sum_{X\in A}\prod_{i=1}^n \exp\left(\frac{t^2(2X_i)^2}{8}\right) \\ &= \sum_{X\in A}\exp\left(\frac{t^2}{2}\sum_{i=1}^n X_i^2\right) \\ &\leq \sum_{X\in A}\exp\left(\frac{t^2 r^2}{2}\right) = |A|\cdot\exp\left(\frac{t^2 r^2}{2}\right)\end{aligned}$$

The last inequality holds by applying he definition of r, which is

$$r := \sup_{X\in\mathcal{A}}||X||_2 = \sup_{X\in\mathcal{A}}\sqrt{\sum_{i=1}^n X_i^2}$$

Taking the logarithm on both sides and dividing by t yields:

$$\mathbb{E}_\sigma\left[\sup_{X\in A}\sum_{i=1}^n \sigma_i X_i\right] \leq \frac{\log|A|}{t} + \frac{tr^2}{2} \tag{5.3}$$

Since such inequality holds for every t, we can minimize over t and get $t = \frac{\sqrt{2\log|A|}}{r}$ and get

$$\mathbb{E}_\sigma\left[\sup_{X\in A}\sum_{i=1}^n \sigma_i X_i\right] \leq r\sqrt{2\log|A|}$$

Dividing both sides by n leads to the desired result.

□

Using this result, we can bound the Rademacher complexity $R_n(\mathcal{H})$ in terms of the growth function $\pi_\mathcal{H}$.

Corollary 5.2. Let $\mathcal{H}$ be a family of functions taking values in $\{-1, +1\}$. Then the following holds:

$$R_n(\mathcal{H}) \leq \sqrt{\frac{2\log\Pi_\mathcal{H}(n)}{n}} \tag{5.4}$$

Proof. For a fixed sample $S_n = (X_1, \ldots, X_n) \sim P$, we denote by H_{S_n} the set of vectors of function values $(h(X_1), \ldots, h(X_n))^\top$ where h is in $\mathcal{H}$. Singce $h \in \mathcal{H}$ take values in $\{-1, +1\}$, the norm of these vectors is

bounded by $\sqrt{n}$. We can then apply Massart's lemma as follows:

$$R_n(\mathcal{H}) = \mathbb{E}_P\left[\mathbb{E}_\sigma\left[\sup_{u\in H_{S_n}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i u_i\right]\right] \leq \mathbb{E}_P\left[\frac{\sqrt{n}\sqrt{2\log|H_{S_n}|}}{n}\right]$$

By definition, $|H_{S_n}|$ is bounded by the growth function $\Pi_{\mathcal{H}}(n)$, thus,

$$R_n(\mathcal{H}) \leq \mathbb{E}_P\left[\frac{\sqrt{n}\sqrt{2\log\Pi_{\mathcal{H}}(n)}}{n}\right] = \sqrt{\frac{2\log\Pi_{\mathcal{H}}(n)}{n}}$$

which concludes the proof.

□

Combining the generalization bound via Rademacher complexity in Thm. 4.2 with the corollary above yields immediately the following generalization bound in terms of the growth function.

Theorem 5.3 (Generalization Bound via Growth Function)**.** Let $\mathcal{H}$ be a family of functions taking values in $\{-1,+1\}$. Then, for any $\delta > 0$, with probability at least $1-\delta$, we have

$$\mathcal{R}(h) \leq \hat{\mathcal{R}}(h) + 2\sqrt{\frac{2\log\Pi_{\mathcal{H}}(n)}{n}} + \sqrt{\frac{\log(2/\delta)}{2n}} \tag{5.5}$$

$h \in \mathcal{H}$. Again, recall that $\mathcal{R}(h) = \mathbb{E}[h(Z)]$ and $\hat{\mathcal{R}}(h) = \frac{1}{n}\sum_{i=1}^{n} h(Z_i)$.

Growth function bounds can also be derived directly (withougt using Rademacher complexity bounds first). The resulting bound is

$$\mathbb{P}\left(\left|\mathcal{R}(h) - \hat{\mathcal{R}}(h)\right| > \varepsilon\right) \leq 4\Pi_{\mathcal{H}}(2n)\exp\left(-\frac{n\varepsilon}{8}\right) \tag{5.6}$$

The computation of the growth function may not be always convenient since, by definition, it requires computing $\Pi_{\mathcal{H}}(n)$ for all $n \geq 1$. The next section introduces an alternative measure of the complexity of a hypothesis set H that is based instead on a single scalar, which will turn out to be in fact deeply related to the behavior of the growth function.

5.2 VC-dimension

Here, we introduce the notion of VC-dimension (Vapnik-Chervonenkis dimension). The VC-dimension is also a purely combinatorial notion but it is often easier to compute than the growth function (or the Rademacher Complexity). As we shall see, the VC-dimension is a key quantity in learning and is directly related to the growth function.

To define the VC-dimension of a hypothesis class $\mathcal{H}$, we first introduce the concepts of *dichotomy* and that of *shattering*. Given a hypothesis set $\mathcal{H}$, a dichotomy of a set S is one of the possible ways of labeling the points of S using a hypothesis in $\mathcal{H}$. A set S of $n \geq 1$ is said to be shattered by a hypothesis class $\mathcal{H}$ when $\mathcal{H}$ realizes all possible dichotomies of S, that is when $\Pi_{\mathcal{H}}(n) = 2^n$.

Definition 5.2 (VC-dimension)**.** The VC-dimension of a hypothesis class $\mathcal{H}$ is the size of the largest sample set that can be fully shattered by $\mathcal{H}$:

$$\mathrm{VCdim}(H) = \max\{n : \Pi_{\mathcal{H}}(n) = 2^n\} \tag{5.7}$$

Remark 5.1. Note that, by definition, if $\text{VCdim}(\mathcal{H}) = d$, there exists a sample set of size d that can be fully shattered. But, this does not imply that all sets of size d or less are fully shattered, in fact, this is typically not the case.

In general, to compute the VC-dimension, we will typically show a lower bound for its value and then a matching upperbouond. To given a lower bound d for $\text{VCdim}(\mathcal{H})$, it suffices toshow that a set S of cardinality d can be shattered by $\mathcal{H}$. To give an upper bound, we need to prove that no set S of cardinality $d+1$ can be shattered by $\mathcal{H}$, which is typically more difficult. The followings are some examples of classifers and their VC dimension.

- **Interval Classifier on real line**. Consider a hypothesis class $\mathcal{H} = \{h_{[a,b]} \mid \forall\ a < b\}$ where

$$h_{[a,b]}(X) = \begin{cases} +1, & x \in [a,b] \\ -1, & x \notin [a,b] \end{cases}$$

  It can be shown that $\text{VCdim}(\mathcal{H}) = 2$ in this case. Suppose there is a sample set S with two random variables $(X_1, X_2) = (x_1, x_2)$. Without loss of generality, we suppose $x_1 < x_2$. Then, all four possible dichotomies $(+1,+1)$, $(-1,+1)$, $(+1,-1)$ and $(-1,-1)$ can be realized by some classifer $h_{[a,b]}$. In contrast, by the definition of intervals, no set of three point can be shattered since the case $(+1,-1,+1)$ labeling cannot be realized by any $h_{[a,b]}$. Hence,

$$\text{VCdim(intervals in } \mathbb{R}) = 2$$

- **Hyperplane Classifier**.

$$\text{VCdim(hyperpalnes in } \mathbb{R}^d) = d+1$$

  To begin with, we derive a lower bound by starting with a set of $d+1$ points in $\mathbb{R}^d$, setting X_0 to be the origin and defining X_i, for $i \in \{1,\ldots,d\}$, as the point whose ith coordinate is 1 and other points are 0, that is,

$$X_{ii} = 1 \quad \text{and} \quad X_{ij} = 0\ \forall j \neq i$$

  for all $d+1$ points. Let $Y_0, Y_1, \ldots, Y_d \in \{-1,+1\}$ be an arbitrary set of labels for $X_0, X_1, \ldots, X_d$. Let w be the vector whose ith coordinate is Y_i. Then the classifier defined by the hyperplane $\{x \mid w^\top X + Y_0/2 = 0\}$ shatters $X_0, X_1, \ldots, X_d$ since for any $i \in \{0,1,\ldots,d\}$,

$$\text{sgn}\left(w^\top X_i + \frac{Y_0}{2}\right) = \text{sgn}\left(Y_i + \frac{Y_0}{2}\right) = Y_i \tag{5.8}$$

  To obtain an upper bound, it suffices to show that no set of $d+2$ points can be shattered by halfspaces. Concretely, let S be a set of $d+2$ points. By Radon's Theorem 5.4, it can be partitioned into two sets X_1 and X_2 such that their convex hulls intersect. Observe that when two sets of points S_1 and S_2 are separaetd by a hyperplane, and each of their convex hulls are also separated by that hyperplane. Thus, S_1 and S_2 cannot be separated by a hyperplane and S is not shattered.

  Combining our lower and upper bounds, we have proven that VC-dimension is $d+1$ in this case.

- **Axis-ligned Rectangles**.

$$\text{VCdim}(\mathcal{H}) = 4$$

  We first show that the VC-dimension is at least four, by considering four points in a diamond pattern. Then it is clear that all 16 dichotomies can be realized.

- **Convex Polygons**.
- **Sine Functions**.

  The previous examples could suggest that the VC-dimension of $\mathcal{H}$ coincides with the number of free parameters defining $\mathcal{H}$. For example, the number of parameters defining hyperplanes matches their VC-dimension. However, this does not hold in general.

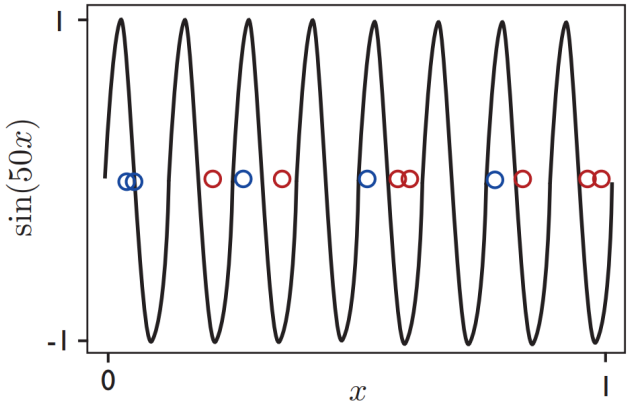


Figure 7: An example of a sine function used for classification

Here is a striking example. Consider the following of sine functions:

$$\mathcal{H} = \{t \mapsto \sin(\omega t) \mid \omega \in \mathbb{R}\}$$

These singe function can be used to classifiy the points on the real line: a point is labeled positively if it is above the curve, negatively otherwise. Altough this family of sine function is defined via a single parameter, ω, it can be shown that VCdim(sine functions) $= +\infty$.

Theorem 5.4 (Radon's theorem)**.** Any set S of $d+2$ points in $\mathbb{R}^d$ can be partitioned into two subsets S_1 and S_2 such that the convex hulls of S_1 and S_2 intersect.

Proof. Let $S = \{X_1, \ldots, X_{d+2}\} \subset \mathbb{R}^d$. The following is a system of $d+1$ linear equations in $\alpha_1, \ldots, \alpha_{d+2}$,

$$\sum_{i=1}^{d+2} \alpha_i X_i = 0 \quad \text{and} \quad \sum_{i=1}^{d+2} \alpha_i = 0$$

This is because the first equality leads to d equations, one for each component of point. The number of unknown, $d+2$, is large than the number of equations, $d+1$, and therefore the system admits a non-zero solution $\beta_1, \ldots, \beta_{d+2}$.

Since $\sum_{i=1}^{d+2} \beta_i = 0$, both $I_1 = \{i \in [1, d+2] \mid \beta_i > 0\}$ and $I_2 = \{i \in [1, d+2] \mid \beta_i < 0\}$ are non-empty sets. Then $S_1 = \{X_i \mid i \in I_1\}$ and $S_2 = \{X_i \mid i \in I_2\}$ form a partition of S. By the last equation above, we have

$$\sum_{i \in I_1} \beta_i = -\sum_{i \in I_2} \beta_i$$

Let $\beta = \sum_{i \in I_1} \beta_i$, then the first part $\sum_{i=1}^{d+2} \alpha_i X_i = 0$ implies that

$$\sum_{i \in I_1} \frac{\beta_i}{\beta} X_i = \sum_{i \in I_2} \frac{-\beta_i}{\beta} X_i$$

with $\sum_{i \in I_1} \beta_i/\beta = 1 = \sum_{i \in I_2} -\beta_i/\beta$, and $\beta_i/\beta \geq 0$ for $i \in I_1$ and $\frac{-\beta_i}{\beta} \geq 0$ for $i \in I_2$. By definition of the convex hulls, this implies that $\sum_{i \in I_1} \frac{\beta_i}{\beta} X_i$ belongs both to the convex hull of S_1 and to that of S_2.

□

5.3 Link Growth Function and VC-dimension

We have shown that the VC-dimension of many other hypothesis sets can be determined or uppder bounded in a similar way. In particular, the VC-dimension of any vector space of dimension $r < \infty$ can be shown to be at most r. The next result known as Sauer's lemma clarifies the connection between the notions of growth function and VC-dimension.

Theorem 5.5 (Sauer's lemma)**.** Let $\mathcal{H}$ be a hypothesis set with $\text{VCdim}(\mathcal{H}) = d$. Then, for all $m \in \mathbb{N}$, the following inequality holds:

$$\Pi_{\mathcal{H}}(n) \leq \sum_{i=0}^{d} \binom{n}{i} \tag{5.9}$$

Proof. The proof is by induction on $n + d$. The statement clearly holds for $n = 1$ and $d = 0$ or $d = 1$. Now, assume that it holds for $(n-1, d-1)$ and $(n-1, d)$. Fix a sample set $S = \{x_1, \ldots, x_n\}$ with $\Pi_{\mathcal{H}}(n)$ dichotomies and let $G = \mathcal{H}_S$ be the set of concepts $\mathcal{H}$ induces by restriction to S.

Now consider the following families over $S' = \{x_1, \ldots, x_{n-1}\}$. We define $G_1 = G_S$ as the set of concepts $\mathcal{H}$ includes by restriction to S'. Next, ...

□

Corollary 5.6. Let $\mathcal{H}$ be a hypothesis set with $\text{VCdim}(\mathcal{H}) = d$. Then for all $n \geq d$,

$$\Pi_{\mathcal{H}}(n) \leq \left(\frac{en}{d}\right)^d = \mathcal{O}(n^d) \tag{5.10}$$

Proof. The proof begins by using Sauer's lemma. The first inequality multiplies each summand by a factor that is greater than or equal to one since $n \geq d$, while the second inequality adds non-negative summands to the summation.

$$\begin{aligned}
\Pi_{\mathcal{H}}(n) &\leq \sum_{i=0}^{d} \binom{n}{i} \\
&\leq \sum_{i=0}^{d} \binom{n}{i} \left(\frac{n}{d}\right)^{d-i} \\
&\leq \sum_{i=0}^{n} \binom{n}{i} \left(\frac{n}{d}\right)^{d-i} \\
&= \left(\frac{n}{d}\right)^d \sum_{i=0}^{d} \binom{n}{i} \left(\frac{d}{n}\right)^i \\
&= \left(\frac{n}{d}\right)^d \left(1 + \frac{d}{n}\right)^n \leq \left(\frac{n}{d}\right)^d e^d
\end{aligned}$$

After simplifying the expression using the binomial theorem, the final inequality follows using the general identity $(1+x) \leq e^x$.

□

The explicit relationship just formulated between VC-dimension and the growth function combined with corollary above leads immediately to the following generalization bounds based on the VC-dimension.

Theorem 5.7 (Generalization Bounds via VC-dimension)**.** Let $\mathcal{H}$ be a family of functions taking values in $\{-1, +1\}$ with VC-dimension d. Then, for any $\delta > 0$, with probability at least $1 - \delta$, the following holds for all $h \in \mathcal{H}$:

$$\mathcal{R}(h) \leq \hat{\mathcal{R}}(h) + \sqrt{\frac{2d \log(en/d)}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}} \tag{5.11}$$

Thus, the form of this generalization bound is

$$\mathcal{R}(h) \leq \hat{\mathcal{R}}(h) + \mathcal{O}\left(\sqrt{\frac{\log(n/d)}{(n/d)}}\right) \tag{5.12}$$

which emphasizes the importance of the ratio m/d for generalization. The theorem provides another instance of Occam's razor principle where simplicity is measured in terms of smaller VC-dimension.

VC-dimension bounds can be derived directly without using an intermediate Rademacher complexity bound, as shown in (5.6). Combining Sauer's lemma with (5.6) leads to the following high-probability bound

$$\mathcal{R}(h) \leq \hat{\mathcal{R}}(h) + \sqrt{\frac{8d \log(2em/d) + 8 \log(4/\delta)}{n}} \tag{5.13}$$

which has the general form we derive above. The log factor plays only a minor role in these bounds. A finer analysis can be used in fact to eliminate that factor.

5.4 Lower Bounds

See the details in Mohri, Rostamizadeh, and Talwalkar 2018.

6 Covering Number and Chaining

In the case of (binary) classification, we established that only a finite number of elements in the hypothesis class $\mathcal{F}$ really matter, as far as establishing a notion of complexity of $\mathcal{F}$ that can be used to bound uniform deviations in expectation $(\hat{\mathcal{R}}(f) - \mathcal{R}(f))$: only the classifiers yielding different labelings matter. We did so using combinatorial arguments, leading to the notion of complexity given by the growth function, which measures the maximal size of $\mathcal{F}$ when restricted to a given number of points. This quantity, in turn, can be upper-bounded in terms of the VC dimension.

We will now apply the same idea in the setting of regression, where we consider real-valued predictors. We will isolate a few (finitely many) predictors of interest, bound the Rademacher complexity of the set of restrictions to samples in terms of the Rademacher complexity of these representative predictors, and control the error that we commit by only considering a subset of $\mathcal{F}$

Our goal is to find a finite set that explains "most of" the deviation in expectation, up to a certain precision parameter ε. To do so, we will use metric arguments and the notion of covering numbers. This

analysis, in fact, will yield improvements also in the setting of binary classification, allowing remove the term $\log(en/d)$ in Eq.(5.11).

6.1 Covering and Packing

We begin by defining the notions of packing and covering a set in a metric space. Recall that a metric space (T, ρ) consists of a non-empty set T, equipped with a mapping $\rho : T \times T \to \mathbb{R}$ that satisfies the following properties:

(a) It is non-negative: $\rho(\theta, \theta')$ for all paris (θ, θ'), with equality if and only if $\theta = \theta'$

(b) It is symmetric: $\rho(\theta, \theta') = \rho(\theta', \theta)$ for all pairs (θ', θ)

(c) The triangle inequality holds: $\rho(\theta, \theta') \leq \rho(\theta, \tilde{\theta}) + \rho(\theta', \tilde{\theta})$

Familiar examples of metric spaces include the real space $\mathbb{R}^d$ wiht the *Euclidean metric*

$$\rho(\theta, \theta') = ||\theta - \theta'||_2 := \sqrt{\sum_{j=1}^{d} (\theta_j - \theta'_j)^2}$$

and the discrete cube $\{0, 1\}^d$ with the *rescaled Hamming metric*

$$\rho_H(\theta, \theta') := \frac{1}{d} \sum_{j=1}^{d} \mathbb{1}\{\theta_j \neq \theta'_j\}$$

Also of interest are various metric spaces of functions, among them usual spaces $L^2(\mu, [0, 1])$ with its metric

$$||f - g||_2 := \left[\int_0^1 (f(x) - g(x))^2 d\mu(x) \right]^{1/2}$$

as well as the space $C[0, 1]$ of all continuous functions on $[0, 1]$ equipped with the sup-norm metric

$$||f - g||_\infty = \sup_{x \in [0,1]} |f(x) - g(x)|$$

Given a metric spaces (T, ρ), a natural way in which to measure its size is in terms of number of balls of a fixed radius ε required to cover it, a quantity known as the covering number.

Definition 6.1. A ε-cover of a set T with respect to a metric d is a set $\{\theta_1, \ldots, \theta_n\} \subset T$ such that for each $\theta \in T$, there exists some $i \in \{1, \ldots, n\}$ such that $\rho(\theta, \theta_i) \leq \varepsilon$. The ε-covering number $N(\varepsilon; T, \rho)$ is the cardinality of the smallest ε-cover.

It is easy to see that covering number is decreasing in ε, namely, $N(\varepsilon) \geq N(\varepsilon')$ for all $\varepsilon \leq \varepsilon'$. Typically, the covering number diverges as $\varepsilon \to 0^+$, and of interest to us is this growth rate on a logarithmic scale. More specifically, the quantity

$$\log N(\varepsilon; T, \rho)$$

is known as the *metric entroppy* of the set T with respect to ρ. Here are some examples which show how covering number can be bounded.

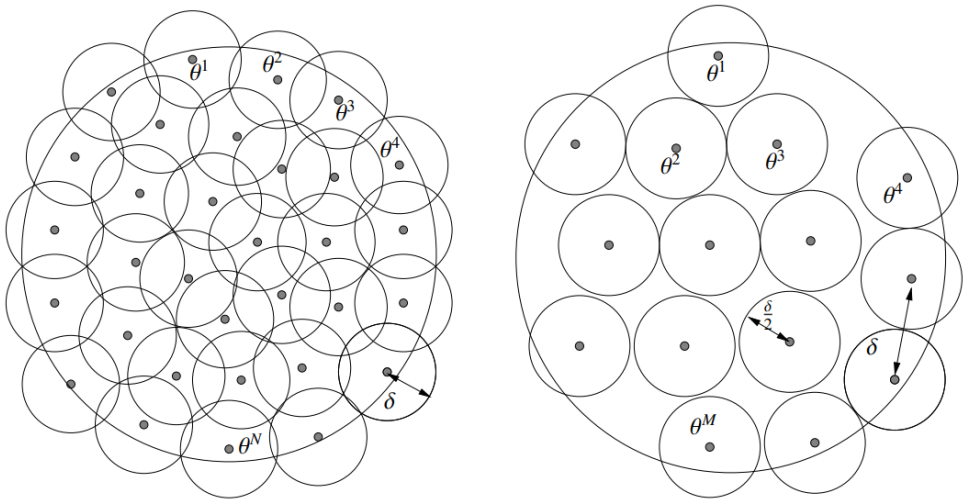


Figure 8: Covering and packing sets

- **Covering numbers of unit cubes**. Consider the interval $[-1, 1]$ in $\mathbb{R}$, equiped with the metric $\rho(\theta, \theta') = |\theta - \theta'|$. Suppose that we divide the interval $[-1, 1]$ into $L := \lfloor 1/\varepsilon \rfloor + 1$ sub intervals, centered at the points $\theta_i = -1 + 2(i-1)\varepsilon$ for $i \in [L] := \{1, 2, \ldots, L\}$, and each of length at most 2ε. By construction, for any point $\theta' \in [0, 1]$, there is some $j \in [L]$ such that $|\theta_j - \theta'| \leq \varepsilon$, which shows that

$$N(\varepsilon; [-1, 1], |\cdot|) \leq \frac{1}{\varepsilon} + 1$$

  We can easily generalize this analysis for the d-dimensional cube $[-1, 1]^d$, we have

$$N(\varepsilon; [-1, 1]^d, ||\cdot||_\infty) \leq \left(1 + \frac{1}{\varepsilon}\right)^d$$

- **Covering the binary hypercube**. Consider the binary hypercube $\mathcal{H} = \{0, 1\}^d$ equiped with the rescaled Hamming metrics.

  – First, let us upper bound its ε-covering number. Let $S = \{1, 2, \ldots, \lceil (1-\varepsilon)d \rceil\}$. Consider the set of binary vectors

$$T(\varepsilon) := \{\theta \in H \mid \theta_j = 0, \ \forall \ j \in S\}$$

  By construction, for any binary vector $\theta' \in H$, we can find a vector $\theta \in T(\varepsilon)$ such that $\rho_H(\theta, \theta') \leq \varepsilon$. Namely, we can match θ' exactly on all entries $j \in S$, and, in the worst case, disagree on all the remaining $\lfloor \varepsilon d \rfloor$ positions. Since $T(\varepsilon)$ contains $2^{\lceil (1-\varepsilon)d \rceil}$ vectors, we can conclude that

$$\frac{\log N_H(\varepsilon; \mathcal{H}^d)}{\log 2} \leq \lceil (1-\varepsilon)d \rceil$$

  – Now let us lower bound its ε-covering number, where $\varepsilon \in (0, 1/2)$. If $\{\theta_1, \ldots, \theta_n\}$ is a ε-covering, then the (unrescaled) Hamming balls of radius εd around each θ_l must contain all 2^d vectors in the binary hypercube.

  Let $s = \lfloor \varepsilon d \rfloor$, then for each θ_l, there are exactly $\sum_{j=0}^{s} \binom{d}{j}$ binary vectors lying within distance εd

from it, and hence we must have

$$n \cdot \sum_{j=0}^{s} \binom{d}{j} \geq 2^d$$

where n is the cardinality of delta-covering set. Noq let $X_i \in \{0, 1\}$ be i.i.d. Bernoulli variables with parameter $1/2$. Rearranging the previous inequality, we have

$$\frac{1}{n} \leq \sum_{j=0}^{s} \binom{d}{j} 2^{-d} = \mathbb{P}\left(\sum_{i=1}^{d} X_i \leq \varepsilon d\right) \leq e^{-2d(1/2-\varepsilon)^2}$$

where the last inequality follows by applying Hoeffding's bound to the sum of d i.i.d. Bernoulli variables. Following some algebra, we obtain the lower bound

$$\log N_H(\varepsilon; \mathcal{H}^d) \geq 2d\left(\frac{1}{2} - \varepsilon\right)^2$$

valid for $\varepsilon \in (0, 1/2)$. This lower bound is qualitatively correct, but can be tightened by using a better upper bound on the binomial tail probability.

Definition 6.2 (Packing Number)**.** A ε-packing number of a set T with respect to a metric ρ is a set $\{\theta_1, \ldots, \theta_m\} \subset T$ such that $\rho(\theta_i, \theta_j) > \varepsilon$ for all distinct $i, j \in \{1, \ldots, m\}$. The ε-packing number $M(\varepsilon; T, \rho)$ is the cardinality of the largest ε-packing.

Lemma 6.1. For all $\varepsilon > 0$, the packing and covering numbers are related as follows:

$$M(2\varepsilon; T, \rho) \leq N(\varepsilon; T, \rho) \leq M(\varepsilon; T, \rho) \tag{6.1}$$

6.2 Bound Rademacher Complexity via Covering Number

Recall that in Section 4 where we introduce the Rademacher compelxity, we define a new variable Z over domain $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, and a function h from hypothesis space $\mathcal{H} = \{h : (Y, X) \mapsto l(Y, f(X)) \mid f \in \mathcal{F}\}$, where $l : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$ is some loss function.

Given a sample set $S_n = \{Z_1, \ldots, Z_n\} \in \mathcal{Z}^n$, define the following pseudo-norms on the space $\mathcal{H}$: for any $h \in \mathcal{H}$,

$$\begin{aligned} ||h||_p &:= \left(\frac{1}{n}\sum_{i=1}^{n} |h(X_i)|^p\right)^{1/p}, \quad p \geq 1 \\ ||h||_\infty &:= \max_i |h(X_i)| \end{aligned} \tag{6.2}$$

Note that the reason why we call it pseudo-norm is the space $\mathcal{H}$ is not required to be finite. Now, let $\rho_p(\theta, \theta') := ||\theta - \theta'||_p$ be the pseudo-metric that is induced by the pseudonorms $||\cdot||_p$ on function space $\mathcal{H}$. Then, we can similarily define the covering number $N(\delta; \mathcal{H}, \rho)$ and packing numbers $M(\delta; \mathcal{H}, \rho)$ on pseudo-metric space $(\mathcal{H}, \rho)$.

Lemma 6.2 (Monotonicity)**.** For any $S_n = \{Z_1, \ldots, Z_n\} \in \mathcal{Z}^n$, $1 \leq p \leq q$ and $\delta > 0$, we have

$$N(\delta; \mathcal{H}, \rho_p) \leq N(\delta; \mathcal{H}, \rho_q) \tag{6.3}$$

$$M(\delta; \mathcal{H}, \rho_p) \leq M(\delta; \mathcal{H}, \rho_q) \tag{6.4}$$

Proof. That is because p-norm is decreasing in p when $p \geq 1$. Consider two norms $||\cdot||_p$ and $||\cdot||_q$ of vector x where $1 \leq p \leq q$.

- If $x = 0$, then $||x||_p \geq ||x||_q$ trivially holds;
- If $x > 0$, then let y be a vector such that $y_k = ||x_k||/||x||_q \leq 1$, which means that $y_k^p \geq y_k^q$. Notice that $\sum_k y_k^q = 1$, we have

$$||y||_p \geq 1$$

Hence we have shown that p-norm is decreasing in p, and therefore the covering number with respect to the metric ρ_p induced by p-norm is less than the covering number with respect to metric ρ_q induced by q-norm. □

We next show that the covering numbers of the pseudo-metric space $(\mathcal{F}, \rho_1)$ can be used to bound the empirical Rademacher complexity.

Theorem 6.3 (Bounding Rademacher Complexity via Covering Number)**.** For any fixed sample set $S_n = \{Z_1, \ldots, Z_n\} \in \mathcal{Z}^n$ with size n, let $\sup_{h\in\mathcal{H}} ||h(Z)||_2 \leq c(Z)$, then the empirical Rademacher complexity is bounded by

$$\hat{R}_{S_n}(\mathcal{H}) \leq \inf_{\varepsilon>0} \left\{ \varepsilon + c(Z)\sqrt{\frac{2\log N(\varepsilon; \mathcal{H}, \rho_1)}{n}} \right\} \tag{6.5}$$

recall that ρ_1 is a pseudo-metirc induced by pseudo-norm $||\cdot||_1$, and note that the hypothesis space $\mathcal{H}$ is not required to be finite.

Proof. For a fixed sample $S_n = \{Z_1, \ldots, Z_n\}$ drawn from a unknown joint probability P and $\varepsilon > 0$, we denote by $\mathcal{H}_{S_n}$ the set of vectors of function $H = (h(Z_1), \ldots, h(Z_n))^\top$ where $h \in \mathcal{H}$. Since we know the l_2-norm of function h with respect to Z is bounded by $c(Z)$, i.e. $\sup_{h\in\mathcal{H}} ||h(Z)||_2 \leq c(Z)$, then the norm of these vectors is bounded by $\sqrt{n}c(Z)$, namely

$$\sup_{H\in\mathcal{H}_{S_n}} ||H||_2 = \sup_{h\in\mathcal{H}} \sqrt{\sum_{i=1}^n h(Z_i)^2} = \sup_{h\in\mathcal{H}} \sqrt{n} \cdot ||h||_2 \leq \sqrt{n} \cdot c(Z)$$

Now, Let $\mathcal{C} \subset \mathcal{H}$ be a minimal ε-cover of space $(\mathcal{H}; \rho_1)$, and for any $h \in \mathcal{H}$, let $h_0 \in \mathcal{C}$ such that $||h - h_0||_1 \leq \varepsilon$. Apply the Massart's lemma, we have

$$\begin{aligned}
\hat{R}_{S_n}(\mathcal{H}) &= \mathbb{E}_\sigma \left[\sup_{h\in\mathcal{H}} \frac{1}{n} \sum_{i=1}^n \sigma_i h(Z_i) \right] \\
&\leq \mathbb{E}_\sigma \left[\sup_{h\in\mathcal{H}} \frac{1}{n} \sum_{i=1}^n \sigma_i (h(Z_i) - h_0(Z_i)) \right] + \mathbb{E}_\sigma \left[\sup_{h\in\mathcal{H}} \frac{1}{n} \sum_{i=1}^n \sigma_i h_0(Z_i) \right] \\
&\leq \varepsilon + \mathbb{E}_\sigma \left[\sup_{h\in\mathcal{C}} \frac{1}{n} \sum_{i=1}^n \sigma_i h(Z_i) \right] \\
&\leq \varepsilon + \sup_{h\in\mathcal{C}} \sqrt{\sum_{i=1}^n h(Z_i)^2} \cdot \frac{\sqrt{2\log|\mathcal{C}|}}{n} \quad \text{(Massart's lemma in Thm. 5.1)} \\
&\leq \varepsilon + c(Z)\sqrt{\frac{2\log N(\varepsilon; \mathcal{H}, \rho_1)}{n}}
\end{aligned} \tag{6.6}$$

It is crucial to notice that Massart's lemma only holds on the finite set and here the ε-cover $\mathcal{C}$ meets the requirement. The final result follows by taking the infimum over $\varepsilon > 0$.

□

The bound in Theorem 6.3 establishes a tradeoff with respect to the precision parameter ε, as the decrease of ε woud lead to the increase of covering number $N(\varepsilon; \mathcal{H}, \rho_1)$. In addition, this bound is sample-dependent, namely a random variable, as the right-hand-side depends on $S_n \in \mathcal{Z}^n$

6.3 Chaining

Theorem 6.3 is established by using one fixed level of granularity ($\varepsilon > 0$) at a time, and taking the infimum over $\varepsilon > 0$ to obtain the final bound. An improved version of this result can be established by integrating over different levels of granularity. In this case, we need to work with covering numbers for the pseudo-metric space $(\mathcal{H}, \rho_2)$ where ρ_2 is induced by the pseudo-norm $||\cdot||_2$.

Theorem 6.4 (Dudley's Entropy Integral Bound)**.** For any fixed sample set $S_n = (Z_1, \ldots, Z_n) \in \mathcal{Z}^n$ and $\sup_{h\in\mathcal{H}} ||h(Z)||_2 \leq c(Z)$, we have

$$\hat{R}_{S_n}(\mathcal{H}) \leq \inf_{\varepsilon\in[0,c(Z)/2]} \left\{ 4\varepsilon + \frac{12}{\sqrt{n}} \int_{\varepsilon}^{c(Z)/2} d\nu \sqrt{\log N(\nu; \mathcal{H}, \rho_2)} \right\} \tag{6.7}$$

note that the hypothesis space $\mathcal{H}$ is not required to be finite.

Proof. Fix the n-size sample $S_n = \{Z_1, \ldots, Z_n\} \in \mathcal{Z}^n$. For each $j \in \mathbb{N}_+$, let

$$\varepsilon_j := c(Z)/2^j$$

and let $\mathcal{C}_j \subset \mathcal{H}$ be a minimal ε_j-cover of pseudo-metric space $(\mathcal{H}, \rho_2)$. We then have $|\mathcal{C}_j| = N(\varepsilon_j; \mathcal{H}, \rho_2)$. For any $h \in \mathcal{H}$ and $j \in \mathbb{N}_+$, let $h_j \in \mathcal{C}_j$ such that $||h - h_j||_2 \leq \varepsilon_j$. The sequence $h_1, h_2, \ldots$ (of elements of cover with decreasing radius) converges towards h. This sequence can be used to define the following telescoping sum, for a given $m \in \mathbb{N}$ to be choose later:

$$h = h - h_m + \sum_{j=1}^{m} (h_j - h_{j-1})$$

with $h_0 := 0$. This telescoping sum can be thought of as a "chain" connecting $h_0 = 0$ to h. This is the reason why the technique we are going to describe is called *chaining*. Upon these, we have

$$\begin{aligned}
\hat{R}_{S_n}(\mathcal{H}) &= \mathbb{E}_\sigma \left[\sup_{h\in\mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i h(Z_i) \right] \\
&\leq \mathbb{E}_\sigma \left[\sup_{h\in\mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i (h(Z_i) - h_m(Z_i)) \right] + \mathbb{E}_\sigma \left[\sup_{h\in\mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \sum_{j=1}^{m} (h_j(Z_i) - h_{j-1}(Z_i)) \right]
\end{aligned}$$

Next, we bound the two summands separately. The first summand is bounded by ε_m as

$$\sum_{i=1}^{n} \sigma_i (h(Z_i) - h_m(Z_i)) \leq \sum_{i=1}^{n} |h(Z_i) - h_m(Z_i)| = n \cdot ||h - h_m||_1 \leq n \cdot ||h - h_m||_2 \leq n \cdot \varepsilon_m$$

Since there are at most $|\mathcal{C}_j| \cdot |\mathcal{C}_{j-1}|$ different ways to create a vector in $\mathbb{R}^n$ of the form

$$\begin{bmatrix} h_j(Z_1) - h_{j-1}(Z_1) \\ \vdots \\ h_j(Z_n) - h_{j-1}(Z_n) \end{bmatrix}$$

with $h_j \in \mathcal{C}_j$ and $h_{j-1} \in \mathcal{C}_{j-1}$, using Massart's lemma in Theorem 5.1 and let $\mathcal{C} = \bigcup_{j=1}^m$ be the union of all covers, the second summand can be upper bounded by

$$\begin{aligned}
\sum_{j=1}^m \mathbb{E}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n\sum_{i=1}^n \sigma_i(h_j(Z_i) - h_{j-1}(Z_i))\right] &= \sum_{j=1}^m \mathbb{E}\left[\sup_{h\in\mathcal{C}} \frac{1}{n}\sum_{i=1}^n\sum_{i=1}^n \sigma_i(h_j(Z_i) - h_{j-1}(Z_i))\right] \\
&\leq \sum_{j=1}^m \sup_{h\in\mathcal{C}} \sqrt{\sum_{i=1}^n (h_j(Z_i) - h_{j-1}(Z_i))^2} \cdot \frac{\sqrt{2\log|\mathcal{C}_j| \cdot |\mathcal{C}_{j-1}|}}{n} \\
&= \sum_{j=1}^m \sup_{h\in\mathcal{C}} ||h_j - h_{j-1}||_2 \cdot \sqrt{\frac{2\log|\mathcal{C}_j| \cdot |\mathcal{C}_{j-1}|}{n}}
\end{aligned}$$

Here we can see that the $||\cdot||_2$ norm naturally appears in the application of Massart's lemma. With the triangular inequality for the pseudo-norm $||\cdot||_2$, we have (using that $\varepsilon_{k-1} = 2\varepsilon_k$)

$$\begin{aligned}
||h_j - h_{j-1}||_2 &\leq ||h_j - h||_2 + ||h - h_{j-1}||_2 \\
&\leq \varepsilon_j + \varepsilon_{j-1} = 3\varepsilon_j = 6(\varepsilon_j - \varepsilon_{j+1})
\end{aligned}$$

Also, $|\mathcal{C}_j| = N(\varepsilon_j; \mathcal{H}, \rho_1)$ and $|\mathcal{C}_{j-1}| \leq |\mathcal{C}_j|$. Putting things together, we have

$$\begin{aligned}
\hat{R}_{S_n}(\mathcal{H}) &\leq \varepsilon_m + 12\sum_{j=1}^m (\varepsilon_j - \varepsilon_{j+1})\sqrt{\frac{\log N(\varepsilon_j; \mathcal{H}, \rho_1)}{n}} \\
&\leq 2\varepsilon_{m+1} + 12\int_{\varepsilon_{m+1}}^{c(Z)/2} d\nu \sqrt{\log N(\delta; \mathcal{H}, \rho_1)}
\end{aligned}$$

where the last inequality follows as the integral is lower-bound by its lower Riemann sum as the function $\nu \mapsto N(\nu; \mathcal{H}, \rho_1)$ is decreasing. For any $\varepsilon \in [0, c(Z)]/2$, choose m such that $\varepsilon < \varepsilon_{m+1} \leq 2\varepsilon$. The statement of the theorem thus follows by taking the infimum over $\varepsilon \in [0, c(Z)/2]$.

□

References

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

Bartlett, Peter L, Michael I Jordan, and Jon D McAuliffe (2006). "Convexity, classification, and risk bounds". In: *Journal of the American Statistical Association* 101.473, pp. 138–156.

Mohri, Mehryar, Afshin Rostamizadeh, and Ameet Talwalkar (2018). *Foundations of machine learning*. MIT press.

Shalev-Shwartz, Shai and Shai Ben-David (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Sridharan, Karthik, Shai Shalev-Shwartz, and Nathan Srebro (2008). "Fast rates for regularized objectives". In: *Advances in neural information processing systems* 21.

Tropp, Joel A (2012). "User-friendly tail bounds for sums of random matrices". In: *Foundations of computational mathematics* 12.4, pp. 389–434.

Wainwright, Martin J (2019). *High-dimensional statistics: A non-asymptotic viewpoint*. Vol. 48. Cambridge University Press.

A Norms

A.1 Norms

A function $f : \mathbb{R}^n \to \mathbb{R}$ with dom $f = \mathbb{R}^n$ is called a norm if

- f is nonnegative: $f(x) \geq 0$ for all $x \in \mathbb{R}$
- f is definite: $f(x) = 0$ only if $x = 0$
- f is homogeneous: $f(tx) = |t|f(x)$, for all $x \in \mathbb{R}^n$ and $t \in \mathbb{R}$
- f satisfies the triangle inequality: $f(x + y) \leq f(x) + f(y)$, for all $x, y \in \mathbb{R}^n$

A norm is a measure of the length of a vector x; we can measure the distance between two vectors x and y as the length of their difference, *i.e.*

$$\mathrm{dist(x, y)} = ||\mathrm{x} - \mathrm{y}||$$

The set of all vectros with norm less than or equal to one,

$$\mathcal{B} = \{x \in \mathbb{R}^n \mid ||x|| \leq 1\}$$

is called the unit ball of the norm $||\cdot||$. The unit ball satisfies the following properties:

- $\mathcal{B}$ is symmetric about the origin; $x \in \mathcal{B}$ iff $-x \in \mathcal{B}$
- $\mathcal{B}$ is convex
- $\mathcal{B}$ is closed, bounded, and has nonempty interior

A.2 Examples of Norm

Here we consider the norm for vector $x \in \mathbb{R}^n$

- l_1-norm
$$||x||_1 = |x_1| + \cdots + |x_n|$$
- l_∞-norm
$$||x||_\infty = \max\{|x_1|, \ldots, |x_n|\}$$
- l_p-norm
$$||x||_p = \left(|x_1|^p + \cdots + |x_n|^p\right)^{1/p}$$
- P-quadratic norms: for matrix $P \in \mathbb{S}^n_{++}$,
$$||x||_P = (x^\top P x)^{1/2} = ||P^{1/2} x||_2$$
The unit ball of a quadratic norm is an ellipsoid (and conversely, if the unit ball of a norm is an ellipsoid, the norm is a quadratic norm)
- Frobenius norm: for matrix $X \in \mathbb{R}^{m \times n}$,
$$||X||_F = \left(\mathrm{tr}(\mathrm{X}^\top \mathrm{X})\right)^{1/2} = \left(\sum_{\mathrm{i}=1}^{\mathrm{m}} \sum_{\mathrm{j}=1}^{\mathrm{n}} \mathrm{X}_{\mathrm{ij}}^2\right)^{1/2} \tag{A.1}$$

The Frobenius norm is the Euclidean norm of the vector obtained by listing the coefficients of the matrix. It is different from the l_2-norm of matrix.

A.3 Equivalence of Norms

Suppose that $||\cdot||_a$ and $||\cdot||_b$ are norms on $\mathbb{R}^n$. A basic result of analysis is that there exist positive constants α and β such that, for all $x \in \mathbb{R}$,

$$\alpha||x||_a \leq ||x||_b \leq \beta||x||_a$$

This means that the norms are equivalent, i.e., they define the same set of open subsets, the same set of covergent sequences, and so on. Using convex analysis, we can give a more specific result: if $||\cdot||$ is any norm on $\mathbb{R}^n$, then there exists a quadratic norm $||\cdot||_P$ for which

$$||x||_P \leq ||x|| \leq \sqrt{n}||x||_P$$

holds for all x. In other words, any norm on $\mathbb{R}^n$ can be uniformly approximated, within a factor of $\sqrt{n}$, by a P-quadratic norm. **We conclude that any norms on all finite-dimensional vector space are equivalent, but on infinite-dimensional vector spaces, the result need not hold.**

A.4 Operator Norms

Supose $||\cdot||_a$ and $||\cdot||_b$ are norms on $\mathbb{R}^m$ and $\mathbb{R}^n$, respectively. We define the *opertaor norm* of $X \in \mathbb{R}^{m\times n}$, induced by the norms $||\cdot||_a and ||\cdot||_b$, as

$$||X||_{a,b} = \sup\left\{||Xu||_a \mid ||u||_b \leq 1\right\} \tag{A.2}$$

It can be shown that this defines a norm on $\mathbb{R}^{m\times n}$.

- When $||\cdot||_a$ and $||\cdot||_b$ are both Euclidean norms, the operator norm of X is its *maximum singular value*, and is denoted $||X||_2$:

$$||X||_2 = \sigma_{\max}(X) = \left(\lambda_{\max}(X^\top X)\right)^{1/2} \tag{A.3}$$

  That is because, $X^\top X$ is a symmetric matrix, which satisfy

$$u^\top(X^\top X)u \leq \lambda_{\max}(X^\top X)u^\top u$$

  This agress with the Euclidean norm on $\mathbb{R}^m$, when $X \in \mathbb{R}^{m\times 1}$, so there is not clash of notation. This norm is also called the *spectral norm* or l_2*-norm* of X.

- The norm induced by the l_∞-norm on $\mathbb{R}^m$ and $\mathbb{R}^n$, denoted $||X||_\infty$, is the *max-row-sum* norm

$$||X||_\infty = \sup\left\{||Xu||_\infty \mid ||u||_\infty \leq 1\right\} = \max_{i=1,\ldots,m}\sum_{j=1}^{n}|X_{ij}|$$

- The norm induced by the l_1-norm on $\mathbb{R}^m$ and $\mathbb{R}^n$, denoted $||X||_1$, is the *max-column-sum* norm

$$||X||_1 = \max_{j=1,\ldots,n}\sum_{i=1}^{m}|X_{ij}|$$

B Probability Theory

B.1 Independence

Definition B.1 (Independent)**.** Two random variables X and Y are independent if, for every A and B,

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$$

and we write $X \perp\!\!\!\perp Y$.

In principle, to check whether X and Y are independent we need to check the above equation for all subsets A and B. Fortunately, we have the following result which we state for continuous random variables though it is true for discrete random variables too.

Theorem B.1. Let X and Y have joint PDF $f_{X,Y}$. Then $X \perp\!\!\!\perp Y$ if and only if

$$f_{X,Y}(x, y) = f_X(x) f_Y(y)$$

for all values x and y.

Definition B.2 (Conditional Independent)**.** Let X, Y and Z be random variables. X and Y are conditionally independent given Z, wirtten $X \perp\!\!\!\perp Y \mid Z$, if

$$f_{X,Y|Z}(x, y \mid z) = f_{X|Z}(x \mid z) f_{Y|Z}(y \mid z)$$

for all x, y and z.

Intuitively, this means that, once you know Z, Y provides no extra information about X. An equivalent definition is that

$$f_{X|Y,Z}(x \mid y, z) = f_{X|Z}(x \mid z)$$

Here are some rules of the conditional independence:

- **Symmetry**

$$X \perp\!\!\!\perp Y \quad \Rightarrow \quad Y \perp\!\!\!\perp X$$

- **Decomposition**

$$X \perp\!\!\!\perp (A, B) \quad \Rightarrow \quad \text{and} \begin{cases} X \perp\!\!\!\perp A \\ X \perp\!\!\!\perp B \end{cases}$$

  Proof:

$$\begin{aligned} f_{X,A}(x, a) &= \int_B f_{X,A,B}(x, a, b) db \\ &= \int_B f_X(x) f_{A,B}(a, b) db \\ &= f_X(x) f_A(a) \end{aligned}$$

  A similar proof shows the independence of X and B.

- **Weak Union**

$$X \perp\!\!\!\perp (A, B) \quad \Rightarrow \quad \text{and} \begin{cases} X \perp\!\!\!\perp A \mid B \\ X \perp\!\!\!\perp B \mid A \end{cases}$$

  Proof:

– by assumption, we have $\mathbb{P}(X) = \mathbb{P}(X \mid A, B)$

– due to the property of decomposition $X \perp\!\!\!\perp B$, we have $\mathbb{P}(X) = \mathbb{P}(X \mid B)$

Combining the above two equalities yields

$$\mathbb{P}(X \mid B) = \mathbb{P}(X \mid A, B)$$

which establishes $X \perp\!\!\!\perp A \mid B$. A similar proof shows the second condition.

- **Contraction**

$$\left.\begin{array}{r} X \perp\!\!\!\perp A \mid B \\ X \perp\!\!\!\perp B \end{array}\right\} \text{ and } \quad \Rightarrow \quad X \perp\!\!\!\perp (A, B)$$

or similarily

$$\left.\begin{array}{r} X \perp\!\!\!\perp B \mid A \\ X \perp\!\!\!\perp A \end{array}\right\} \text{ and } \quad \Rightarrow \quad X \perp\!\!\!\perp (A, B)$$

Proof: this property can be proved by noticing that

$$\mathbb{P}(X \mid A, B) = \mathbb{P}(X \mid B) = \mathbb{P}(X)$$

each equality of which is asserted by $X \perp\!\!\!\perp A \mid B$ and $X \perp\!\!\!\perp B$, respectively. A similar proof shows the second one.

- **Intersection**

for strictly positive probability distributions, the following also hold

$$\left.\begin{array}{r} X \perp\!\!\!\perp Y \mid Z, W \\ X \perp\!\!\!\perp W \mid Z, Y \end{array}\right\} \text{ and } \quad \Rightarrow \quad X \perp\!\!\!\perp (W, Y) \mid Z$$

Proof: by assumption

$$\mathbb{P}(X \mid Z, W, Y) = \mathbb{P}(X \mid Z, W) = \mathbb{P}(X \mid Z, Y)$$

Using this equality, together with the law of total probability applied to $\mathbb{P}(X \mid Z)$

$$\begin{aligned} \mathbb{P}(X \mid Z) &= \sum_{w \in W} \mathbb{P}(X \mid Z, W = w) \mathbb{P}(W = w \mid Z) \\ &= \sum_{w \in W} \mathbb{P}(X \mid Y, Z) \mathbb{P}(W = w \mid Z) \\ &= \mathbb{P}(X \mid Z, Y) \sum_{w \in W} \mathbb{P}(W = w \mid Z) \\ &= \mathbb{P}(X \mid Z, Y) \end{aligned}$$

This suggest

$$\mathbb{P}(X \mid Z, W, Y) = \mathbb{P}(X \mid Z)$$

which establishs $X \perp\!\!\!\perp (W, Y) \mid Z$.

In general, we have

$$\begin{aligned} X \perp\!\!\!\perp (Y, Z) \quad &\Leftrightarrow \quad X \perp\!\!\!\perp Y \text{ and } X \perp\!\!\!\perp Y \mid Z \\ &\Leftrightarrow \quad X \perp\!\!\!\perp Z \text{ and } X \perp\!\!\!\perp Z \mid Y \end{aligned} \tag{B.1}$$

B.2 Expectations

Definition B.3. The expectation, or mean, or first moment, of random variable X is defined to be

$$\mathbb{E}[X] = \int_{\mathcal{X}} x dF(x) = \begin{cases} \sum_{\mathcal{X}} x f(x) \\ \int_{\mathcal{X}} x f(x) dx \end{cases}$$

assuming that the sum (or integral) is well defined.

Theorem B.2. Let $Y = r(X)$, then

$$\mathbb{E}[Y] = \mathbb{E}[r(X)] = \int_{\mathcal{X}} r(x) dF_X(x)$$

Definition B.4. The conditional expectation of X given $Y = y$ is

$$\mathbb{E}[X \mid Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x) dx$$

If $r(x, y)$ is a function of x and y then

$$\mathbb{E}[r(X, Y) \mid Y = y] = \int_{-\infty}^{\infty} r(x, y) f_{X|Y}(x) dx$$

Theorem B.3 (Law of Total Expectations)**.** For random variables X and Y, assumping $\mathbb{E}[X]$ and $\mathbb{E}[Y]$ exists, then we have

$$\mathbb{E}\big[\mathbb{E}[X \mid Y]\big] = \mathbb{E}[X]$$

and more generally for any function $r(x, y)$

$$\mathbb{E}\big[\mathbb{E}[r(X, Y) \mid X]\big] = \mathbb{E}[r(X, Y)]$$

Proof. By definition B.4, we have

$$\begin{aligned} \mathbb{E}\big[\mathbb{E}[X \mid Y]\big] &= \int_{-\infty}^{\infty} \mathbb{E}[X \mid Y = y] f_Y(y) dy \\ &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} x f_{X|Y}(x) dx \right) f_Y(y) dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x \frac{f_{X,Y}(x, y)}{f_Y(y)} f_Y(y) dx dy \\ &= \int_{-\infty}^{\infty} x dx \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy \\ &= \int_{-\infty}^{\infty} x f_X(x) dx = \mathbb{E}[X] \end{aligned}$$

and similarily, we have

$$\begin{aligned}\mathbb{E}\big[\mathbb{E}[r(X,Y) \mid Y]\big] &= \int_{-\infty}^{\infty} \mathbb{E}[r(X,Y) \mid Y = y] f_Y(y) dy \\ &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} r(x,y) f_{X|Y}(x) dx \right) f_Y(y) dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} r(x,y) f_{X,Y}(x,y) dx dy \\ &= \mathbb{E}[r(X,Y)]\end{aligned} \tag{B.2}$$

□

Corollary B.4 (Law of Iterated Expectation)**.** For random variables X, Y, Z, we have

$$\mathbb{E}\big[\mathbb{E}[Z \mid X,Y] \mid Y\big] = \mathbb{E}[Z \mid Y] = \mathbb{E}\big[\mathbb{E}[Z \mid Y] \mid X,Y\big]$$

Proof. The first equality holds because of the fact that, for any y

$$\begin{aligned}\mathbb{E}\big[\mathbb{E}[Z \mid X,Y] \mid Y = y\big] &= \mathbb{E}[r(X,Y) \mid Y = y] \\ &= \int_{-\infty}^{\infty} r(X,Y) f_{X|Y}(x) dx \\ &= \int_{-\infty}^{\infty} \mathbb{E}[Z \mid X,Y] f_{X|Y}(x) dx \\ &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} z f_{Z|X,Y}(z) dz \right) f_{X|Y}(x) dx \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} z \frac{f_{Z,X,Y}(z,x,y)}{f_{X,Y}(x,y)} \frac{f_{X,Y}(x,y)}{f_Y(y)} dx dz \\ &= \int_{-\infty}^{\infty} z dz \int_{-\infty}^{\infty} f_{Z,X|Y}(z,x) dx \\ &= \int_{-\infty}^{\infty} z f_{Z|Y}(z) dz = \mathbb{E}[Z \mid Y = y]\end{aligned}$$

and for any y

$$\begin{aligned}\mathbb{E}\big[\mathbb{E}[X \mid Y] \mid X, Y = y\big] &= \mathbb{E}\big[\mathbb{E}[X \mid Y = y] \mid X, Y = y\big] \\ &= \mathbb{E}[X \mid Y = y] \cdot \mathbb{E}[1 \mid X, Y = y] = \mathbb{E}[X \mid Y = y]\end{aligned}$$

□

Theorem B.5 (Independencec)**.** For random variables X and Y, we have

$$\mathbb{E}[XY] = \mathbb{E}[X \mid Y] \cdot \mathbb{E}[Y]$$

If X is independent of Y, i.e. $X \perp\!\!\!\perp Y$, then we have

$$\mathbb{E}[X \mid Y] = \mathbb{E}[X]$$

and consequently

$$\mathbb{E}[XY] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$$

Proof. By definition

$$\begin{aligned}\mathbb{E}[X \mid Y] &= \int x f_{X|Y}(x)dx \\ &= \int x \frac{f_{X,Y}(x,y)}{f_Y(y)}dx \\ &= \int x f_X(x)dx \qquad (X \perp\!\!\!\perp Y) \\ &= \mathbb{E}[X]\end{aligned}$$

□

Definition B.5. The condtional variance is defined as

$$\begin{aligned}\mathrm{Var}(X \mid Y = y) &= \mathbb{E}\big(X - \mathbb{E}[X \mid Y = y]\big)^2 \\ &= \int_{-\infty}^{\infty} \big(x - \mathbb{E}[X \mid Y = y]\big)^2 f_{X|Y}(x)dx\end{aligned} \tag{B.3}$$

Theorem B.6 (Law of Total Variance)**.** For random variables X and Y,

$$\mathrm{Var}(X) = \mathbb{E}[\mathrm{Var}(X \mid Y)] + \mathrm{Var}(\mathbb{E}[X \mid Y]) \tag{B.4}$$

Proof. Notice that

$$\begin{aligned}\mathbb{E}[\mathrm{Var}(X \mid Y)] &= \mathbb{E}\Big[\mathbb{E}[X^2 \mid Y] - \mathbb{E}[X \mid Y]^2\Big] \\ &= \mathbb{E}[X^2] - \mathbb{E}\big[\mathbb{E}[X \mid Y]^2\big]\end{aligned}$$

and

$$\begin{aligned}\mathrm{Var}(\mathbb{E}[X \mid Y]) &= \mathbb{E}\big[\mathbb{E}[X \mid Y]^2\big] - \mathbb{E}\big[\mathbb{E}[X \mid Y]\big]^2 \\ &= \mathbb{E}\big[\mathbb{E}[X \mid Y]^2\big] - \mathbb{E}[X]^2\end{aligned}$$

Adding these two together yields $\mathrm{Var}(X)$. □

B.3 Convergences

Definition B.6 (Type of Convergence)**.** Let $X_1, X_2, \ldots,$ be a sequence of random variables and let X be another random variable. Let F_n denote the CDF of X_n and let F denote the CDF of X_n and F the CDF of X.

1. X_n converges to X **in quadratic mean** (convergence in L_2), written $X_n \xrightarrow{qm} X$, if

$$\mathbb{E}[X_n - X]^2 \to 0 \tag{B.5}$$

   as $n \to \infty$.

2. X_n converges to x in L_1, written $X_n \xrightarrow{L_1} X$, if

$$\mathbb{E}|X_n - X| \to 0 \tag{B.6}$$

   as $n \to \infty$.

3. X_n converges to X **almost surely**, written $X_n \xrightarrow{a.s.} X$, if

$$\mathbb{P}\big(\{s : X_n(s) \to X(s)\}\big) = 1 \tag{B.7}$$

4. X_n converges to X **in probability**, written $X_n \xrightarrow{P} X$, if, for every $\varepsilon > 0$,

$$\mathbb{P}(|X_n - X| > \varepsilon) \to 0 \tag{B.8}$$

as $n \to \infty$.

5. X_n converges to X **in distribution**, written $X_n \rightsquigarrow X$, if

$$\lim_{n\to\infty} F_n(t) = F(t) \tag{B.9}$$

at all t for which F is continuous

Theorem B.7. The following relationships hold:

1. $X_n \xrightarrow{qm} X$ implies that $X_n \xrightarrow{L_1} X$
2. $X_n \xrightarrow{L_1} X$ implies that $X_n \xrightarrow{P} X$
3. $X_n \xrightarrow{a.s.} X$ implies that $X_n \xrightarrow{P} X$
4. $X_n \xrightarrow{qm} X$ implies that $X_n \xrightarrow{P} X$
5. $X_n \xrightarrow{P} X$ implies that $X_n \rightsquigarrow X$
6. If $X_n \rightsquigarrow X$ and if $\mathbb{P}(X = c) = 1$ for some real number c, then $X_n \xrightarrow{P} P$

In general, none of the reverse implications hold except the special case in 3.

Proof. Consider the following

1. Suppose
2.
3.
4. Suppose that $X_n \xrightarrow{qm} X$. Fix $\varepsilon > 0$ and use Markov's inequalit,

$$\mathbb{P}(|X_n - X| > \varepsilon) = \mathbb{P}(|X_n - X|^2 > \varepsilon^2) \leq \frac{\mathbb{E}|X_n - X|^2}{\varepsilon^2} \to 0$$

5. Fix $\varepsilon > 0$ and let x be a continuity point of F, then

$$\begin{aligned} F_n(x) = \mathbb{P}(X_n \leq x) &= \mathbb{P}(X_n \leq x, X \leq x + \varepsilon) + \mathbb{P}(X_n \leq x, X > x + \varepsilon) \\ &\leq \mathbb{P}(X \leq x + \varepsilon) + \mathbb{P}(|X_n - X| > \varepsilon) \\ &= F(x + \varepsilon) + \mathbb{P}(|X_n - X| > \varepsilon) \end{aligned}$$

Also,

$$\begin{aligned} F(x - \varepsilon) = \mathbb{P}(X \leq x - \varepsilon) &= \mathbb{P}(X \leq x - \varepsilon, X_n \leq x) + \mathbb{P}(X \leq x - \varepsilon, X_n > x) \\ &\leq F_n(x) + \mathbb{P}(|X_n - X| > \varepsilon) \end{aligned}$$

Hence,

$$F(x - \varepsilon) - \mathbb{P}(|X_n - X| > \varepsilon) \leq F_n(x) \leq F(x + \varepsilon) + \mathbb{P}(|X_n - X| > \varepsilon)$$

Take the limit as $n \to \infty$ to conclude that

$$F(x-\varepsilon) \leq \lim_{n\to\infty} \inf F_n(x) \leq \lim_{n\to\infty} \sup F_n(x) \leq F(x+\varepsilon)$$

This holds for any $\varepsilon > 0$. Take the limit as $\varepsilon \to 0$ and use the fact that F is continuous at x, we conclude that

$$\lim_{n\to\infty} F_n(x) = F(x)$$

6. Fix $\varepsilon > 0$, then

$$\begin{aligned}
\mathbb{P}(|X_n - c| > \varepsilon) &= \mathbb{P}(X_n < c-\varepsilon) + \mathbb{P}(X_n > c+\varepsilon) \\
&\leq \mathbb{P}(X_n \leq c-\varepsilon) + \mathbb{P}(X_n > c+\varepsilon) \\
&= F_n(c-\varepsilon) + 1 - F_n(c+\varepsilon) \\
&\to F(c-\varepsilon) + 1 - F(c+\varepsilon) \\
&= 0 + 1 - 1 = 0
\end{aligned}$$

□

Theorem B.8 (Slutsky's Theorem)**.** Let X_n, Y_n and X, Y be random variables,

1. If $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$, then $X_n + Y_n \xrightarrow{P} X + Y$
2. If $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$, then $X_n Y_n \xrightarrow{P} XY$
3. If $X_n \rightsquigarrow X$ and $Y_n \rightsquigarrow c$, then $X_n + Y_n \rightsquigarrow X + c$
4. If $X_n \rightsquigarrow X$ and $Y_n \rightsquigarrow c$, then $X_n Y_n \rightsquigarrow cX$

Theorem B.9 (Continuous Mapping Theorem)**.** Let X be a random variable, X_n be a sequence of random variables and g be a continuous function.

1. If $X_n \xrightarrow{a.s.} X$, then $g(X_n) \xrightarrow{a.s.} g(X)$
2. If $X_n \xrightarrow{P} X$, then $g(X_n) \xrightarrow{P} g(X)$
3. If $X_n \rightsquigarrow X$, then $g(X_n) \rightsquigarrow g(X)$

Theorem B.10 (The Weak Law of Large Numbers)**.** If $X_1, \ldots, X_n$ are I.I.D, then

$$\bar{X}_n \xrightarrow{P} \mu$$

WLLN means that the distribution of $\bar{X}_n$ becomes more concentrated around μ as n gets large.

Proof. Assume that $\sigma < \infty$. This is not necessary but it simplifies the proof. Using Chebyshev inequality,

$$\mathbb{P}(|\bar{X}_n - \mu| \geq \varepsilon) \leq \frac{\mathrm{Var}(\bar{X}_n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \to 0$$

□

Theorem B.11 (The Central Limit Theorem)**.** Let $X_1, \ldots, X_n$ be I.I.D with mean μ and variance σ^2, then

$$Z_n \equiv \frac{\bar{X}_n - \mu}{\sqrt{\mathrm{Var}(\bar{X}_n)}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \rightsquigarrow Z \tag{B.10}$$

where $Z \sim N(0,1)$. In other words,

$$\lim_{n\to\infty} \mathbb{P}(Z_n \le z) = \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \tag{B.11}$$

CLT suggests that the distribution (CDF, not PDF) of $\bar{X}_n$ can be approximated using a Normal disitribution. It's the probability statements that we are approximating, not the random variable itself.

Proof. Suppose there are n I.I.D random variables X_i with mean μ and variance σ^2. Let

$$Y_i = \frac{X_i - \mu}{\sigma}$$

and

$$Z_n = \frac{\sum_i Y_i}{\sqrt{n}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$$

Suppose the moment generating function (MGF) of Y_i is $\varphi_Y(t) = \mathbb{E}[e^{tY}]$, and it is finite in a neighborhood around $t = 0$. Then we have

$$\varphi_{Y_1+\cdots+Y_n}(t) = \mathbb{E}\left[e^{t(Y_1+\cdots+Y_n)}\right] = \mathbb{E}\left[e^{tY_i}\right]^n = \big(\varphi_Y(t)\big)^n$$

and consequently

$$\varphi_{Z_n}(t) = \mathbb{E}\left[e^{t\frac{Y_1+\cdots+Y_n}{\sqrt{n}}}\right] = \left[\varphi_Y\left(\frac{t}{\sqrt{n}}\right)\right]^n$$

Notice that

$$\varphi_Y'(0) = \mathbb{E}[Y] = 0 \quad \text{and} \quad \varphi_Y''(0) = \mathbb{E}[Y^2] = \operatorname{Var} Y = 1$$

So the Taylor expansion gives us

$$\begin{aligned}
\varphi_Y(t) &= \varphi_Y(0) + t\varphi_Y'(0) + \frac{t^2}{2!}\varphi_Y''(0) + \frac{t^3}{3!}\varphi_Y'''(0) + \cdots \\
&= 1 + 0 + \frac{t^2}{2} + \frac{t^3}{3!}\varphi_Y'''(0) + \cdots \\
&= 1 + \frac{t^2}{2} + \frac{t^3}{3!}\varphi_Y'''(0) + \cdots
\end{aligned}$$

Therefore,

$$\begin{aligned}
\varphi_{Z_n}(t) &= \left[\varphi_Y\left(\frac{t}{\sqrt{n}}\right)\right]^n \\
&= \left[1 + \frac{t^2}{2!n} + \frac{t^3}{3!n^{3/2}}\varphi_Y'''(0) + \cdots\right]^n \\
&= \left[1 + \frac{\frac{t^2}{2} + \frac{t^3}{3!n^{1/2}}\varphi'''(0) + \cdots}{n}\right]^n \\
&\to e^{t^2/2}
\end{aligned}$$

The last step results from the fact that $\left(1 + \frac{a_n}{n}\right)^n \to e^a$ if $a_n \to a$. Notice that the MGF of standard normal variable $Z \sim N(0,1)$ is just

$$\varphi_Z(t) = \mathbb{E}\left[e^{tZ}\right] = e^{t^2/2}$$

So we have

$$\varphi_{Z_n}(t) \to \varphi_Z(t) \quad \Rightarrow \quad Z_n \rightsquigarrow Z$$

□

Example. CLT implies that $Z_n = \sqrt{n}(\bar{X}_n - \mu)/\sigma$ approximately follows $N(0,1)$. However, we rarely know σ. Instead, we can estimate σ^2 from i.i.d samples $X_1, \ldots, X_n$ by

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

Suppose the assumptions in CLT hold, prove the following

- $S_n^2 \xrightarrow{P} \sigma^2$
- $\sqrt{n}(\bar{X}_n - \mu)/S_n \rightsquigarrow N(0,1)$

Proof. For the first statement, notice that by CLT, we have

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} \mathbb{E}[X]$$

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{P} \mathbb{E}[X^2]$$

Therefore, we can utilize the continuous mapping theorem (or Slutsky's theorem) and get

$$\begin{aligned}
S_n^2 &= \frac{n}{n-1}\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \\
&= \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2\right) \\
&= \frac{n}{n-1}\left(\mathbb{E}[X^2] - \mathbb{E}[X]^2\right) \quad \text{(continous mapping theorem)} \\
&= \frac{n}{n-1}\sigma^2 \to \sigma^2 \quad (\text{as } n \to \infty)
\end{aligned}$$

Now, the second statement can be shown trivially by Slutsky's theorem, that is

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}\frac{\sigma}{S_n} \rightsquigarrow N(0,1)$$

by noticing that $\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \rightsquigarrow N(0,1)$ and $\sigma/S_n \xrightarrow{P} 1$. □

Theorem B.12 (Multivariate Central Limit Theorem)**.** Let $X_1, \ldots, X_n$ be IID random vectors where

$$X_i = (X_{1i}\ X_{2i} \cdots X_{ki})^\top$$

with mean

$$\mu = (\mu_1\ \mu_2 \cdots \mu_k)^\top = (\mathbb{E}[X_{1i}]\ \mathbb{E}[X_{2i}] \cdots \mathbb{E}[X_{ki}])^\top$$

and variance matrix Σ. Let

$$\bar{X} = (\bar{X}_1\ \bar{X}_2 \cdots \bar{X}_k)$$

where $\bar{X}_j = n^{-1}\sum_{i=1}^n X_{ji}$. Then

$$\sqrt{n}(\bar{X} - \mu) \rightsquigarrow N(0, \Sigma)$$

If Y_n has a limiting Normal distribution then the delta method allows us to find the limiting distribution of $g(Y_n)$ where g is any smooth function (differentiable).

Theorem B.13 (The Delta Method)**.** Suppose that

$$\frac{\sqrt{n}(Y_n - \mu)}{\sigma} \rightsquigarrow N(0,1)$$

and that g is a differentiable function such that $g'(\mu) \neq 0$. Then

$$\frac{\sqrt{n}(g(Y_n) - g(\mu))}{|g'(\mu)|\sigma} \rightsquigarrow N(0,1)$$

In other words,

$$Y_n \simeq N\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{implies that} \quad g(Y_n) \simeq N\left(g(\mu), (g'(\mu))^2\frac{\sigma^2}{n}\right) \tag{B.12}$$

Theorem B.14 (The Multivariate Delta Method)**.** Suppose that $Y_n = (Y_{n1}, \ldots, Y_{nk})$ is a sequence of random vectors such that

$$\sqrt{n}(Y_n - \mu) \rightsquigarrow N(0, \Sigma)$$

Let $g : \mathbb{R}^k \to \mathbb{R}$ and let

$$\nabla g(y) = \left(\frac{\partial g}{\partial y_1} \cdots \frac{\partial g}{\partial y_k}\right)^\top$$

Let ∇_μ denote $\nabla g(y)$ evaluated at $y = \mu$ and assume that the elements of ∇_μ are nonzero. Then

$$\sqrt{n}(g(Y_n) - g(\mu)) \rightsquigarrow N(0, \nabla_\mu^\top \Sigma \nabla_\mu)$$

C Concentration of Measure

C.1 Markov Inequality

Theorem C.1 (Markov Inequality)**.** For any nonnegative random variable $X \geq 0$

$$P(X \geq t) \leq \frac{\mathbb{E}[X]}{t} = O\left(\frac{1}{t}\right) \tag{C.1}$$

Proof.

$$\mathbb{E}[X] = \int_0^\infty xp(x)dx \geq \int_t^\infty xp(x)dx \geq t\int_t^\infty = t\mathbb{P}(X \geq t)$$

□

Similarily, we can apply the same calculation and get

$$\mathbb{E}\left[(X-\mu)^k\right] \geq t^k\mathbb{P}\left((X-\mu)^k \geq t^k\right) = t^k\mathbb{P}(|X-\mu| \geq t)$$

that is to say

$$\mathbb{P}(|X-\mu| \geq t) \leq \frac{\mathbb{E}\left[(X-\mu)^k\right]}{t^k} \tag{C.2}$$

C.2 Chebyshev Inequality

Theorem C.2 (Chebyshev Inequality)**.** For any random variable X with variance σ^2, for any $t \geq 0$

$$\mathbb{P}\big(|X - \mu| \geq t\sigma\big) \leq \frac{1}{t^2} \tag{C.3}$$

Proof. This could be obatained immediatly by choosing $k = 2$ and $t = n\sigma$ from inequality (C.2), that is,

$$\mathbb{P}\big(|X - \mu| \geq t\sigma\big) \leq \frac{\sigma^2}{t^2\sigma^2} = \frac{1}{t^2}$$

□

Here is an example. Consider the average of i.i.d. random variables with mean μ and variance σ^2

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

It has mean μ and variance σ^2/n. Applying Chebyshev inequality, we have

$$\mathbb{P}\left(|\bar{X}_n - \mu| \geq \frac{t\sigma}{\sqrt{n}}\right) \leq \frac{1}{t^2}$$

with 0.99 proabability ($t = 10$), the average $\bar{X}_n$ would not exceed $\mu + 10\sigma/\sqrt{n}$. This would lead to the Weak Law of Large Numbers.

C.3 Chernoff's Methods

Theorem C.3 (Chernoff Bound)**.** Suppose the moment generating function of random variable X exists, and is finite for all $|t| \leq b, b > 0$. Let $\mu = \mathbb{E}[X]$, for any $t > 0$

$$\mathbb{P}((X - \mu) \geq u) \leq \inf_{0 \leq t \leq b} \frac{\mathbb{E}[e^{tX}]}{e^{(u+\mu)t}} \tag{C.4}$$

Proof. By Markov inequality, we have

$$\mathbb{P}((X - \mu) \geq u) = \mathbb{P}\left(e^{t(X-\mu)} \geq e^{tu}\right) \leq \frac{\mathbb{E}\left[e^{t(X-\mu)}\right]}{e^{tu}}$$

Since this bound is true for any t, we have

$$\mathbb{P}((X - \mu) \geq u) \leq \inf_{0 \leq t \leq b} \frac{\mathbb{E}[e^{tX}]}{e^{(u+\mu)t}}$$

□

Bounded Random Variables.

We are going to consider the case of bounded random variables and derive the so called Hoeffding's bound for them. As we know, the bounded random variables are the special case of sub-Gaussian variables.

Lemma C.4 (MGF of Rademacher Variables)**.** The Rademacher variable is the random variable $X \in$

$\{+1, -1\}$ with equally probability. The MGF of Rademacher variable satifies

$$\mathbb{E}[e^{tX}] \leq e^{t^2/2} \tag{C.5}$$

Proof. By definition,

$$\begin{aligned}\mathbb{E}[e^{tX}] = \frac{1}{2}(e^t + e^{-t}) &= \frac{1}{2}\left(\sum_{k=0}^{\infty}\frac{t^k}{k!} + \sum_{k=0}^{\infty}\frac{(-t)^k}{k!}\right) \\ &= \sum_{k=0}^{\infty}\frac{t^{2k}}{(2k)!} \leq \sum_{k=0}^{\infty}\frac{t^{2k}}{2^k k!} = e^{t^2/2}\end{aligned}$$

□

Lemma C.5 (Jensen's inequality)**.** A function g is convex if

$$g(\alpha x + (1-\alpha)y) \leq \alpha g(x) + (1-\alpha)g(y)$$

for all x, y and all $\alpha \in [0, 1]$; then for random variable X we have

$$g(\mathbb{E}[X]) \leq \mathbb{E}[g(X)]$$

Proof. Let $\mu = \mathbb{E}[X]$ and let $L_\mu(x) = a + bx$ be the tangent line for the function g at μ, then we have $L_\mu(\mu) = g(\mu)$. By convexity, we know $g(x) \geq L_\mu(x)$ for all x; thus we have

$$\mathbb{E}[g(X)] \geq \mathbb{E}[L_\mu(X)] = \mathbb{E}[a + bX] = a + b\mu = L_\mu(\mu) = g(\mu)$$

□

Lemma C.6 (MGF of Bounded Variables)**.** The bounded variables is the random variable X with zero mean and with support on some bounded interval $[a, b]$. The MGF of bounded variable is

$$\mathbb{E}_X[e^{tX}] \leq \exp\left(\frac{(b-a)^2 t^2}{2}\right) \tag{C.6}$$

which in turn show that bounded random variables are $(b - a)$ sub-Gaussian.

Proof. Let X be a random variable with zero mean and with support on some bounded interval $[a, b]$, and (note that one can always subtract the means and get a new rv)

$$Y = X - \mathbb{E}[X]$$

using Jensen's inequality and the convexity of $g(x) = e^x$, we have

$$\mathbb{E}_X[e^{tX}] = \mathbb{E}_X\left[e^{t(X - \mathbb{E}[X'])}\right] \leq \mathbb{E}_{X,X'}\left[e^{t(X-X')}\right]$$

now let ε be a Rademacher random variable, and note that the distribution

$$X - X' \overset{d}{=} X' - X \overset{d}{=} \varepsilon(X - X')$$

so we have

$$\mathbb{E}_{X,X'}\left[e^{t(X-X')}\right] = \mathbb{E}_{X,X'}\left[\mathbb{E}_{\varepsilon}[e^{\varepsilon t(X-X')}]\right] \leq \mathbb{E}_{X,X'}\left[e^{t^2(X-X')^2/2}\right] \leq e^{t^2(b-a)^2/2}$$

with the notice that X is bounded and $(X - X')$ is at most $(b - a)$. □

This in turn yields the simple version of Hoeffding's bound.

Gaussian Random Variables.

Corollary C.7 (Gaussian Tail Bound)**.** Suppose random variable $X \sim N(\mu, \sigma^2)$, the MGF of X is then $\mathbb{E}[e^{tX}] = e^{\mu t+\sigma^2 t^2/2}$. Applying the Chernoff bound, we have one-sided upper bound

$$\mathbb{P}(X - \mu \geq u) \leq \exp\left(-\frac{u^2}{2\sigma^2}\right) \tag{C.7}$$

and *lower tail bound*

$$\mathbb{P}(-X + \mu \geq u) \leq \exp\left(-\frac{u^2}{2\sigma^2}\right)$$

putting these together, we have the *two-sided Gaussian tail bound*:

$$\mathbb{P}(|X - \mu| \geq u) \leq 2\exp\left(-\frac{u^2}{2\sigma^2}\right)$$

Proof. Suppose $X \sim N(\mu, \sigma^2)$, then the MGF of X is

$$\begin{aligned} M_X(t) = \mathbb{E}\left[e^{tX}\right] &= \frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} e^{tx}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx \\ &= \frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} e^{-\frac{(x-(\mu+\sigma^2 t))^2}{2\sigma^2}}e^{\frac{\sigma^2 t^2}{2}+\mu t}dx \\ &= e^{\mu t+\sigma^2 t^2/2} \end{aligned}$$

or equivalently

$$\mathbb{E}[e^{t(X-\mu)}] = \exp\left(\frac{\sigma^2 t^2}{2}\right) \tag{C.8}$$

to apply the Chernoff bound we then need to compute

$$\inf_{t\geq 0}\ \frac{e^{\mu t+\sigma^2 t^2/2}}{e^{(u+\mu)t}} = \inf_{t\geq 0}\ e^{-ut+\sigma^2 t^2/2} = e^{-ut+\sigma^2 t^2/2}|_{t=u/\sigma^2} = e^{-\frac{u^2}{2\sigma^2}}$$

therefore, we obtain one-sided upper tail bound,

$$\mathbb{P}(X - \mu \geq u) \leq \exp\left(-\frac{u^2}{2\sigma^2}\right)$$

□

The Gaussian tail bound is much sharper than Chebyshev's inequality; consider the average of i.i.d. Gaussian random variables, $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ and we construct the estimate

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

where $\bar{X}_n \sim N(\mu, \sigma^2/n)$, in this case, the Gaussian tail bound is

$$\mathbb{P}\left(\left|\bar{X}_n - \mu\right| \geq t\frac{\sigma}{\sqrt{n}}\right) \leq 2\exp\left(-\frac{t^2}{2}\right)$$

with probability 0.99 ($t = \sqrt{2\ln(1/0.0005)} = 3.25$), that the average $\bar{X}_n$ is within $3.25\sigma/\sqrt{n}$. More generally, with probability at least $1 - \delta$

- Chebyshev tells us that
$$|\bar{X}_n - \mu| \leq \frac{\sigma}{\sqrt{n\sigma}}$$
- Chernoff tail bound tells us that
$$|\bar{X}_n - \mu| \leq \sigma\sqrt{\frac{2\ln(2/\delta)}{n}}$$

Sub-Gaussian Random Variables.

Corollary C.8 (Sub-Gaussian Tail Bound)**.** Formally, a random variable X with mean μ is called σ-sub-Gaussian if there exists a positive number σ such that

$$\mathbb{E}[e^{t(X-\mu)}] \leq \exp\left(\frac{\sigma^2 t^2}{2}\right) \tag{C.9}$$

Roughly, these are random variables whose tails decay faster than a Gaussian. Similar to Gaussian tail bound, here we can derive the two-sided sub-Gaussian tail bound

$$\mathbb{P}(|X - \mu| \geq u) \leq 2\exp\left(-\frac{u^2}{2\sigma^2}\right) \tag{C.10}$$

Now, suppose we have n i.i.d. σ sub-Gaussian random variables $X_1, X_2, \ldots, X_n$, again

$$\bar{X}_n = \frac{1}{n}\sum_i X_i$$

by independence we have

$$\mathbb{E}\left[e^{t(\bar{X}_n - \mu)}\right] = \mathbb{E}\left[e^{\frac{t}{n}\sum_i (X_i - \mu)}\right] = \prod_{i=1}^{n} \mathbb{E}[e^{\frac{t}{n}(X_i - \mu)}] \leq \prod_{i=1}^{n} e^{\frac{\sigma^2 t^2}{2n^2}} = \exp\left(\frac{\sigma^2 t^2}{2n}\right)$$

alternatively, $\bar{X}_n$ is $\sigma/\sqrt{n}$ sub-Gaussian, this yields the tail bound for the average of sub-Gaussian rvs:

$$\mathbb{P}\left(\left|\bar{X}_n - \mu\right| \geq k\frac{\sigma}{\sqrt{n}}\right) \leq 2\exp\left(-\frac{k^2}{2}\right)$$

Exponential Random Variables

Theorem C.9 (Exponential Tail Bound)**.** Suppose that we have $X_1, \ldots, X_n$ which are each $\sigma_1, \ldots, \sigma_n$ sub-Gaussian; they are not identically distributed, but using just *independence*, one can verify that the average $\bar{X}_n$ is σ sub-Gaussian where

$$\sigma = \frac{1}{n}\sqrt{\sum_{i=1}^{n} \sigma_i^2}$$

this yields the *exponential tail inequality*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_i)\right|\geq t\right)\leq \exp\left(-\frac{t^2}{2\sigma^2}\right) \tag{C.11}$$

note the these random variables still need to be independent.

C.4 Hoeffding's Inequality

Here we use the information of bounded variable (first-order info) to bound the MGF of random variable, then we utilize the methodds of Chernoff bound.

Theorem C.10 (Hoeffding's Inequality)**.** Suppose $X_1,\ldots,X_n$ are i.i.d bounded random variables, with $X_i\in[a,b]$, then the sample average, $\bar{X}_n=\frac{1}{n}\sum_i X_i$ has the bound

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{k=1}^{n}X_k-\mu\right|\geq t\right)\leq 2\exp\left(-\frac{2nt^2}{(b-a)^2}\right) \tag{C.12}$$

The Hoeffding's inequality tells us that, with probability at least $1-\delta$,

$$\left|\frac{1}{n}\sum_{k=1}^{n}X_k-\mu\right|\leq (b-a)\sqrt{\frac{\ln(2/\delta)}{2n}} \tag{C.13}$$

where $\hat{\mu}=\bar{X}_n$ is the sample-average estimator of mean μ.

Proof. Suppose the random variable X has mean μ and is bounded by $[a,b]$. The logarithmic moment generating function of X is then

$$\varphi(s)=\log\mathbb{E}\left[e^{s(X-\mu)}\right]$$

The usual derivatives of $\varphi(s)$ is then

$$\begin{aligned}
\varphi'(s)&=\frac{\mathbb{E}[(X-\mu)e^{s(X-\mu)}]}{\mathbb{E}[e^{s(X-\mu)}]}\\
\varphi''(s)&=\frac{\mathbb{E}[(X-\mu)^2e^{s(X-\mu)}]}{\mathbb{E}[e^{s(X-\mu)}]}-\left(\frac{\mathbb{E}[(X-\mu)e^{s(X-\mu)}]}{\mathbb{E}[e^{s(X-\mu)}]}\right)^2\\
&=\frac{\int_a^b(x-\mu)^2e^{s(x-\mu)}dP(x)}{\int_a^b e^{s(x-\mu)}dP(x)}-\left(\frac{\int_a^b(x-\mu)e^{s(x-\mu)}dP(x)}{\int_a^b e^{s(x-\mu)}dP(x)}\right)^2
\end{aligned}$$

where we assume $P(x)$ is the distribution of X. It is essential to notice that $\varphi''(s)$ is the varaince of some random variable $\tilde{X}\in[a,b]$ with the distribution proportional to $e^{s(x-\mu)}dP(x)$. Therefore, we can bound the variance of $\tilde{X}$ as

$$\mathrm{Var}(\tilde{X})=\inf_{\mu\in[a,b]}\mathbb{E}[\tilde{X}-\mu]^2\leq\mathbb{E}\left[\tilde{X}-\frac{a+b}{2}\right]^2\leq\frac{(b-a)^2}{4}$$

for any s almost surely. Since $\varphi(0)=0$ and $\varphi'(0)=0$, the Taylor's expansion with Lagrange remainder of $\varphi(s)$ at point $s=0$ satisfies

$$\varphi(s)=\varphi(0)+\frac{\varphi'(0)}{1!}s+\frac{\varphi''(\xi)}{2!}s^2\leq\frac{(b-a)^2}{8}s^2\quad\text{almost surely}$$

where $\xi \in [0, s]$. That means

$$\mathbb{E}\left[e^{s(X-\mu)}\right] \leq \exp\left(\frac{(b-a)^2}{8}s^2\right) \tag{C.14}$$

Now we have complete the key part of the proof. Next, recall the Markov's inequality for any non-negative random X and $\varepsilon > 0$,

$$\begin{aligned}
\mathbb{P}(\bar{X}_n - \mu \geq t) &= \mathbb{P}(e^{s(\bar{X}_n-\mu)} \geq e^{st}) \\
&\leq \inf_s \; e^{-st}\mathbb{E}\left[e^{s(\bar{X}_n-\mu)}\right] \\
&= \inf_s \; e^{-st}\prod_{i=1}^{n}\mathbb{E}\left[e^{\frac{s}{n}(X_i-\mu)}\right] \\
&\leq \inf_s \; e^{-st}\prod_{i=1}^{n}\exp\left(\frac{s^2(b-a)^2}{8n}\right) \\
&= \inf_s \; \exp\left(-st + \frac{s^2(b-a)^2}{8n}\right) \\
&= \exp\left(-\frac{2nt^2}{(b-a)^2}\right), \qquad s = \frac{4nt}{(b-a)^2}
\end{aligned}$$

Repeating this in the other direction we get

$$\mathbb{P}\big(|\bar{X}_n - \mu| \geq t\big) \leq 2\exp\left(-\frac{2nt^2}{(b-a)^2}\right)$$

□

C.5 Bernstein's Inequality

The Hoeffding's bound depended only on the bounds of the random variable but not explicityly on the variance. The bound $b - a$, provides a (possibly loss) upper bound on the standard deviation. One might at least hope that if the random variables were bounded, and additionally had small variance, we might be able to imporve Hoeffding's bound.

Theorem C.11 (Berstein's Inequality)**.** Suppose we have $X_1, \ldots, X_n$ which were i.i.d from a distribution with mean μ, bounded suppport $[a, b]$, with variance $\mathbb{E}(X - \mu)^2 = \sigma^2$, then

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{k=1}^{n}X_k - \mu\right| \geq t\right) \leq 2\exp\left(-\frac{nt^2}{2\sigma^2 + 2(b-a)t}\right) \tag{C.15}$$

The inequality implies that, with probability at least $1 - \delta$,

$$\left|\frac{1}{n}\sum_{k=1}^{n}X_k - \mu\right| \leq \sigma\sqrt{\frac{2\ln(1/\delta)}{n}} + \frac{2(b-a)\ln(1/\delta)}{3n}$$

Proof. Using the Taylor's expansion of the exponential, we can bound the moment generating function of X

by

$$\begin{aligned}
\mathbb{E}\left[e^{s(X-\mu)}\right] &= 1 + s\mathbb{E}(X-\mu) + \sum_{k=2}^{\infty}\frac{s^k}{k!}\mathbb{E}(X-\mu)^k = 1 + \sum_{k=2}^{\infty}\frac{s^k}{k!}\mathbb{E}(X-\mu)^k \\
&\leq 1 + \sum_{k=2}^{\infty}\frac{s^k}{k!}\mathbb{E}\left[|X-\mu|^{k-2}|X-\mu|^2\right] \\
&\leq 1 + \sum_{k=2}^{\infty}\frac{s^k}{k!}\mathbb{E}|X-\mu|^{k-2}\sigma^2 \qquad \text{(Cauchy-Schwartz inequality)} \\
&= 1 + \frac{\sigma^2}{c^2}\sum_{k=2}^{\infty}\frac{s^k}{k!}, c^k \qquad \text{let } c = \mathbb{E}|X-\mu| \leq (b-a) \\
&= 1 + \frac{\sigma^2}{c^2}(e^{sc} - 1 - sc) \\
&\leq \exp\left(\frac{\sigma^2}{c^2}(e^{sc} - 1 - sc)\right)
\end{aligned}$$

With $\sigma^2 = \text{Var}(X_i)$, we have

$$\begin{aligned}
\mathbb{P}\left(\bar{X}_n - \mu \geq t\right) &= \mathbb{P}\left(e^{s(\bar{X}_n-\mu)} \geq e^{st}\right) \qquad \text{(Markov's inequality)} \\
&\leq \inf_s \; e^{-st}\mathbb{E}\left[e^{s(\bar{X}_n-\mu)}\right] \\
&= \inf_s e^{-st}\prod_{i=1}^{n}\mathbb{E}\left[e^{\frac{s}{n}(X_i-\mu)}\right] \\
&\leq \inf_s \; e^{-st}\prod_{i=1}^{n}\exp\left(\frac{\sigma^2}{c^2}\left(e^{\frac{sc}{n}} - 1 - \frac{sc}{n}\right)\right) \\
&= \inf_s \; \exp\left(-st + \frac{\sigma^2}{c^2}\left(ne^{\frac{sc}{n}} - n - sc\right)\right) \\
&= \exp\left(\frac{nt}{c} - \frac{nt}{c}\ln\left(1 + \frac{tc}{\sigma^2}\right) - \frac{n\sigma^2}{c^2}\ln\left(1 + \frac{tc}{\sigma^2}\right)\right), \qquad s = \frac{n}{c}\ln\left(1 + \frac{tc}{\sigma^2}\right) \\
&= \exp\left(-\frac{n\sigma^2}{c^2}\Big((1+\alpha)\ln(1+\alpha) - \alpha\Big)\right), \qquad \alpha = \frac{tc}{\sigma^2} \leq \frac{t(b-a)}{\sigma^2}
\end{aligned}$$

With the knowing that

$$(1+\alpha)\ln(1+\alpha) \geq \frac{\alpha^2}{2 + 2\alpha/3}$$

we can get what we desired. □

C.6 McDiarmid's Inequality

So far we have focused on the concentration of averages. A natural question is whether other functions of i.i.d. random variables also show exponential concentration. It turns out that many other functions do concentrate sharply, and roughly the main property of the function that we need is that if we change the value of one random variable the function does not change dramatically.

Theorem C.12 (McDiarmid's Inequality)**.** Suppose we have i.i.d random variables $X_1, \ldots, X_n$ where each $X_i \in \mathbb{R}^n$. We have a Lipschitz function $f : \mathbb{R}^n \to \mathbb{R}$, that satisfies the property that:

$$|f(x_1, \ldots, x_{k-1}, x_k, x_{k+1}, \ldots, x_n) - f(x_1, \ldots, x_{k-1}, x_k', x_{k+1}, \ldots, x_n)| \leq L_k$$

for every $x, x' \in \mathbb{R}^n$. Then for any $t \geq 0$

$$\mathbb{P}\big(|f(X_1, \ldots, X_n) - \mathbb{E}f(X_1, \ldots, X_n)| \geq t\big) \leq 2\exp\left(-\frac{2t^2}{\sum_{k=1}^n L_k^2}\right) \tag{C.16}$$

Proof. The proof generalizes Hoeffding's inequality, which corresponds to $f(x) = \frac{1}{n}\sum_{i=1}^n x_i$. Now, we introduce the random variables V_k for $k = 1, \ldots, n$,

$$V_k = \mathbb{E}[f(X_1, \ldots, X_n) \mid X_1, \ldots, X_k] - \mathbb{E}[f(X_1, \ldots, X_n) \mid X_1, \ldots, X_{k-1}]$$

By the law of iterated expectation, we have

$$\mathbb{E}[V_k \mid X_1, \ldots, X_{k-1}] = \mathbb{E}[f(X_1, \ldots, X_n) \mid X_1, \ldots, X_{k-1}] - \mathbb{E}[f(X_1, \ldots, X_n) \mid X_1, \ldots, X_{k-1}] = 0$$

and in the mean time,

$$\sum_{k=1}^n V_k = f(X_1, \ldots, X_n) - \mathbb{E}f(X_1, \ldots, X_n)$$

Since $|V_k| \leq L_k$ almost surley, V_k is also an bounded variable. Using the exact method as in the proof of Hoeffding's inequality, we have

$$\mathbb{E}\left[e^{sV_k}\right] \leq \exp\left(\frac{L_k^2}{8}s^2\right)$$

Then

$$\begin{aligned}
\mathbb{P}\left(\sum_{k=1}^n V_k \geq t\right) &= \mathbb{P}\left(e^{s\sum_{k=1}^n V_k} \geq e^{st}\right) \\
&\leq \inf_s \; e^{-st}\mathbb{E}\left[e^{s\sum_{k=1}^n V_k}\right] \qquad \text{(Markov's inequality)} \\
&= \inf_s \; e^{-st}\prod_{k=1}^n \mathbb{E}\left[sV_k\right] \\
&\leq \inf_s \; e^{-st}\prod_{k=1}^n \exp\left(\frac{L_k^2}{8}s^2\right) \\
&= \inf_s \exp\left(-st + \frac{\sum_{k=1}^n L_k^2}{8}s^2\right) \\
&= \exp\left(-\frac{2t^2}{\sum_{k=1}^n L_k^2}\right), \qquad s = \frac{4t}{\sum_{k=1}^n L_k^2}
\end{aligned}$$

□

C.7 Expectation of the Maximum

Theorem C.13 (Expectation of the Maximum)**.** If $Z_1, \ldots, Z_n$ are (potentially dependent) random variables which are σ-sub-Gaussian, then

$$\mathbb{E}[\max\{Z_1 - \mathbb{E}[Z_1], \ldots, Z_n - \mathbb{E}[Z_n]\}] \leq \sqrt{2\sigma^2 \log n}$$

Proof. By using the Jensen's inequality for logarithm, which is concave, we have

$$\begin{aligned}
\mathbb{E}\Big[\max\big\{Z_1-\mathbb{E}[Z_1],\dots,Z_n-\mathbb{E}[Z_n]\big\}\Big] &\le \frac{1}{t}\log\mathbb{E}\left[e^{t\max\{Z_1-\mathbb{E}[Z_1],\dots,Z_n-\mathbb{E}[Z_n]\}}\right] \quad \text{(by Jensen's inequality)}\\
&= \frac{1}{t}\log\mathbb{E}\left[\max\left\{e^{t(Z_1-\mathbb{E}[Z_1])},\dots,e^{t(Z_n-\mathbb{E}[Z_n])}\right\}\right]\\
&\le \frac{1}{t}\log\mathbb{E}\left[e^{t(Z_1-\mathbb{E}[Z_1])}+\cdots+e^{t(Z_n-\mathbb{E}[Z_n])}\right] \quad \text{(bounding the max by the sum)}\\
&\le \frac{1}{t}\log\left(ne^{\sigma^2t^2/2}\right)=\frac{\log n}{t}+\sigma^2\frac{t}{2} \quad \text{(by sub-Gaussian property)}
\end{aligned}$$

Since such inequality is hold for any $t\in\mathbb{R}$, then we can minimize over t and get $t=\sigma^{-1}\sqrt{2\log n}$. Thus

$$\mathbb{E}\Big[\max\big\{Z_1-\mathbb{E}[Z_1],\dots,Z_n-\mathbb{E}[Z_n]\big\}\Big]\le\sigma\sqrt{2\log n}$$

□

D Concentration for Matrices

D.1 Matrix Analysis

D.1.1 Matrix Functions

Consider a function $f:\mathbb{R}\to\mathbb{R}$. We define a map on diagonal matrices by applying the function to each diagonal entry. We then extend f to a function on Hermitian matrices using the eigenvalue decomposition:

$$f(A):=Qf(\Lambda)Q^* \tag{D.1}$$

where $A=Q\Lambda Q^*$. The *spectral mapping theorem* states that each eigenvalue of $f(A)$ is equal to $f(\lambda)$ for some eigenvalue λ of A. This point is obvious from our definition.

Standard inequalities for real functions typically do not have parallel versions that hold for the semi-definite ordering. Nevertheless, there is one type of relation for real functions that always extends to the semi-definite setting.

$$f(a)\le g(a)\quad\forall\ a\in I\quad\Longrightarrow\quad f(A)\preceq g(A) \tag{D.2}$$

when the eigenvalues of A lie in I. We sometimes refer to this as the *transfer rule*.

D.1.2 Matrix Exponential

The exponential of an Hermitian matrix A can be defined by applying (D.1) with the function $f(x)=e^x$. Alternatively, we may use the power series expansion

$$\exp(A):=I+\sum_{p=1}^{\infty}\frac{A^p}{p!}$$

The exponential of an Hermitian matrix H is always **positive definite** from the spectral mapping theorem. Here is a sketch proof,

$$x^\top e^Hx=x^\top e^{H/2}e^{H/2}x=\left(e^{H/2}x\right)^\top\left(e^{H/2}x\right)=\left\|e^{H/2}x\right\|_2^2\ge0 \tag{D.3}$$

because of the eigenvalue decomposition of Hermitian matrix. On account of the transfer rule (D.2), the matrix exponential satisfies some simple semidefinite relations that we collect here. For each Hermitian matrix A, it holds that

$$\begin{aligned} I + A &\preceq e^A \\ \cosh(A) &\preceq e^{A^2/2} \end{aligned} \tag{D.4}$$

We often work with the trace of the matrix exponential, $\operatorname{tr}\exp : A \mapsto \operatorname{tr} e^A$. The trace exponential function is **convex**. It is also monotone with respect to the semi-definite order:

$$A \preceq H \quad \Rightarrow \quad \operatorname{tr} e^A \leq \operatorname{tr} e^H \tag{D.5}$$

The matrix exponential does not convert sums into products, but the trace exponential has a related property that serves as a limited substitue. The Golden-Thompson inequality states that

$$\operatorname{tr} e^{A+H} \leq \operatorname{tr}\left(e^A e^H\right) \quad \text{for all Hermitian } A, H \tag{D.6}$$

The obvious generalization of the bound (D.6) to three matrices is **false**. The pperator monotone functions and operator convex functions are depressingly rare. In particular, the matrix exponential does not belong to either.

D.1.3 Matrix Logarithm

We define the matrix logarithm as the functional inverse of the matrix exponential

$$\log e^A := A \quad \text{for all Hermitian } A$$

This formular determines the logarithm on the positive definite cone, which is adequate for our purposes.

The matrix logarithm interacts beautifully with the semidefinite order. Indeed, the logarithm is opeartor monotone:

$$0 \prec A \preceq H \quad \Longrightarrow \quad \log(A) \preceq \log(H) \tag{D.7}$$

The logarithm is also operator concave:

$$\alpha \log A + (1-\alpha) \log H \preceq \log(\alpha A + (1-\alpha) H) \tag{D.8}$$

for all PD A, H and $\alpha \in [0, 1]$.

D.1.4 Expectation and the Semidefinite Order

Since the expectation of a random matrix can be viewed as a convex combination and the PSD cone is convex, expectation preserves the semi-definite order:

$$X \preceq Y \text{ almost surely} \quad \Longrightarrow \quad \mathbb{E}[X] \preceq \mathbb{E}[Y] \tag{D.9}$$

Every operator convex function admits an operator Jensen's inequality. In particular, the matrix square is operator convex, which implies that

$$(\mathbb{E}X)^2 \preceq \mathbb{E}X^2 \tag{D.10}$$

D.1.5 Matrix Martingales

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a master probability space. Consider a filtration $\{\mathcal{F}_k\}$ contained in the master sigma algebra:

$$\mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}_\infty \subset \mathcal{F}$$

Given such a filtration, we define the conditional expectation $\mathbb{E}_k[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_k]$. A sequence of $\{X_k\}$ of random matrices is adapated to the filtration when each X_k is measurable with respect to $\mathcal{F}_k$. Loosely speaking, an adapted sequence is one where the present depends only upon the past.

An adapted sequence $\{Y_k\}$ of Hermitian matrices is called a matrix martingale when

$$\mathbb{E}_{k-1}[Y_k] = Y_{k-1} \quad \text{and} \quad \mathbb{E}||Y_k|| < \infty \quad \text{for} \quad k = 1, 2, 3, \dots$$

We obtain a scalar martingale if we track any fixed coordinate of a matrix martingale $\{Y_k\}$. Given a matrix martingale $\{Y_k\}$, we can construct the difference sequence

$$X_k := Y_k - Y_{k-1} \quad \text{for} \quad k = 1, 2, 3$$

Note that the difference sequence is conditionally zero mean, $\mathbb{E}_{k-1} X_k = 0$.

D.2 Tail Bounds via the Matrix Laplace Transform Method

D.2.1 Matrix Moments and Cumulants

Consider a random Hermitian matrix X that has momens of all orders. By analogy with the classical scalar definitions, we may construct matrix extensions of the moment-generating function (MGF) and the cumulant-generating function (CGF):

$$M_X(\theta) := \mathbb{E}e^{\theta X} \quad \text{and} \quad C_X(\theta) := \log\left(\mathbb{E}e^{\theta X}\right) \tag{D.11}$$

We admit the possibility that these expectations do not exist for all value of θ. The matrix MGF and CGF have formal power series expansions:

$$M_X(\theta) = I + \sum_{p=1}^{\infty} \frac{\theta^p}{p!} \cdot \mathbb{E}[X^p] \quad \text{and} \quad C_X(\theta) = \sum_{p=1}^{\infty} \frac{\theta^p}{p!} \cdot \Psi_p$$

The coefficients $\mathbb{E}[X^p]$ are caled matrix moments, and we refer to Ψ_p as a matrix cumulant. The first cumulant is the mean and the second cumulant is the variance

$$\Psi_1 = \mathbb{E}[X] \quad \text{and} \quad \Psi_2 = \mathbb{E}[X^2] - (\mathbb{E}X)^2$$

Higher-order cumulants are harder to write down and interpret.

D.2.2 Laplace Transform Method

Proposition D.1 (The Laplace Transform Method)**.** Let Y be a random Hermitian matrix. Then for all $t \in \mathbb{R}$,

$$\mathbb{P}\left(\lambda_{\max}(Y) \geq t\right) \leq \inf_{\theta > 0} \left\{ e^{-\theta t} \cdot \mathbb{E}\left[\operatorname{tr} e^{\theta Y}\right] \right\} \tag{D.12}$$

In words, we can control tail probabilities for the maximum eigenvalue of a random matrix by producing a bound for the trace of the matrix MGF.

Proof. Fix a positive number θ, we have the chain of relations

$$\mathbb{P}(\lambda_{\max} \geq t) = \mathbb{P}(\lambda_{\max}(\theta Y) \geq \theta t) = \mathbb{P}\left(e^{\lambda_{\max}(\theta Y)} \geq e^{\theta t}\right) \leq e^{-\theta t} \cdot \mathbb{E} e^{\lambda_{\max}(\theta Y)}$$

The first identity uses the homogeneity of the maximum eigenvalue map, and the second relies on the monotonicity of the scalar exponential function; the third relation is Markov's inequality. To bound the exponential, note that

$$e^{\lambda_{\max}(\theta Y)} = \lambda_{\max}\left(e^{\theta Y}\right) \leq \operatorname{tr} e^{\theta Y}$$

The identity is the spectral mapping theorem; the inequality holds because the exponential of an Hermitian matrix is positive definite and the maximum eigenvalue of a positive definitie matrix is dominated by the trace. Combine the latter two relations to reach

$$\mathbb{P}(\lambda_{\max}(Y) \geq t) \leq e^{-\theta t} \cdot \mathbb{E}\left[\operatorname{tr} e^{\theta Y}\right]$$

This inequality holds for any positive θ, wo we take an infimum to complete the proof. □

D.2.3 Failure of the Matrix MGF

In the scalar setting, the Laplace transform method is very effective for studying sums of independent random variables because the MGF decomposes. Consider an independent sequence $\{Z_k\}$ of real random variables. We see that the scalar MGF of the sum satisfies a multiplication rule

$$M_{\sum X_k}(\theta) = \mathbb{E} \exp\left(\sum_k \theta X_k\right) = \mathbb{E} \prod_k e^{\theta X_k} = \prod_k \mathbb{E} e^{\theta X_k} = \prod_k M_{X_k}(\theta) \tag{D.13}$$

This calculation relies on the fact that the scalar exponential function converts sums to products, a property th matrix exponential does not share. As a consequence, there is no immediate analog of (D.13) in the matrix setting

D.2.4 A Concave Trace Function

Theorem D.1 (Lieb)**.** Fix a Hermitian matrix H. The function

$$f : A \mapsto \operatorname{tr} \exp(H + \log A)$$

is **concave** on the positive definite cone.

We require a simple but powerful corollary of Lieb's theorem. This result describes how expectation interacts with the trace exponential.

Corollary D.2. Let H be a fixed Hermitian matrix, and let X be a random Hermitian matrix. Then

$$\mathbb{E}\left[\operatorname{tr} \exp(H + X)\right] \leq \operatorname{tr} \exp\left(H + \log\left(\mathbb{E} e^{X}\right)\right)$$

Proof. Define the random matrix $Y = e^X$, and calculate that

$$\mathbb{E}[\operatorname{tr}\exp(H+X)] = \mathbb{E}[\operatorname{tr}\exp(H+\log Y)] \leq \operatorname{tr}\exp(H+\log(\mathbb{E}Y)) = \operatorname{tr}\exp\left(H+\log\left(\mathbb{E}e^X\right)\right)$$

The first identity follows from the definition of the matrix logarithm because Y is always PD. Lieb's result, ensures that the trace function is concave in Y, so we may invoke Jensen's inequality to draw the expectation inside the logarithm. □

D.2.5 Subadditivity of the Matrix CGF

Although the multiplication rule (D.13) of MGF is a dead end in the matrix case, the scalar CGF has a related property that submits to generalization. For an independent family $\{X_k\}$ or real random variables, the scalar CGF is additive:

$$C_{\sum_K X_k}(\theta) = \log\mathbb{E}\exp\left(\sum_k \theta X_K\right) = \sum_k \log\mathbb{E}e^{\theta X_k} = \sum_k C_{X_k}(\theta) \tag{D.14}$$

where the second identity comes from (D.13) when take logarithms.

One key insight is that Corollary D.2 offers a completely satisfactory way to extend the addition rule (D.14) for scalar CGF's the matrix setting. We have the following result.

Lemma D.3 (Subadditiveity of Matrix CGF's)**.** Consider a finite sequence $\{X_k\}$ of independent, random, Hermitian matrices. Then

$$\mathbb{E}\left[\operatorname{tr}\exp\left(\sum_k \theta X_k\right)\right] \leq \operatorname{tr}\exp\left(\sum_k \log\mathbb{E}\left[e^{\theta X_k}\right]\right) \quad \forall\ \theta\in\mathbb{R} \tag{D.15}$$

Proof. It does not harm to assume $\theta = 1$. Let $\mathbb{E}_k$ denote the expectation, conditioned on $X_1,\dots,X_k$. Abbreviate

$$C_k := \log\mathbb{E}_{k-1}\left[e^{X_k}\right] = \log\mathbb{E}\left[e^{X_k}\right] = X_k$$

where the equality holds because the family $\{X_k\}$ is independent. We see that

$$\begin{aligned}
\mathbb{E}\operatorname{tr}\exp\left(\sum_{k=1}^n X_k\right) &= \mathbb{E}_0\cdots\mathbb{E}_{n-1}\operatorname{tr}\exp\left(\sum_{k=1}^{n-1} X_k + X_n\right)\\
&\leq \mathbb{E}_0\cdots\mathbb{E}_{n-2}\operatorname{tr}\exp\left(\sum_{k=1}^{n} X_k + \log\mathbb{E}_{n-1}\left[e^{X_n}\right]\right) \quad \text{by Corollary D.2}\\
&= \mathbb{E}_0\cdots\mathbb{E}_{n-2}\operatorname{tr}\exp\left(\sum_{k=1}^{n-2} + X_{n-1} + C_n\right)\\
&\leq \mathbb{E}_0\cdots\mathbb{E}_{n-3}\operatorname{tr}\exp\left(\sum_{k=1}^{n-2} X_k + C_{n-1} + C_n\right)\\
&\cdots\\
&\leq \operatorname{tr}\exp\left(\sum_{k=1}^{n} C_k\right)
\end{aligned}$$

□

To make the parallel with the addition rule (D.14) clearer, we can rewrite the conclusion of this lemma in the form

$$\operatorname{tr}\exp\left(C_{\sum_k X_K}(\theta)\right) \le \operatorname{tr}\exp\left(\sum_k C_{X_k}(\theta)\right) \tag{D.16}$$

by applyting the definition of the matrix CGF.

D.2.6 Tail Bounds of Independent Sums

This section contains abstract tail bounds for the sum of independent random matrices. Later, we will specialize these results to some specific situations. We begin with a very general inequality, which is the progenitor of other results.

Theorem D.4 (Master Tail Bound for Independence Sums)**.** Consider a finite sequence $\{X_k\}$ of independent, random, Hermitian matrices. For all $t \in \mathbb{R}$

$$\mathbb{P}\left(\lambda_{\max}\left(\sum_k X_k\right) \ge t\right) \le \inf_{\theta\ge 0}\left\{e^{-\theta t}\cdot \operatorname{tr}\exp\left(\sum_k \log \mathbb{E}\left[e^{\theta X_k}\right]\right)\right\} \tag{D.17}$$

Proof. From Laplace transform bound, for random Hermitian matrice Y, we have

$$\mathbb{P}(\lambda_{\max}(Y) \ge t) \le \inf_{\theta>0}\left\{e^{-\theta t}\cdot \mathbb{E}\left[\operatorname{tr}\exp(\theta Y)\right]\right\}$$

Notice that the sum of i.i.d random Hermitian matrices $\sum_k X_k$ is still a random Hermitian matrice, and hence we have

$$\mathbb{P}\left(\lambda_{\max}\left(\sum_k X_k\right) \ge t\right) \le \inf_{\theta>0}\left\{e^{-\theta t}\cdot \mathbb{E}\left[\operatorname{tr}\exp\left(\sum_k \theta X_k\right)\right]\right\}$$

By applying the subadditivity of matrix CGF in Lemma D.3, we have

$$\mathbb{P}\left(\lambda_{\max}\left(\sum_k X_k\right) \ge t\right) \le \inf_{\theta>0}\left\{e^{-\theta t}\cdot \operatorname{tr}\exp\left(\sum_k \log \mathbb{E}\left[e^{\theta X_k}\right]\right)\right\}$$

□

Corollary D.5. Consider a finite sequence $\{X_k\}$ of independent, random, self-adjoint matrices with dimension d. Assume there is a function $g : (0,\infty) \to [0,\infty]$ and a sequence $\{A_k\}$ of fixed Hermitian matrices that satisfy the relations

$$\mathbb{E}\left[e^{\theta X_k}\right] \preceq e^{g(\theta)\cdot A_k} \tag{D.18}$$

for $\theta > 0$. Define the scale parameter $\rho := \lambda_{\max}(\sum_k A_k)$. Then for all $t \in \mathbb{R}$,

$$\mathbb{P}\left(\lambda_{\max}\left(\sum_k X_k\right) \ge t\right) \le d\cdot \inf_{\theta>0}\left\{e^{-\theta t + g(\theta)\cdot\rho}\right\} \tag{D.19}$$

Proof. The hypothesis implies that

$$\log \mathbb{E}\left[e^{\theta X_k}\right] \preceq g(\theta)\cdot A_k$$

for $\theta > 0$ becasue of the property that the matrix logarithm is operator monotone. Recall that the trace

exponential is monotone with respect to the semidefinite order, i.e.

$$A \preceq H \quad \Rightarrow \quad \operatorname{tr} e^A \preceq \operatorname{tr} e^H$$

As a consequence, we can introduce such relation into the master inequality (D.17), that is, for each $\theta > 0$

$$\begin{aligned}
\mathbb{P}\left(\lambda_{\max}\left(\sum_k X_k\right) \geq t\right) &\leq e^{-\theta t} \cdot \operatorname{tr} \exp\left(\sum_k \log \mathbb{E}\left[e^{\theta X_k}\right]\right) \\
&\leq e^{-\theta t} \cdot \operatorname{tr} \exp\left(g(\theta) \sum_k A_k\right) \\
&\leq e^{-\theta t} \cdot d \cdot \lambda_{\max}\left(\exp\left(g(\theta) \sum_k A_k\right)\right) \\
&= d \cdot e^{-\theta t} \cdot \exp\left(g(\theta) \cdot \lambda_{\max}\left(\sum_k A_k\right)\right)
\end{aligned}$$

The third inequality holds because the trace of a PD matrix, such as the exponential of matrix), is bounded by the dimension d times the maximum eigenvalue. The last line depends on the spectral mapping theorem and the fact that the function g is nonnegative. Identify the quantity $\rho := \lambda_{\max}(\sum_k A_k)$, and take the infimum over positive θ to reach the conclusion. □

Remark D.1 (Minimum Eigenvalue)**.** We can study the minimum eigenvalue of a sum of random Hermitian matrices because $\lambda_{\min}(X) = -\lambda_{\max}(-X)$. As a result

$$\mathbb{P}\left(\lambda_{\min}\left(\sum_k X_k\right) \leq t\right) = \mathbb{P}\left(\lambda_{\max}\left(\sum_k -X_k\right) \geq -t\right)$$

Remark D.2 (Maximum Singular Value)**.** We can also analyze the maximum singular value of a sum of random rectangular matrices B by applying these results to the Hermitian dilation, that is

$$\varphi(B) := \begin{bmatrix} 0 & B \\ B^* & 0 \end{bmatrix}$$

For a finite sequence $\{Z_k\}$ of independent, random, rectangular matrices, we have

$$\mathbb{P}\left(\left\|\sum_k Z_k\right\| \geq t\right) = \mathbb{P}\left(\lambda_{\max}\left(\sum_k \varphi(Z_k)\right) \geq t\right)$$

and the property that dilation is real-linear. This device allows us to extend most of the tail bounds in this paper to rectangular matrices.

D.3 Matrix Gaussian and Rademacher

We begin with the scalar case. Consider a finite sequence $\{a_k\}$ of real numbers and a finite sequence $\{\gamma_k\}$ of independent standard Gaussian variables. We have the probability inequality

$$\mathbb{P}\left(\sum_k a_k \gamma_k \geq t\right) \leq \exp\left(-\frac{t^2}{2\sigma^2}\right) \tag{D.20}$$

where $\sigma^2 := \sum_k a_k^2$. This result testifies that a Gaussian series with real coefficients satisfies a normal-type tail bound where the variance is controlled by the sum of the squared coefficients. The relation follows easily from the scalar Laplace transform method.

Lemma D.6 (Rademacher and Gaussian MGF's)**.** Suppose that A is an Hermitian matrix. Let σ be a Rademacher random variable, and let γ be a standard normal random variable. Then

$$\mathbb{E}\left[e^{\sigma\theta A}\right] \preceq e^{\theta^2 A^2/2} \quad \text{and} \quad \mathbb{E}\left[e^{\gamma\theta A}\right] = \mathbb{E}e^{\theta^2 A^2/2} \quad \theta \in \mathbb{R} \tag{D.21}$$

Proof of Lemma D.6. Absorbing θ into A, we may assume $\theta = 1$ in each case. We begin with the Rademacher MGF. By direct calculation,

$$\mathbb{E}e^{\varepsilon A} = \cosh(A) \preceq e^{A^2/2}$$

where the second realtion is (D.4). For the Gaussian case, recall that the moments of a standard normal variable satisfy

$$\mathbb{E}[\gamma^{2p+1}] = 0 \quad \text{and} \quad \mathbb{E}[\gamma^{2p}] = \frac{(2p)!}{p!2^p} \quad p = 0, 1, 2 \dots$$

Therefore,

$$\mathbb{E}e^{\gamma A} = I + \sum_{p=1}^{\infty} \frac{\mathbb{E}[\gamma^{2p}]A^{2p}}{(2p)!} = I + \sum_{p=1}^{\infty} \frac{(A^2/2)^p}{p!} = e^{A^2/2}$$

□

Theorem D.7 (Matrix Gaussian and Rademacher Series)**.** Consider a finite sequence $\{A_k\}$ of fixed (non-random) Hermitian matrices with dimension dm, and let $\{\gamma_k\}$ be a finite sequence of independent standard normal variables. Compute the variance parameter $\sigma^2 := ||\sum_k A_k^2||_2$. Then, for all $t \geq 0$

$$\mathbb{P}\left(\lambda_{\max}\left(\sum_k \gamma_k A_k\right) \geq t\right) \leq d \cdot \exp\left(-\frac{t^2}{2\sigma^2}\right) \tag{D.22}$$

In particular,

$$\mathbb{P}\left(\left\|\sum_k \gamma_k A_k\right\|_2 \geq t\right) \leq 2d \cdot \exp\left(-\frac{t^2}{2\sigma^2}\right) \tag{D.23}$$

The same bounds hold when we replace $\{\gamma_k\}$ by a finite sequence of independent Rademacher random variables.

Proof of Theorem D.7. Let ξ_k be a finite sequence of independent standard normal variables or independent Rademacher variables. Invoke Lemma (D.6) to obtain

$$\mathbb{E}e^{\xi_k \theta A_k} \preceq e^{g(\theta)A_k^2}$$

where $g(\theta) := \theta^2/2$ for $\theta > 0$. Recall that

$$\sigma^2 = \left\|\sum_k A_k^2\right\| = \lambda_{\max}\left(\sum_k A_k^2\right)$$

Corollary D.5 delivers

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum_k \xi_k A_k\right) \geq t\right\} \leq d \cdot \inf_{\theta>0} e^{-\theta t + g(\theta)\cdot\sigma^2} = d \cdot e^{-t^2/2\sigma^2} \tag{D.24}$$

where the infimum is attained when $\theta = t/\sigma^2$.

To obtain the norm bound (D.23), recall that $||Y||_2 = \max(\lambda_{\max}(Y), -\lambda_{\min}(Y))$. Standard Gaussian variables and Rademacher variables are symmetric, so the inequality above implies

$$\mathbb{P}\left\{-\lambda_{\min}\left(\sum_k \xi_k A_k\right) \geq t\right\} = \mathbb{P}\left\{\lambda_{\max}\left(\sum_k (-\xi_k) A_k\right) \geq t\right\} \leq d \cdot e^{-t^2/2\sigma^2}$$

Apply the union bound we have

$$\mathbb{P}\left\{\left\|\sum_k \gamma_k A_k\right\|_2 \geq t\right\} \leq \mathbb{P}\left\{-\lambda_{\min}\left(\sum_k \xi_k A_k\right) \geq t\right\} + \mathbb{P}\left\{\lambda_{\max}\left(\sum_k (-\xi_k) A_k\right) \geq t\right\} \leq 2d \cdot e^{-t^2/2\sigma^2}$$

□

D.4 Matrix Bennett and Bernstein Bounds

In the scalar setting, Bennett and Bernstein inequalities describe the upper tail of a sum of independent, zero-mean random variables that are either bounded or subexponential. In the matrix case, the analogous results concern a sum of zero-mean random matrices.

Lemma D.8 (Bounded Bernstein MGF)**.** Suppose that X is a random Hermitian matrix that satisfies

$$\mathbb{E}[X] = 0 \quad \text{and} \quad \lambda_{\max}(X) \leq 1 \quad \text{a.s.}$$

Then we have, for $\theta > 0$,

$$\mathbb{E}e^{\theta X} \preceq e^{\left(e^\theta - \theta - 1\right)\cdot \mathbb{E}[X^2]}$$

Proof of Lemma D.8. Fix the parameter $\theta > 0$, and define a smooth function f on the real line:

$$f(x) = \frac{e^{\theta x} - \theta x - 1}{x^2}$$

for $x \neq 0$ and $f(0) = \theta^2/2$. An exercise in differential calculus verifies that f is increasing. Therefore, $f(x) \leq f(1)$ when $x \leq 1$. The eigenvalues of X do not exceed one, so the transfer rule (D.2) implies that

$$f(X) \preceq f(I) = f(1) \cdot I$$

Expanding the matrix exponential and applying the latter relation, we discover that

$$e^{\theta X} = I + \theta X + X \cdot f(X) \cdot X$$

To complete the proof, we take the expectation of this semidefinite bound:

$$\mathbb{E}e^{\theta X} \preceq I + f(1) \cdot \mathbb{E}[X^2] \preceq \exp\left(f(1) \cdot \mathbb{E}[X^2]\right) = \exp\left(\left(e^\theta - \theta - 1\right) \cdot \mathbb{E}[X^2]\right)$$

The second semidefinite relation follows from (D.4). □

Theorem D.9 (Matrix Bernstein - Bounded Case)**.** Consider a finite sequence $\{X_k\}$ of independent, random, Hermitian matrices with dimension d. Assume that

$$\mathbb{E}[X_k] = 0 \quad \text{and} \quad \lambda_{\max}(X_k) \le R \quad \text{a.s.}$$

Compute the norm of the total variance,

$$\sigma^2 := \left\| \sum_k \mathbb{E}[X_k^2] \right\|_2$$

Then the following chain of inequalities holds for all $t \ge 0$:

$$\begin{aligned} \mathbb{P}\left\{ \lambda_{\max}\left(\sum_k X_k \right) \ge t \right\} &\le d \cdot \exp\left(-\frac{\sigma^2}{R^2} \cdot h\left(\frac{Rt}{\sigma^2} \right) \right) \quad \text{(Bennett inequality)} \\ &\le d \cdot \exp\left(-\frac{t^2/2}{\sigma^2 + Rt/3} \right) \quad \text{(Bernstein inequality)} \\ &\le \begin{cases} d \cdot \exp(-3t^2/8\sigma^2), & \text{for } t \le \sigma^2/R \\ d \cdot \exp(-3t/8R), & \text{for } t \ge \sigma^2/R \end{cases} \quad \text{(split Bernstein inequality)} \end{aligned} \tag{D.25}$$

The function $h(u) := (1+u)\log(1+u) - u$ for $u \ge 0$.

Proof of Theorem D.9. We assume that $R = 1$; the general result follows by a scaling argument once we note that the summands are 1-homogeneous and the variance σ^2 is 2-homogeneous.

The main challenge is to establish the Bennett inequality, part (i); the ramining bounds are consequences of simple numerical estimates. Invoke Lemma D.8 to see that

$$\mathbb{E}[e^{\theta X_k}] \preceq e^{g(\theta)\cdot \mathbb{E}[X_k^2]}$$

where $g(\theta) := e^\theta - \theta - 1$ for $\theta > 0$. For each $\theta > 0$, Corollary D.5 implies that

$$\begin{aligned} \mathbb{P}\left\{ \lambda_{\max}\left(\sum_k X_k \right) \ge t \right\} &\le d \cdot \exp\left(-\theta t + g(\theta) \cdot \lambda_{\max}\left(\sum_k \mathbb{E}[X_k^2] \right) \right) \\ &= d \cdot \exp\left(-\theta + g(\theta) \cdot \sigma^2 \right) \end{aligned}$$

The right-hand side attains its minimal value when $\theta = \log(1 + t/\sigma^2)$. Substitute and simplify this value yields

$$-\theta t + g(\theta \cdot \sigma^2) = \sigma^2 \left(\frac{t}{\sigma^2} - \left(1 + \frac{t}{\sigma^2} \right) \log\left(1 + \frac{t}{\sigma^2} \right) \right) = -\sigma^2 h\left(\frac{t}{\sigma^2} \right)$$

where $h(u) := (1+u)\log(1+u) - u$ for $u \ge 0$, which leads to the result in establish part (i)

$$\mathbb{P}\left\{ \lambda_{\max}\left(\sum_k X_k \right) \ge t \right\} \le d \cdot \exp\left(-\sigma^2 h\left(\frac{t}{\sigma^2} \right) \right)$$

The Bennett inequality (i) implies the Bernstein inequality (ii) because of the numerical bound

$$h(u) \geq \frac{u^2/2}{1+u/3} \quad \text{for} \quad u \geq 0$$

The latter relation is established by comparing derivatives. The Bernstein inequality (ii) implies the split Bernstein inequality (iii). To obtain the sub-Gaussian piece of (iii), observe that

$$\frac{1}{\sigma^2 + Rt/3} \geq \frac{1}{\sigma^2 + R(\sigma^2/R)/3} = \frac{3}{4\sigma^2} \quad \text{for} \quad t \leq \sigma^2/R$$

becasue the left-hand side is a decreasing function of t for $t \geq 0$. Similarily, we obatin the subexponential piece of (iii) from the fact that

$$\frac{t}{\sigma^2 + Rt/3} \geq \frac{\sigma^2/R}{\sigma^2 + R(\sigma^2/R)3} = \frac{3}{4R} \quad \text{and} \quad t \geq \sigma^2/R$$

which holds because the left-hand side is an increasing function of t for $t \geq 0$. □

Theorem D.10 (Matrix Bernstein - Subexponential Case)**.** Consider a finite sequence $\{X_k\}$ of independent, random, Hermitian matrices with dimension d. Assume that

$$\mathbb{E}[X_k] = 0 \quad \text{and} \quad \mathbb{E}[X_k^p] \preceq \frac{p!}{2} \cdot R^{p-2} A_k^2$$

for $p = 2, 3, 4, \ldots$. Coompute the variance parameter $\sigma^2 := ||\sum_k A_k^2||_2$. Then the following chain of inequalities holds for all $t \geq 0$:

$$\begin{aligned} \mathbb{P}\left(\lambda_{\max}\left(\sum_k X_k\right) \geq t\right) &\leq d \cdot \exp\left(-\frac{t^2/2}{\sigma^2 + Rt}\right) \\ &\leq \begin{cases} d \cdot \exp(-t^2/4\sigma^2) & \text{for } t \leq \sigma^2/R \\ d \cdot \exp(-t/4R) & \text{for } t \geq \sigma^2/R \end{cases} \end{aligned} \tag{D.26}$$

The hypotheses of Theorem D.10 are not fully comparable with the hypotheses of Theorem D.9, because Theorem D.10 allows the random matrices to be unbounded but it also demands that we control the fluctuation of the maximum and minimum eigenvalues.

D.5 Matrix Hoeffding and Azuma and McDiarmid

The scalar version of Azuma's inequality states that a scalar martingale exhibits normal concentration about its mean value, and the scale for deviations is controlled by the total maximum squared range of the difference sequence. Here is a matrix extension.

Lemma D.11 (Symmetrization)**.** Let H be a fixed Hermitian matrix, and let X bea random Hermitian matrix with $\mathbb{E}[X] = 0$. Then

$$\mathbb{E}\left[\operatorname{tr} e^{H+X}\right] \leq \mathbb{E}\left[\operatorname{tr} e^{H+2\sigma X}\right] \tag{D.27}$$

where σ is a Rademacher variable independent from X

Proof. Consturct an independent copy X' of the random matrix, and let $\mathbb{E}'$ denote integration with respect to the new variable. Since the matrix is zero mean,

$$\mathbb{E}\left[\operatorname{tr} e^{H+X}\right] = \mathbb{E}\left[\operatorname{tr}\ e^{H+X-\mathbb{E}'[X']}\right] \le \mathbb{E}\left[\operatorname{tr}\ e^{H+(X-X')}\right] = \mathbb{E}\left[\operatorname{tr}\ e^{H+\sigma(X-X')}\right]$$

We have use the convexity of the trace exponential (both $\operatorname{tr}\exp(A)$ and $\operatorname{tr}\exp(-A)$ are convex) to justify the Jensen's inequality. Since $X - X'$ is a symmetric random matrix, we can modulate it by an independent Rademacher variable σ without changing its distribution. The final bound depends on a short sequence of inequalities:

$$\begin{aligned}
\mathbb{E}\left[\operatorname{tr} e^{H+X}\right] &\le \mathbb{E}\operatorname{tr}\left(e^{H/2+\sigma X}\cdot e^{H/2-\sigma X'}\right) \quad \text{(Golden–Thompson (D.6))} \\
&\le \mathbb{E}\left[\left(\operatorname{tr}\ e^{H+2\sigma X}\right)^{1/2}\cdot\left(\operatorname{tr}\ e^{H-2\sigma X}\right)^{1/2}\right] \quad \text{(Cauchy-Schwarz for the trace)} \\
&\le \left(\mathbb{E}\left[\operatorname{tr}\ e^{H+2\sigma X}\right]\right)^{1/2}\cdot\left(\mathbb{E}\left[\operatorname{tr}\ e^{H+2\sigma X}\right]\right)^{1/2} \\
&= \mathbb{E}\left[\operatorname{tr}\ e^{H+2\sigma X}\right]
\end{aligned}$$

□

Lemma D.12 (Azuma CGF)**.** Suppose that X is a random Hermitian matrix and A is fixed Hermitian matrix satisfy $X^2 \preceq A^2$. Let σ be a Rademacher random variable independent from X. Then

$$\log \mathbb{E}\left[e^{2\sigma\theta X} \mid X\right] \preceq 2\theta^2 A^2 \quad \text{for} \quad \theta \in \mathbb{R}$$

Proof. We apply the Rademacher MGF bound, Lemma D.6, conditionally to obtain

$$\mathbb{E}\left[e^{2\theta\sigma X} \mid X\right] \preceq e^{2\theta^2 X^2}$$

The fact that the logarithm is operator monotone implies that

$$\log \mathbb{E}\left[e^{2\theta\sigma X} \mid X\right] \preceq 2\theta^2 X^2 \preceq 2\theta^2 A^2$$

where the second realtion follows from the hypothesis on X. □

Theorem D.13 (Matrix Azuma)**.** Consider a finite adpated sequence $\{X_k\}$ of Hermitian matrices in dimension d, and a fixed sequence $\{A_k\}$ of Hermitian matrices that satisfy

$$\mathbb{E}_{k-1}[X_k] = 0 \quad \text{and} \quad X_k^2 \preceq A_k^2 \quad \text{a.e.}$$

Compute the variance parameter

$$\sigma^2 := \left\|\sum_k A_k^2\right\|_2 = \lambda_{\max}\left(\sum_k A_k^2\right)$$

Then, for all $t \ge 0$,

$$\mathbb{P}\left(\lambda_{\max}\left(\sum_k X_k\right) \ge t\right) \le d\cdot e^{-t^2/8\sigma^2} \tag{D.28}$$

Proof. The matrix Laplace transform method, Proposition D.1, states that

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum_k X_k\right) \geq t\right\} \leq \inf_{\theta>0}\left\{e^{-\theta t}\cdot \mathbb{E}\left[\operatorname{tr}\exp\left(\sum_k \theta X_k\right)\right]\right\}$$

The main difficulty in the proof is to bound the matrix MGF, which we accomplish by an iterative argument that alternates between symmetrization and cumulant bounds.

Let us detail the first step of the iteration. Define the natural filtration $\mathcal{F}_k := \mathcal{F}(X_1, \dots, X_k)$ of the process $\{X_k\}$. Then we may compute

$$\begin{aligned}
\mathbb{E}\left[\operatorname{tr}\exp\left(\sum_k \theta X_k\right)\right] &= \mathbb{E}\left[\mathbb{E}\left[\operatorname{tr}\exp\left(\sum_{k=1}^{n-1}\theta X_k + \theta X_n\right)\Big|\mathcal{F}_{n-1}\right]\right] \quad \text{(iterated law)}\\
&\leq \mathbb{E}\left[\mathbb{E}\left[\operatorname{tr}\exp\left(\sum_{k=1}^{n}\theta X_k + 2\sigma\theta X_n\right)\Big|\mathcal{F}_{n}\right]\right] \quad \text{(symmetrization)}\\
&\leq \mathbb{E}\left[\operatorname{tr}\exp\left(\sum_{k=1}^{n-1}\theta X_k + \log\mathbb{E}[e^{2\sigma\theta X_n}\mid\mathcal{F}_n]\right)\right] \text{(concavity of trace exponential)}\\
&\leq \mathbb{E}\left[\operatorname{tr}\exp\left(\sum_{k=1}^{n-1}\theta X_k + 2\theta^2 A_n^2\right)\right] \quad \text{(Azuma CGF)}
\end{aligned}$$

- the first identity is the tower property of the conditional expectation
- in the second line, we winvoke the symmetrization method, Lemma D.11, conditional on $\mathcal{F}_{n-1}$, and then we relax the conditioning on the inner expectation to the larger algebra $\mathcal{F}_n$
- by construction, the Rademacher variable σ is independent from $\mathcal{F}_n$, so we can appy the concavity result, Corollary D.2, conditional on $\mathcal{F}_n$
- finally we use the fact (D.5) that trace exponential is monotone to introduce the Azuma CGF bound, Lemma D.12, in the last inequality

By iteration, we achieve

$$\mathbb{E}\left[\operatorname{tr}\exp\left(\sum_k \theta X_k\right)\right] \leq \operatorname{tr}\exp\left(2\theta^2\sum_k A_k^2\right)$$

Note that this procedure relies on the fact that the sequence $\{A_k\}$ of upper bounds does not depend on the values of the random sequence $\{X_k\}$. Substitute the MGF bound into the Laplace transform bound above, and observe that the infimum is achieved when $\theta = t/4\sigma^2$, we have

$$\begin{aligned}
\mathbb{P}\left\{\lambda_{\max}\left(\sum_k X_k\right) \geq t\right\} &\leq \inf_{\theta>0}\left\{e^{-\theta t}\cdot\mathbb{E}\left[\operatorname{tr}\exp\left(\sum_k\theta X_k\right)\right]\right\}\\
&\leq \inf_{\theta>0}\left\{e^{-\theta t}\cdot\operatorname{tr}\exp\left(2\theta^2\sum_k A_k^2\right)\right\}\\
&\leq \inf_{\theta>0}\left\{e^{-\theta t}\cdot d\cdot\lambda_{\max}\left(\exp\left(2\theta^2\sum_k A_k^2\right)\right)\right\}\\
&= d\cdot\inf_{\theta>0}\left\{e^{-\theta t}e^{2\theta^2\sigma^2}\right\}\\
&= d\cdot e^{-\frac{t^2}{8\sigma^2}}
\end{aligned}$$

□