A Minimal Note on Statistical Foundations of Modern Machine Learning

Thanh Nguyen-Tang

February 27, 2021

Contents

Forewords

This is my study note when I am taking this course: Algorithmic Foundations of Learning by Prof. Patrick Rebeschini, University of Oxford, Michaelmas Term 2019.

With the goal of keeping in place important results for my own reference and of practicing derivation skills for generalization bounds, in this note, I

1. rearrange important concepts and results taught in the course (in the same order as the course materials);
2. derive proofs for important results where I have tried my best to prove them myself first before referring to the lecture's proofs (when I failed to prove).

1 Preliminary

1.1 Matrices

1.1.1 Eigenvalues and eigenvectors

Remark. We can compute eigenvectors using only the information from eigenvalues https://arxiv.org/pdf/1908.03795.pdf.

Theorem 1. *The eigenvectors $x_1, x_2, ..., x_n$ of a square matrix A associated with distinct eigenvalues $\lambda_1, \lambda_2, ..., \lambda_n$ forms a basis of $\mathbb{R}^n$.*

Theorem 2 (Eigendecomposition). *Let $A \in \mathbb{R}^{n\times n}$ with n linearly independent eigenvectors $(q_i)_{i=1}^n$ associated with eigenvalues $(\lambda_i)_{i=1}^n$, and let $Q = [q_1, ..., q_n] \in \mathbb{R}^{n\times n}$ and $\Lambda = diag(\lambda_1, ..., \lambda_n)$. Then, A can be decomposed as*

$$A = Q\Lambda Q^{-1}.$$

In this case, A is called a diagonalizable matrix.

Remark 1. We can normalize eigenvectors to form an orthornomal basic. In this case, $Q^{-1} = Q^T$.
Remark 2. Any real symmetric matrix is diagonalizable.

1.1.2 Singular value decomposition (SVD)

Theorem 3 (SVD theorem). *Let $A \in \mathbb{R}^{m\times n}$ of any rank $r \in [0, m \wedge n]$, then A can be always be composed into*

$$A = U\Sigma V^T,$$

where $U = [u_1, ..., u_m] \in \mathbb{R}^{m\times m}$ and $V = [v_1, ..., v_n] \in \mathbb{R}^{n\times n}$ are orthogonal matrices, and $\Sigma \in \mathbb{R}^{m\times n}$ with $\Sigma_{ii} = \sigma_i \geq 0$ and $\Sigma_{i,j} = 0, \forall i \neq j$.

Remark 1. σ_i are call singular values of A, u_i are called left-singular vectors and v_j are called right-singular vectors.
Remark 2. The singular matrix Σ is unique for each matrix A.

1.1.3 Matrix norm

Definition 1 (**Matrix norm induced by vector norm**).

$$\|A\|_p := \sup_{x\neq 0} \frac{\|Ax\|_p}{\|x\|_p}, \forall p \in [1, \infty].$$

We are interested in the special case where $p = 2$. In this case, the matrix norm is called *spectral norm.* A spectral norm can be alternatively characterized as

$$\|A\|_2 := \sup\{x^T Ay : \|x\|_2 = \|y\|_2 = 1\},$$

and it has a closed form

$$\|A\|_2 = \sqrt{\lambda_{\max}(A^* A)} = \sigma_{\max}(A).$$

Remark. Since A and A^T have the same singular values, $\|A\|_2 = \|A^T\|_2$.

Lemma 1. *Let A be a matrix with rank k. We have*

$$\|A\|_2 \leq \|A\|_F := \sqrt{trace(A^* A)} \leq \sqrt{k}\|A\|_2.$$

Lemma 2. *Let $x \in \mathbb{R}^d$ such that $x^T x \leq a$ for a given $a > 0$. Then, we have*

$$xx^T \preceq aI.$$

Proof. Note that $xx^T \preceq aI$ iff $\lambda_{\max}(xx^T) \leq a$, but we have

$$a \geq x^T x = \|x^T\|_2^2 = \lambda_{\max}(xx^T).$$

□

Lemma 3. *For any matrix $A \in \mathbb{R}^{m\times n}$, we have*

$$\|A\|_2 \geq \frac{x^T Ax}{\|x\|_2}, \forall x \in \mathbb{R}^n/\{0\}.$$

Proof. We use an alternative definition of matrix norm (https://en.wikipedia.org/wiki/Matrix_norm) instead of the original definition where

This alternative definition immediately yields the inequality to be proved. □

A positive-definite (p.d.) symmetric matrix A induces an inner product:

$$\langle x, y\rangle_A := x^T Ay.$$

Remark 1. In practice, we often encounter $A = \sum_{i=1}^n u_i u_i^T$ where $u_i \in \mathbb{R}^d$. This is Gram matrix and tt is easy to check that such A is a symmetric p.d. matrix .
Remark 2: It follows from the Cauchy-Schwartz inequality that $\langle x, y\rangle_A \leq \sqrt{\langle x, x\rangle_A \langle y, y\rangle_A}$.

Lemma 4. *For square matrices M and A where A is invertible, and for any $n \in \mathbb{N}$, we have*

$$M^n = A^{-1}(AMA^{-1})^n A.$$

1.2 Smoothness notions for nonparametric regression

In regression, we need regularity assumptions and whenever we state them, we need to analyze how mild they are. Boundedness and smoothness are two such common assumptions (for the regression function).

Continuity. Absolutely continuous $\subseteq$ uniformly continuous $\subseteq$ continuous.

Strong derivative. A real-valued function is differentiable at x_0 if its derivative at x_0 exists. **Notations**. C^0: the class of all continuous functions; C^1: The class of all functions whose derivative exists and in C^0. Similarly defined C^k.

Weak derivatives. Differentiability is a strong notion of smoothness. We often work with weaker notions of smoothness that broaden the application. Now, instead of working with differentiable functions, we only require the functions to be integral. Recall that Lebesgue space $L^p(\mathbb{R})$ is defined as follows. Consider a measure space $(\Omega, \mathcal{F}, \mu)$ where μ is Lebesgue measure and let $p \in [1, \infty)$. $L^p(\mathbb{R})$ is the space of functions f such that

$$\|f\|_p := (\int |f|^p d\mu)^{1/p} < \infty$$

Now we define a notion of weak derivatives. Let $U \subset \mathbb{R}^n$ be an open set, $L^1_{loc}(U)$ be the space of locally integral functions (i.e., functions that are integral on every compact subset of U. Note that compactness in $\mathbb{R}^n$ means closed and bounded, e.g., $[a, b]$ is a compact subset of $\mathbb{R}$). Consider $u, v \in L^1_{loc}(U)$. For multi-index α, v is said to be the α^{th} weak derivative of u if

$$\int_U u D^\alpha \phi = (-1)^{|\alpha|} \int_U v\phi,$$

for all infinitely differentiable function ϕ with compact support in U, denoted by $\phi \in C_c^\infty(U)$. Here

$$D^\alpha \phi = \frac{\partial^{|\alpha|}\phi}{\partial^{\alpha_1} x_1 ... \partial^{\alpha_n} x_n}.$$

A weak derivative is unique up to a set of measure zero. A function that is not differentiable at some point or everywhere can still have weak derivative. Canto function is differentiable almost everywhere but does not have weak derivative.

1.3 Simple techniques from measure theory

The expectation is defined by 5 axioms in http://www.stat.umn.edu/geyer/8501/measure.pdf.

The expectation can also be expressed in Lebesgue's integral:

$$\mathbb{E}X = \int_\Omega X(\omega) dP(\omega).$$

Proposition 1. *Let X be a non-negative random variable. We have $\mathbb{E}X = 0$ iff $X \overset{a.s.}{=} 0$.*

Proof. Assume that $\mathbb{E}X = 0$. Let $E_\epsilon := \{X \geq \epsilon\}$ for any $\epsilon \geq 0$. We have $\epsilon 1_{E_\epsilon} \leq X$. Thus, we have $0 = EX \geq E[\epsilon 1_{E_\epsilon}] = \epsilon P(E_\epsilon)$. Thus, $P(E_\epsilon) = 0, \forall \epsilon > 0$ or $P(X = 0) = 1$.

Now, assume that $P(X = 0) = 1$. Let $X_n = \min\{X, n\}, \forall n \geq 0$. We have $\forall \omega \in \Omega, \lim_{n\to\infty} X_n(\omega) = X(\omega)$ which implies that $\mathbb{E}X = \lim_{n\to\infty} E[X_n]$. We also have

$$0 \leq \mathbb{E}[X_n] = \int_\Omega X_n(\omega) dP(\omega) = \int_{\{X_n > 0\}} X_n(\omega) dP(\omega) \leq nP(X_n > 0) \leq nP(X > 0) = 0.$$

Thus, $\mathbb{E}[X_n] = 0, \forall n$, or $\mathbb{E}X = \lim_{n\to\infty} E[X_n] = 0$. □

1.3.1 Union bound

We have

$$\{\exists a \in \mathcal{A} : U_a\} \subseteq \cup_{a\in\mathcal{A}}\{U_a\} \text{ or } Pr(\exists a \in \mathcal{A} : U_a) \leq \sum_{a\in\mathcal{A}} Pr(U_a).$$

An important remark is that the union bound considers all $a \in \mathcal{A}$ equally without focusing on a particular a. This seemly trivial observation is in fact very useful to break the dependence structure in establishing error bounds as a particular a in $\{\exists a \in \mathcal{A} : U_a\}$ can be data dependent or unknown. Another remark is that if $\mathcal{A}$ is infinite, the union bound could be infinity; thus t he union bound is often meaningful when $\mathcal{A}$ is a finite set, e.g., $\mathcal{A}$ is a ϵ-covering

2 Concentration inequalities

2.1 Independent random variables

Lemma 5 (Markov's inequality). *For any non-negative random variable X and $t > 0$, we have*

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

Lemma 6. *Let X be a random variable with $\mathbb{E}X = 0$ and $a \leq X \leq b$ a.s. Then, for any $s > 0$, we have*

$$\mathbb{E}[e^{sX}] \leq e^{s^2(b-a)^2/8}.$$

refs. Chapter 2 of "Combinatorial methods in density estimation".

Definition 2 (Sub-Gaussian random variables). A real-valued random variable X is sub-Gaussian with variance proxy $\sigma^2 > 0$ if

$$\mathbb{E}\left[e^{\lambda(X-\mathbb{E}X)}\right] \leq e^{\sigma^2\lambda^2/2}, \forall \lambda \in \mathbb{R}.$$

Remark. A random variable bounded in $[a, b]$ is sub-Gaussian with variance proxy $\sigma^2 = (b-a)^2/4$.

Proposition 2 (Hoeffding's inequality). *Let $X_1, ..., X_n \sim X$ be i.i.d. real-valued sub-Gaussian random variables with variance proxy σ^2. Then, for any $\epsilon > 0$, we have the following upper-tail bound*

$$P\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}X \geq \epsilon\right) \leq e^{-n\epsilon^2/(2\sigma^2)},$$

and for any $\delta \in [0, 1]$, we have the following upper confidence bound

$$P\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}X < \sqrt{\frac{2\sigma^2\log(1/\delta)}{n}}\right) \geq 1 - \delta.$$

ref. http://www.stats.ox.ac.uk/~rebeschi/teaching/AFoL/19/material/lecture6.pdf

Definition 3 (One-sided Bernstein's condition). A real-valued random variable X is said to satisfy the one-sided Bernstein's condition with parameter $b > 0$ if

$$\mathbb{E}e^{\lambda(X-\mathbb{E}X)} \leq \exp(\frac{(VarX)\lambda^2/2}{1-b\lambda}), \forall \lambda \in [0, 1/b).$$

Remark. If $X - \mathbb{E}X \leq c$ for a given $c > 0$, then X satisfies the one-sided Bernstein's condition with parameter $b = c/3$.

Proposition 3 (Bernstein's inequality). *Let $X_1, ..., X_n \sim X$ be i.i.d. real-valued random variables that satisfy the one-sided Bernstein's condition with parameter $b > 0$. Then, for any $\epsilon > 0$ and $\delta \in [0, 1]$, we have*

$$P\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}X \geq \epsilon\right) \leq \exp\left(-\frac{n\epsilon^2/2}{VarX + b\epsilon}\right),$$

$$P\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}X < \frac{b}{n}\log(1/\delta) + \sqrt{\frac{2(VarX)\log(1/\delta)}{n}}\right) \geq 1 - \delta.$$

ref: http://www.stats.ox.ac.uk/~rebeschi/teaching/AFoL/19/material/lecture7.pdf

Proposition 4 (Matrix Bernstein's inequality). *Let $X_1, X_2, ..., X_n$ be zero-mean, independent, symmetric, $d \times d$ random matrices such that $\|X_i\|_2 \leq c, \forall i$ for a given $c > 0$. Then, for any $\epsilon > 0$, we have*

$$P\left(\|\sum_{i=1}^{n} X_i\|_2 \geq \epsilon\right) \leq 2d\exp\left(\frac{\epsilon^2/2}{\sigma^2 + c\epsilon/3}\right),$$

where $\sigma^2 = \|\sum_{i=1}^{n} \mathbb{E}[X_i^2]\|_2$.

ref. http://www.stat.cmu.edu/~arinaldo/Teaching/36709/S19/Scribed_Lectures/Mar5_Tim.pdf

2.2 Martingale structures

First, we briefly review some relevant terminologies in measure theory. Consider a probability space $(\Omega, \mathcal{F}, P)$ where $\mathcal{F}$ is a σ-field of Ω. A function $X : \Omega \to \mathbb{R}$ is said to be $\mathcal{F}$-measurable if $X(\omega) = X(\omega')$ for any ω' in the smallest set in $\mathcal{F}$ containing ω, i.e., X is called a random variable. The conditional expectation $\mathbb{E}[X|\mathcal{F}] : \Omega \to \mathbb{R}$ is another random variable which is defined by

$$\mathbb{E}[X|\mathcal{F}](\omega) := \frac{\int_{\mathcal{F}(\omega)} X(\omega')dP(\omega')}{P(\mathcal{F}(\omega))},$$

where $\mathcal{F}(\omega)$ is the smallest set in $\mathcal{F}$ containing ω. A special case is when $\mathcal{F} = \{\emptyset, \Omega\}$, we have $\mathbb{E}[X|\mathcal{F}] = \mathbb{E}[X]$ which is deterministic.

A filtration is an increasing sequence of sub-σ-fields: $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq ... \subseteq \mathcal{F}$.

Definition 4 (Martingales). Let $(\Omega, \mathcal{F})$ be a measurable space, $\mathcal{F}_0 := \{\emptyset, \Omega\}$, and $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq ... \subseteq \mathcal{F}$ be a sequence of sub-σ-fields. Let $(X_k)_{k=0,1,...}$ be a sequence of random variables such that X_k is $\mathcal{F}_k$-measurable. The sequence (X_k) is said to be a martingale adapted to the filtration $(\mathcal{F}_k)$ if $\mathbb{E}|X_k| \leq \infty$ and $X_k = \mathbb{E}[X_{k+1}|\mathcal{F}_k]$ for all $k \geq 0$.

Remark (Boob construction of martingales). Let $Y_1, ..., Y_n$ be a sequence of random variables. Let $X = f(Y_1, ..., Y_n)$ for some function f such that X is integrable. We construct a filtration as follows: let $\mathcal{F}_0 = \{\emptyset, \Omega\}$, and define the generated σ-field $\mathcal{F}_k := \sigma(Y_1, ..., Y_k), \forall k \in [1, n]$. Then let $X_k = \mathbb{E}[X|\mathcal{F}_k], k \in [0, n]$. It is not hard to verify that (X_k) is a martingale adapted to the filtration $(\mathcal{F}_k)$. Note that $X_0 = \mathbb{E}[X|\mathcal{F}_0] = \mathbb{E}[X]$ is deterministic and $X_n = \mathbb{E}[X|\mathcal{F}_n] = X$.

Theorem 4 (**Azuma's inequality**). *Consider a Boob construction of martingale (X_k) as above. Assume that $|X_k - X_{k-1}| \le c_k$ a.s. for some $0 < c_k < \infty$ for all $1 \le k \le n$. For any $t \ge 0$, we have*

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

Proof. Let $s > 0$. It follows from Markov's inequality that

$$P(X - \mathbb{E}X \ge t) = P(e^{s(X-\mathbb{E}X)} \ge e^{st}) \le e^{-st}\mathbb{E}[e^{s(X-\mathbb{E}X)}].$$

It follows from Lemma 6 that for any $k \in [1, n]$, we have

$$\mathbb{E}\left[e^{s(X_k - X_{k-1})}|\mathcal{F}_{k-1}\right] \le e^{s^2c_k^2/2},$$

since $X_{k-1} = \mathbb{E}[X_k|\mathcal{F}_{k-1}]$. Thus, we have

$$\begin{aligned}\mathbb{E}[e^{sX}] = \mathbb{E}[e^{sX_n}] = \mathbb{E}\left[\mathbb{E}\left[e^{sX_n}|\mathcal{F}_{n-1}\right]\right] &\le \mathbb{E}\left[e^{s^2c_n^2/2}e^{sX_{n-1}}|\mathcal{F}_{n-1}\right] = e^{s^2c_n^2/2}\mathbb{E}[e^{sX_{n-1}}] \\ &\le e^{s^2\sum_{k=1}^n c_k^2/2}\mathbb{E}[e^{sX_0}] = e^{s^2\sum_{k=1}^n c_k^2/2}e^{s\mathbb{E}X}.\end{aligned}$$

Hence, we have

$$P(X - \mathbb{E}X \ge t) \le e^{-st}e^{s^2\sum_{k=1}^n c_k^2/2}.$$

Minimizing the RHS of the inequality above w.r.t. s results in $s = \frac{t}{\sum_{k=1}^n c_k^2}$. We have

$$P(X - \mathbb{E}X \ge t) \le e^{-\frac{t^2}{2\sum_{k=1}^n c_k^2}}.$$

Considering $-X_k$ instead of X_k leads to a similar inequality:

$$P(X - \mathbb{E}X \le -t) \le e^{-\frac{t^2}{2\sum_{k=1}^n c_k^2}}.$$

Combining these two inequalities above results in the Azuma's inequality. □

Refs: http://www.math.ucsd.edu/~fan/wp/concen.pdf, http://www.stat.cmu.edu/~arinaldo/Teaching/36755/F17/Scribed_Lectures/F17_0913.pdf.

Proposition 5 (**Freedman's inequality** [Tropp et al., 2011]). *Let X be the martingale adapted to a filtration $\mathcal{F}$ (i.e., $X_k = \mathbb{E}[X|\mathcal{F}_k], \forall k = 1, 2, ..., n$ where $X_n = \mathbb{E}[X|\mathcal{F}_n] = X$, and $X_0 = \mathbb{E}[X]$) satisfying that $X_k - X_{k-1} = X_k - \mathbb{E}[X_k|\mathcal{F}_{k-1}] \overset{a.s.}{\le} M, \forall k = 1, ..., n$ where M can be a random variable. Denote the variance process $W := \sum_{i=1}^n Var[X_k|F_{k-1}]$. Then, for all $\epsilon > 0, \sigma^2 > 0$, we have*

$$P\left(X - \mathbb{E}[X] \ge \epsilon, W \le \sigma^2\right) \le \exp\left(\frac{-\epsilon^2/2}{\sigma^2 + M\epsilon/3}\right).$$

In addition, if $|X_k - X_{k-1}| \le \overset{a.s.}{\le} c, \forall k$, we have

$$P\left(|X - \mathbb{E}[X]| \ge \epsilon, W \le \sigma^2\right) \le 2\exp\left(\frac{-\epsilon^2/2}{\sigma^2 + M\epsilon/3}\right).$$

Remark 1. For intuition, we can imagine $X = f(Z_1, ..., Z_n)$ where $Z_1, ..., Z_n$ are random variables for some integrable f, $\mathcal{F}_k = \sigma(Z_1, Z_2, ..., Z_k)$ for all $1 \leq k \leq n$, and $\mathcal{F}_0 = \{\emptyset, \Omega\}$.
Remark 2. Freeman's inequality is a martingale counterpart to Bernstein's inequality. Let $Y_k = X_k - X_{k-1}, \forall k = 1, .., n$. We have $\mathbb{E}[Y_k | \mathcal{F}_{k-1}] = 0, \forall k = 1, ..., n$ and $X - \mathbb{E}[X] = \sum_{k=1}^n Y_k$. If (Y_k) are independent, W is deterministic and the Freedman's inequality reduces to the Bernstein's inequality.

Proposition 6 (**Matrix Freedman's inequality**). *Let $\{Y_k : k = 0, 1, ...\}$ be a matrix martingale adapted to filtration $\{\mathcal{F}_k : k = 0, 1, ...\}$ whose values are Hermitian matrices of dimension d with difference sequence $\{X_k : k = 1, 2, ...\}$. Assume that $\lambda_{\max}(X_k) \overset{a.s.}{\leq} c, \forall k \geq 1$ for a given $c > 0$. Define*

$$W_k := \sum_{j=1}^{k} \mathbb{E}_{j-1}[X_j^2], \forall k \geq 1,$$

where $\mathbb{E}_j[\cdot] := \mathbb{E}[\cdot | \mathcal{F}_j]$. Then, for all $\epsilon > 0, \sigma^2 > 0$, we have

$$P\left(\exists k \geq 0 : \lambda_{\max}(Y_k) \geq \epsilon \text{ and } \|W_k\|_2 \leq \sigma^2\right) \leq d \exp\left(\frac{-\epsilon^2/2}{\sigma^2 + c\epsilon/3}\right).$$

ref. https://arxiv.org/abs/1101.3039.

3 Uniform convergence bounds

A uniform convergence bound can be useful when an error quantity involves an unknown decision point (e.g., in the population risk minimizer is unknown in the risk minimization problem). It can also be used to break a dependency structure in algorithm analysis. This is useful because often in practice (especially in RL settings), an algorithm might have a complicated dependence structure that is different from the Independence or martingale structure. A uniform convergence bound can reduce such complicated dependence structure into a martingale structure where concentration inequalities are available.

3.1 Empirical risk minimization

Let $\mathcal{B}$ be a set of all admissible actions [1], $\mathcal{A} \subseteq \mathcal{B}$ is a subset of admissible actions (think of it as a hypothesis space or function approximations), $S = \{Z_1, ..., Z_n\}$ be a set of independent r.v.s on $\mathcal{Z}$, $l : \mathcal{A} \times \mathcal{Z} \to \mathbb{R}_+$ is a loss function. Let $r(a) := \mathbb{E} l(a, Z)$ be the population risk and $R(a) := \frac{1}{n} \sum_{i=1}^n l(a, Z_i)$ be the empirical risk. Let us denote

$$a^* \in \arg\inf_{a \in \mathcal{A}} r(a)$$
$$A^* \in \arg\inf_{A \in \mathcal{A}} R(A).$$

We are interested in bounding the excess risk

$$r(A^*) - \inf_{a \in \mathcal{B}} r(a) = \underbrace{r(A^*) - r(a^*)}_{\text{estimation error}} + \underbrace{r(a^*) - \inf_{a \in \mathcal{B}} r(a)}_{\text{approximation error}},$$

where the estimation error is due to the access of only empirical data S and approximation error is due to using different decision space $\mathcal{A}$ instead of $\mathcal{B}$. In this course we will focus on bounding the estimation error.

[1] The decision space $\mathcal{A}$ is any abstract topology, e.g., $\mathbb{R}^d$ or a set of functions of a particular form.

We bound the estimation error uniformly over the decision space $\mathcal{A}$:

$$\begin{aligned} r(A^*) - r(a^*) &= r(A^*) - R(A^*) + \underbrace{R(A^*) - R(a^*)}_{\leq 0} + R(a^*) - r(a^*) \\ &\leq \sup_{a\in\mathcal{A}}\{r(a) - R(a)\} + \sup_{a\in\mathcal{A}}\{R(a) - r(a)\}. \end{aligned}$$

Now let us take a look at a general recipe to derive a uniform bound for the estimation error. Assume that we have the following bounds in expectation

$$\begin{aligned} \mathbb{E}\sup_{a\in\mathcal{A}}\{r(a) - R(a)\} &\leq \frac{1}{2}B_{expectation} \\ \mathbb{E}\sup_{a\in\mathcal{A}}\{R(a) - r(a)\} &\leq \frac{1}{2}B_{expectation} \end{aligned}$$

and bounds in probability, i.e., for any $\epsilon > 0$,

$$\begin{aligned} P\left(\sup_{a\in\mathcal{A}}\{r(a) - R(a)\} \geq \mathbb{E}\sup_{a\in\mathcal{A}}\{r(a) - R(a)\} + \epsilon\right) &\leq \frac{1}{2}B_{uppertail}(\epsilon), \\ P\left(\sup_{a\in\mathcal{A}}\{R(a) - r(a)\} \geq \mathbb{E}\sup_{a\in\mathcal{A}}\{r(a) - R(a)\} + \epsilon\right) &\leq \frac{1}{2}B_{uppertail}(\epsilon), \end{aligned}$$

where $B_{uppertail}(\epsilon)$ is a strictly decreasing function of ϵ. Then for any $\delta \in [0,1]$, with probability at least $1-\delta$, we have

$$\begin{aligned} r(A^*) - r(a^*) &\leq \sup_{a\in\mathcal{A}}\{r(a) - R(a)\} + \sup_{a\in\mathcal{A}}\{R(a) - r(a)\} \\ &< \mathbb{E}\sup_{a\in\mathcal{A}}\{r(a) - R(a)\} + \mathbb{E}\sup_{a\in\mathcal{A}}\{R(a) - r(a)\} + 2B^{-1}_{uppertail}(\delta) \\ &\leq B_{expectation} + 2B^{-1}_{uppertail}(\delta). \end{aligned}$$

In practice, we often derive bounds in probability by concentration inequalities and bounds in expectation using a complexity notion of $\mathcal{A}$.

4 Rademacher complexity

4.1 Maximal inequalities

We are interested in bounding the expected maximal process:

$$\mathbb{E}\sup_{a\in\mathcal{A}}\{r(a) - R(a)\}.$$

We will bound that via Rademacher complexity.

Definition 5 (**Rademacher complexity**)**.** Let $\Omega_1, ..., \Omega_n \in \{-1, 1\}$ be independent Rademacher r.v.s (and independent of any other r.v.s. in our models), i.e., $P(\Omega_i = 1) = P(\Omega_i = -1) = 0.5$. The Rademacher complexity of a set $\mathcal{T} \subseteq \mathbb{R}^d$ is defined as

$$Rad(\mathcal{T}) := \mathbb{E}\sup_{t\in\mathcal{T}}\frac{1}{n}\sum_{i=t}^{n}\Omega_i t_i.$$

Denote

$$\begin{aligned} \mathcal{L} &:= \{z \in \mathcal{Z} \to l(a,z) \in \mathbb{R} : a \in \mathcal{A}\} \\ \mathcal{L} \circ \{z_1, ..., z_n\} &:= \{(l(z_1,a), ..., l(z_n,a)) \in \mathbb{R}^n : a \in \mathcal{A}\}. \end{aligned}$$

Proposition 7. *We have*

$$\mathbb{E}\sup_{a\in\mathcal{A}}\{r(a)-R(a)\} \leq 2\mathbb{E}Rad(\mathcal{L}\circ\{z_1,...,z_n\}) = 2\mathbb{E}\sum_{a\in\mathcal{A}}\frac{1}{n}\sum_{i=1}^{n}\Omega_i l(a,Z_i).$$

Proposition 8 (Properties of Rademacher complexity). *Let $\mathcal{T}\subseteq\mathbb{R}^n$.*

1. *(Scale and shift) $\forall v\in\mathbb{R}^n, c\in\mathbb{R}$, we have*

$$Rad(c\mathcal{T}+v) = |c|Rad(\mathcal{T}).$$

2. *(Summation)*

$$Rad(\mathcal{T}+\mathcal{T}') = Rad(\mathcal{T})+Rad(\mathcal{T}').$$

3. *(Convex hull) Let $conv(\mathcal{T}) = \{\sum_{j=1}^m w_j t_j : w\in\Delta_m, t_1,...,t_m\in\mathcal{T}, m\in\mathbb{N}\}$. We have*

$$Rad(conv(\mathcal{T})) = Rad(\mathcal{T}).$$

4. *(Massart's lemma) Assume $|\mathcal{T}|<\infty$ and define $\bar{t}=\frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}}t$. We have*

$$Rad(\mathcal{T}) \leq \max_{t\in\mathcal{T}}\|t-\bar{t}\|_2\frac{\sqrt{2\log|\mathcal{T}|}}{n}.$$

5. *(Contraction, or Telegrand's lemma) Let $f_i:\mathbb{R}\to\mathbb{R}$ be a λ-Lipschitz function for each $i=1:n$. We have*

$$Rad((f_1,...,f_n)\circ\mathcal{T}) \leq \gamma Rad(\mathcal{T}),$$

where

$$(f_1,...,f_n)\circ\mathcal{T} := \{(f_1(t_1),...,f_n(t_n))\in\mathbb{R}^n : t\in\mathcal{T}\}.$$

4.2 Rademacher complexity examples

Here we will derive the Rademacher complexity of feed-forward neural networks. The k-th layer $l^{(k)}:\mathbb{R}^{d_{k-1}}\to\mathbb{R}^{d_k}$ is defined by

$$l^{(k)}(x) := \sigma^{(k)}(w^{(k)}x+b^{(k)}),$$

where a given weight matrix w^k and bias vector $b^{(k)}$.

A feed-forward network of depth $H\geq 1$ is defined by $f_{nn}^{(H)}:\mathbb{R}^d\to\mathbb{R}$:

$$f_{nn}^{(H)}(x) := l^{(H)}\circ...\circ l^{(1)}(x),$$

where $d_0=d, d_H=1, \sigma^{(h)}$ is γ-Lipschitz functions for all $h<H$ and $\sigma^{(H)}(x)=x$. Consider the following class of neural networks

$$\mathcal{A}_{nn}^{(H)} = \{x\in\mathbb{R}^d\to f_{nn}^{(H)}(x) : \|w^{(k)}\|_\infty\leq\omega, \|b^{(k)}\|_\infty\leq\beta, \forall k=1:H\},$$

where for a given matrix m, $\|m\|_\infty := \max_i\sum_j|m_{i,j}|$.

Proposition 9. *For any $x_1,...,x_n\in\mathbb{R}^d$, we have*

$$Rad(\mathcal{A}_{nn}^{(H)}\circ\{x_1^n\}) \leq \frac{\beta}{\sqrt{n}}\sum_{k=0}^{H-2}(2\omega\gamma)^k+\gamma^{H-2}(2\omega)^{H-1}\frac{\omega}{\sqrt{n}}\max_i\|x_i\|_\infty\sqrt{2\log 2d}.$$

This result is a direct consequence of the following two lemmas.

Lemma 7. *Let $\mathcal{L}$ be a class of functions $\mathbb{R}^d \to \mathbb{R}$ including the zero function, $\sigma : \mathbb{R} \to \mathbb{R}$ be γ-Lipschitz. Define*

$$\mathcal{L}_1 := \{x \in \mathbb{R}^d \to \sigma(w^T L(x) + b) \in \mathbb{R} : |b| \leq \beta, \|w\|_1 \leq \omega, l_1, ..., l_m \in \mathcal{L}\},$$

where $L(x) := (l_1(x), ..., l_m(x)) \in \mathbb{R}^m$. Then, for any $x_1, ..., x_n \in \mathbb{R}^d$, we have

$$Rad(\mathcal{L}_1 \circ \{x_1^n\}) \leq \gamma \left(\frac{\beta}{\sqrt{n}} + 2\omega Rad(\mathcal{L} \circ \{x_1^n\}) \right).$$

Proof. First, let us define

$$\begin{aligned} \mathcal{A} &:= \{x \in \mathbb{R}^d \to w^T L(x) \in \mathbb{R} : \|w\|_1 \leq \omega, l_1, ..., l_m \in \mathcal{L}\} \\ \mathcal{B} &:= \{x \in \mathbb{R}^d \to b \in \mathbb{R} : |b| \leq \beta\} \end{aligned}$$

We have $\mathcal{L}_1 \circ \{x_1^n\} = \sigma \circ (\mathcal{A} + \mathcal{B}) \circ \{x_1^n\}$. It follows from the contraction and summation property of Rademacher complexity that

$$Rad(\mathcal{L}_1 \circ \{x_1^n\}) = Rad(\sigma \circ (\mathcal{A} + \mathcal{B}) \circ \{x_1^n\}) \leq \gamma \left(Rad(\mathcal{A} \circ \{x_1^n\}) + Rad(\mathcal{B} \circ \{x_1^n\}) \right)$$

We have

$$nRad(\mathcal{B} \circ \{x_1^n\}) = \mathbb{E} \sup_{|b| \leq \beta} \sum_{i=1}^n \Omega_i b \leq \beta \mathbb{E} \sqrt{(\sum_{i=1}^n \Omega_i)^2} \overset{(a)}{\leq} \beta \sqrt{\mathbb{E}(\sum_{i=1}^n \Omega_i)^2} = \beta \sqrt{n},$$

where (a) follows from Jensen's inequality for convex function $x \mapsto \sqrt{x}$.

We also have

$$\begin{aligned} nRad(\mathcal{A} \circ \{x_1^n\}) &= \mathbb{E} \sup_{\|w\|_1 \leq \omega, l_1, ..., l_m \in \mathcal{L}} \sum_{i=1}^n \Omega_i w^T L(x_i) \\ &= \mathbb{E} \sup_{\|w\|_1 \leq \omega, l_1, ..., l_m \in \mathcal{L}} w^T \sum_{i=1}^n \Omega_i L(x_i) \\ &\overset{(b)}{\leq} \mathbb{E} \sup_{\|w\|_1 \leq \omega} \|w\|_1 \| \sum_{i=1}^n \Omega_i L(x_i) \|_\infty \\ &\leq \omega \mathbb{E} \sup_{l_1, ..., l_m \in \mathcal{L}} \max_{j=1:m} | \sum_{i=1}^n \Omega_i l_j(x_i) | \\ &= \omega \mathbb{E} \sup_{l \in \mathcal{L}} | \sum_{i=1}^n \Omega_i l(x_i) | \\ &= \omega \mathbb{E} \sup_{l \in \mathcal{L} \cup \mathcal{L}_-} \sum_{i=1}^n \Omega_i l(x_i) \\ &\overset{(c)}{\leq} \omega \mathbb{E} \sup_{l \in \mathcal{L} - \mathcal{L}} \sum_{i=1}^n \Omega_i l(x_i) \\ &= \omega Rad((\mathcal{L} - \mathcal{L}) \circ \{x_1^n\}) = \omega (Rad(\mathcal{L} \circ \{x_1^n\}) + Rad((-\mathcal{L}) \circ \{x_1^n\})) \\ &= 2\omega Rad(\mathcal{L} \circ \{x_1^n\}), \end{aligned}$$

where (b) follows from Holder's inequality and (c) follows from $\mathcal{L} \cup \mathcal{L}_- \subseteq \mathcal{L} - \mathcal{L}$. Here, $\mathcal{L}_- := \{-l : l \in \mathcal{L}\}$ and $\mathcal{L} - \mathcal{L} := \{l_1 - l_2 : l_1, l_2 \in \mathcal{L}\}$.

Thus we conclude the proof. □

Lemma 8. *Let us consider*

$$\mathcal{A}_1 := \{x \in \mathbb{R}^d \to w^T x : \|w\|_1 \le \omega\}.$$

Then, for any $x_1, ..., x_n \in \mathbb{R}^d$*, we have*

$$Rad(\mathcal{A}_1 \circ \{x_1^n\}) \le \omega \max_i \|x_i\|_\infty \frac{\sqrt{2\log(2d)}}{\sqrt{n}}.$$

Proof. We have

$$\begin{aligned}
nRad(\mathcal{A}_1 \circ \{x_1^n\}) &= \mathbb{E} \sup_{\|w\|_1 \le \omega} \sum_{i=1}^n \Omega w^T x_i = \mathbb{E} \sup_{\|w\|_1 \le \omega} w^T \sum_{i=1}^n \Omega x_i \\
&\overset{(d)}{\le} \mathbb{E} \sup_{\|w\|_1 \le \omega} \|w\|_1 \| \sum_{i=1}^n \Omega_i x_i \|_\infty \\
&\le \omega \mathbb{E} \max_{j=1:d} | \sum_{i=1}^n \Omega_i x_{i,j} | \\
&= \omega \mathbb{E} \max_{t \in \mathcal{T} \cup \mathcal{T}_-} \sum_{i=1}^n \Omega_i t_i \\
&\overset{(e)}{\le} \omega \max_{t \in \mathcal{T} \cup \mathcal{T}_-} \|t - \bar{t}\|_2 \frac{\sqrt{2 \log |\mathcal{T} \cup \mathcal{T}_-|}}{n} \\
&= \omega \max_{t \in \mathcal{T}} \|t\|_2 \frac{\sqrt{2\log(2d)}}{n} \\
&\le \omega \sqrt{n} \max_i \|x_i\|_\infty \sqrt{2 \log(2d)}
\end{aligned}$$

where (d) follows from Holder's inequality, (e) follows from Massart's lemma, and $\mathcal{T} := \{t^{(j)} = (x_{1,j}, ..., x_{n,j}) \in \mathbb{R}^n : j = 1 : d\}, \mathcal{T}_- := \{-t : t \in \mathcal{T}\}$. □

5 Minimax lower bounds

5.1 Hypothesis testing

Consider the binary hypothesis testing: Let P and Q be two probability measures on some measurable space $(\Omega, \mathcal{F})$. Let $X : \Omega \to \mathcal{X}$ be a random variable that is either drawn from P (the null hypothesis H_0) or Q (the alternative hypothesis H_1). A test $f : \mathcal{X} \to \{0, 1\}$ is designed to determine which distribution the sample X is drawn from (i.e., which hypothesis is true). Any test can commit two types of errors: a type I error if $f(X) = 1$ when $X \sim P$, and a type II eror if $f(X) = 0$ when $X \sim Q$.

We say that P and Q has densitites p and q w.r.t a measure ρ if for any measurable event E, we have

$$\begin{aligned}
P(E) &= \int \rho(dx) p(x) 1_E(x), \\
Q(E) &= \int \rho(dx) q(x) 1_E(x).
\end{aligned}$$

Remark. If ρ is the Lebesgue measure, we write $\rho(dx) = dx$. If $\rho(dx) = \sum_{i=1}^n \delta_{x_i}(dx), P(E) = \sum_{i:x_i \in E} p(x_i)$.

Theorem 5 (Neyman Pearson). *For any* $f : \mathcal{X} \to \{0, 1\}$*, we have*

$$P(f(X) = 1) + Q(f(X) = 0) \ge \int \rho(dx) \min\{p(x), q(x)\} = 1 - TV(P, Q),$$

where the equality occurs at $f^* := 1_{q \ge p}$*, and the total variation distance* $T(P, Q) := \sup_E |P(E) - P(Q)|$.

Remark 1. Neyman Pearson test says that unless P and Q have disjoint support under the reference measure ρ, any test f commits at least one of the error types with strictly positive probability.
Remark 2. This lower bound presents a structural limitation of what we could hope to achieve statistically based on the amount of information in the problem instance (determined by P and Q in the binary hypothesis testing).
Remark 3: TV is bounded in $[0,1]$. $TV(P,Q)=0$ iff $P=Q$ and $TV(P,Q)=1$ iff P and Q have disjoint supports.
Remark 4: Proving the lower bounds using the Neyman Pearson lemma reduces to proving upper bounds on the total variation distance.

Proof. For any test f, let $R=\{f=1\}:=\{x\in\mathcal{X}:f(x)=1\}$. Let $R^*=\{f^*=1\}=\{q\geq p\}:=\{x\in\mathcal{X}:q(x)\geq p(x)\}$. We have

$$\begin{aligned}P(f(X)=1)+Q(f(X)=0)&=1+P(R)-Q(R)\\&=1+\int\rho(dx)|p(x)-q(x)|(1_{R\cap(R^*)^c}(x)-1_{R\cap R^*}(x))\\&\geq 1-\int\rho(dx)|p(x)-q(x)|,\end{aligned}$$

where the equality occurs iff $R=R^*$, i.e., $f=f^*$. At $f=f^*$, it is easy to see that $P(f^*(X)=1)+Q(f^*(X)=0)=\int\rho(dx)\min\{p(x),q(x)\}$. □

Definition 6 (**KL divergence**). Let P and Q be two probability measures on the same measurable space, with densities p and q, respectively, w.r.t. some reference measure ρ. The KL divergence is defined by

$$KL(P,Q)=\begin{cases}\int\rho(dx)p(x)\log\frac{p(x)}{q(x)} & \text{if } P\ll Q,\\ \infty & \text{otherwise.}\end{cases}$$

Proposition 10 (**Properties of KL divergence**). *Let P and Q be two probability measures on the same measurable space. We have*

1. *(Gibbs' inequality) $KL(P,Q)\geq 0$ with equality iff $P=Q$.*
2. *(Chain rule for product distributions) If $P=\otimes_{i=1}^n P_i$ and $Q=\otimes_{i=1}^n Q_i$, then*

$$KL(P,Q)=\sum_{i=1}^n KL(P_i,Q_i).$$

3. *(Pinsker's inequality) For any measurable event, we have*

$$\begin{aligned}P(E)-Q(E)&\leq\sqrt{\frac{1}{2}KL(P,Q)}\\P(E)-Q(E)&\leq\sqrt{\frac{1}{2}KL(Q,P)},\end{aligned}$$

which thus yields

$$\begin{aligned}TV(P,Q)&\leq\sqrt{\frac{1}{2}KL(P,Q)}\\TV(P,Q)&\leq\sqrt{\frac{1}{2}KL(Q,P)}.\end{aligned}$$

Corollary. *Let $X_1, ..., X_n$ i.i.d. from either P or Q on $\mathcal{X}$. For any test $f : \mathcal{X}^n \to \{0,1\}$, we have*

$$P(f(X_1, ..., X_n) = 1) + Q(f(X_1, ..., X_n) = 0) \geq 1 - \sqrt{\frac{n}{2} KL(P, Q)}.$$

Remark. This lower bound characterizes the amount of information as a function of the number of data points n and a discrepancy between P and Q.

5.2 Minimax theory

We present the general procedure to establish minimax lower bounds. [2]

Setting. Let P be a model (i.e., a probability measure) on the measurable space $(\Omega, \mathcal{F})$ and let $X_1, ..., X_n$ be samples of P. Let θ be a map from probability measures on $(\Omega, \mathcal{F})$ into some metric space (Θ, d). We are interested in estimating θ via some measurable map $\hat{\theta}(X_1, ..., X_n) \in \Theta$. We judge the goodness of the estimator $\hat{\theta}$ by meas of expected loss: $\mathbb{E}_P \left[d(\hat{\theta}(X_1, ..., X_n), \theta(P)) \right]$. Consider the scenario where we do not know P but only its neighborhood $\mathcal{P} \ni P$. Since we do not know P, we are interested in an estimator that can do well not only for a particular model P but for any models in $\mathcal{P}$. Thus, we seek for an estimator that minimizes the maximum expected loss, the minimax risk:

$$R_n(\mathcal{P}) := \inf_{\hat{\theta}} R_n(\hat{\theta}, \mathcal{P}) := \inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P \left[d(\hat{\theta}(X_1, ..., X_n), \theta(P)) \right].$$

Remark 1. An estimator $\hat{\theta}^* \in \arg\inf_{\hat{\theta}} R_n(\hat{\theta}, \mathcal{P})$ is called **minimax** which performs best in the worst-case scenario.
Remark 2. *The minimax rate of convergence* is $n^{-\alpha}$ for some $\alpha > 0$ if there exits $0 < c < C < \infty$ such that $c \leq R_n(\mathcal{P})/n^{-\alpha} \leq C, \forall n$ sufficiently large.

We often could not find $R_n(\mathcal{P})$ exactly. Instead, we find an upper bound U_n and lower bound L_n of $R_n(\mathcal{P})$. Finding an upper bound is often easier than an lower bound because any estimator $\hat{\theta}$ gives an upper bound on $R_n(\mathcal{P})$. We focus on finding minimax lower bounds here.

All the minimax lower bound methods requires a procedure of reducing the problem to a hypothesis testing problem. We will illustrate this reduction procedure via the following theorem.

Theorem 6 (Reduction procedure). *Let $M = \{P_1, ..., P_N\} \subset \mathcal{P}$ and let $s = \min_{i \neq j} d(\theta(P_i), \theta(P_j))$. Then we have*

$$R_n(\mathcal{P}) \geq \frac{s}{2} \inf_{\psi : \Omega^n \to 1:N} \max_{i=1:N} P_i(\psi(X_1, ..., X_n) \neq i).$$

Remark. We can think of $\psi(X_1, ..., X_n) \in 1 : N$ as a multiple hypothesis test, i.e. a function that tests which of the distributions in M the $X_1, ..., X_n$ comes from.

Proof. We have

$$\begin{aligned} R_n(\mathcal{P}) &= \inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P \left[d(\hat{\theta}(X_1, ..., X_n), \theta(P)) \right] \\ &\geq \inf_{\hat{\theta}} \sup_{i \in 1:N} \mathbb{E}_{P_i} \left[d(\hat{\theta}(X_1, ..., X_n), \theta(P_i)) \right] \\ &\geq \frac{s}{2} \inf_{\hat{\theta}} \sup_{i \in 1:N} P_i(d(\hat{\theta}(X_1, ..., X_n), \theta(P_i)) > s/2), \end{aligned}$$

where the second inequality follows from Markov's inequality for $t = s/2$.

[2] Ref:http://www.stat.cmu.edu/~larry/=sml/minimax.pdf

Denote $\hat{\theta} := \hat{\theta}(X_1, ..., X_n)$ for simplicity. Let $\psi^*(\hat{\theta}) := \arg\min_{j=1:N} d(\hat{\theta}, \theta(P_j))$. For any $i \in 1 : N$, if $\psi^*(\hat{\theta}) \neq i$, we have

$$s \leq d(\theta(P_i), \theta(P_{\psi^*(\hat{\theta})})) \leq d(\theta(P_i), \hat{\theta}) + d(\hat{\theta}, \theta(P_{\psi^*(\hat{\theta})})) \leq 2d(\theta(P_i), \hat{\theta}).$$

Thus, we have

$$\{d(\theta(P_i), \hat{\theta}) > s/2\} \supseteq \{\psi^*(\hat{\theta}) \neq i\}.$$

Hence, we have

$$\begin{aligned} R_n(\mathcal{P}) &\geq \frac{s}{2} \inf_{\hat{\theta}} \sup_{i \in 1:N} P_i(d(\hat{\theta}, \theta(P_i)) > s/2) \\ &\geq \frac{s}{2} \inf_{\hat{\theta}} \sup_{i \in 1:N} P_i(\psi^*(\hat{\theta}) \neq i) \\ &\geq \frac{s}{2} \inf_{\psi} \sup_{i \in 1:N} P_i(\psi \neq i), \end{aligned}$$

where the third inequality follows from that $\{\psi^*(\hat{\theta}(\cdot)) : \Omega^n \to \{1, ..., N\} : \hat{\theta}\} \subseteq \{\psi(\cdot) : \Omega^n \to \{1, ..., N\}\}$. □

References

Joel Tropp et al. Freedman's inequality for matrix martingales. *Electronic Communications in Probability*, 16:262–270, 2011.