Empirical Risk Minimization and Uniform Convergence

Zhehao Li

March 17, 2022

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 consider the methods based on empirical risk minimization.

Risk Minimization Decomposition

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

$$\begin{aligned} \mathcal{R}(\hat{f}) - \mathcal{R}^* &= \left\{ \mathcal{R}(\hat{f}) - \inf_{f' \in \mathcal{F}} \mathcal{R}(f') \right\} + \left\{ \inf_{f' \in \mathcal{F}} \mathcal{R}(f') - \mathcal{R}^* \right\} \\ &= \text{estimation error} \quad + \quad \text{approximation error} \end{aligned} \tag{1}$$

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 $\varphi(x)$.

Approximation Error

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

$$\inf_{\theta \in \Theta} \mathcal{R}(f_\theta) - \mathcal{R}^* = \left(\inf_{\theta \in \Theta} \mathcal{R}(f_\theta) - \inf_{\theta \in \mathbb{R}^d} \mathcal{R}(f_\theta) \right) + \left(\inf_{\theta \in \mathbb{R}^d} \mathcal{R}(f_\theta) - \mathcal{R}^* \right) \tag{2}$$

Lipschitz Continuous Loss

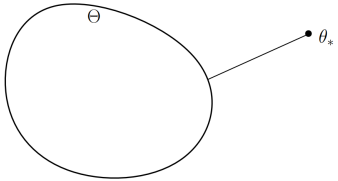


Figure: The distance between minimizer θ_* and set Θ on $\mathbb{R}$

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

An Example

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

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

which equals to zero if $||\theta_*||_2 \leq D$.

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 $\hat{f} \in \arg\min_{f \in \mathcal{F}} \hat{\mathcal{R}}(f) = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} l(y_i, f(x_i))$. That is

$$
\begin{aligned}
\mathcal{R}(\hat{f}) - \inf_{f \in \mathcal{F}} \mathcal{R}(f) &= \mathcal{R}(\hat{f}) - \mathcal{R}(g) \\
&= \left\{ \mathcal{R}(\hat{f}) - \hat{\mathcal{R}}(\hat{f}) \right\} + \left\{ \hat{\mathcal{R}}(\hat{f}) - \hat{\mathcal{R}}(g) \right\} + \left\{ \hat{\mathcal{R}}(g) - \mathcal{R}(g) \right\} \\
&\leq \sup_{f \in \mathcal{F}} \left\{ \mathcal{R}(f) - \hat{\mathcal{R}}(f) \right\} + \left\{ \hat{\mathcal{R}}(\hat{f}) - \hat{\mathcal{R}}(g) \right\} + \sup_{f \in \mathcal{F}} \left\{ \hat{\mathcal{R}}(f) - \mathcal{R}(f) \right\} \\
&\leq \sup_{f \in \mathcal{F}} \left\{ \mathcal{R}(f) - \hat{\mathcal{R}}(f) \right\} + 0 + \sup_{f \in \mathcal{F}} \left\{ \hat{\mathcal{R}}(f) - \mathcal{R}(f) \right\} \\
&\leq 2 \sup_{f \in \mathcal{F}} \left| \hat{\mathcal{R}}(f) - \mathcal{R}(f) \right|
\end{aligned}
\tag{3}
$$

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

Analysis

Now, we can make the following observations about the bound of estimation error

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

- When $\hat{f}$ is not the global minimizer of $\hat{\mathcal{R}}$ but simply satisfies $\hat{\mathcal{R}}(\hat{f}) \leq \inf_{f \in \mathcal{F}} \hat{\mathcal{R}}(f) + \epsilon$, 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.

Basic Concept

Uniform convergence is a technique that helps achieve such bounds. Uniform convergence is a property of a parameter set Θ, which gives us bounds of the form

$$\mathbb{P}\left(\left|\hat{\mathcal{R}}(f) - \mathcal{R}(f)\right| \geq \epsilon\right) \leq \delta \quad \forall\ f \in \mathcal{F} \tag{4}$$

In other words, uniform convergence tells us that for any choice of $\mathcal{F}$, our empirical risk is always close to our population risk with high probability.

- 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 $\mathbb{E}[\sup_{f\in\mathcal{F}} |\hat{\mathcal{R}}(f) - \mathcal{R}(f)|]$

Uniform Deviation from Expectation I

Let $H(Z_1, \ldots, Z_n) = \sup_{f \in \mathcal{F}} \left\{ \hat{\mathcal{R}}(f) - \mathcal{R}(f) \right\}$, 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 $l_\infty = \max_i |l(Y_i, f(X_i))|$ be the maximal absolute value of the loss functions for all (x, y) in the support of the data generating distribution.
When changing a single $Z_i \in \mathcal{X} \times \mathcal{Y}$ into $Z_i' \in \mathcal{X} \times \mathcal{Y}$, the deviation in H is almost surely at most $\frac{2}{n} l_\infty$, that is

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

Thus applying the MacDiarmid inequality,

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

Uniform Deviation from Expectation II

By setting $\delta = \exp\left(-nt^2/2l_\infty^2\right)$, 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 $H(Z_1,\ldots,Z_n) = \sup_{f\in\mathcal{F}}\left\{\hat{\mathcal{R}}(f) - \mathcal{R}(f)\right\}$, we have

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

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

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

$$||X||_{a,b} = \sup\left\{||Xu||_a \mid ||u||_b \leq 1\right\} \tag{5}$$

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 $||X||_2$:

$$||X||_2 = \sigma_{\max}(X) = \left(\lambda_{\max}(X^\top X)\right)^{1/2} \tag{6}$$

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.

Linear Model and Quadratic Loss I

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

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

Linear Model and Quadratic Loss II

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

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

where $||M||_{op}$ is the operator norm of the matrix M.

Bounding the Matrix I

Theorem (The Laplace Transform Method)

Let Y be a random self-joint matrix. Then for all $t \in \mathbb{R}$,

$$\mathbb{P}\left(\lambda_{\max}(Y) \geq t\right) \leq \inf_{\theta > 0} e^{-\theta t} \cdot \mathbb{E}\left[\operatorname{tr} e^{\theta Y}\right] \tag{7}$$

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

$$\mathbb{P}(\lambda_{\max} \geq t) = \mathbb{P}(\lambda_{\max}(\theta Y) \geq \theta t) = \mathbb{P}\left(e^{\lambda_{\max}(\theta Y)} \geq e^{\theta t}\right) \leq e^{-\theta t} \cdot \mathbb{E} e^{\lambda_{\max}(\theta Y)}$$

Bounding the Matrix II

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.

Bounding the Matrix III

Theorem (Matrix Hoeffding's Bound)

Given n independent symmetric matrices $M_i \in \mathbb{R}^{d\times d}$, such that for all $i \in \{1,\dots,n\}$, $\mathbb{E}[M_i] = 0$, $M_i^2 \preceq C_i^2$ almost surely. Then for all $t \geq 0$,

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

where $\sigma^2 = \lambda_{\max}(\frac{1}{n}\sum_{i=1}^{n} C_i^2)$ is the maximum eigenvalue of sample average matrices.

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

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

Bounding the Matrix IV

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 matrices $\{M_i\}$, we can apply matrix Hoeffding's inequality and get

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

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

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

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

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

Bounding the Vector I

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

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

Bounding the Vector II

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

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n} Y_i\varphi(X_i) - \mathbb{E}[Y_i\varphi(X_i)]\right\|_2 \geq t\right) \leq \sum_{j=1}^{d} 2\exp\left(-\frac{2nt^2}{dc^2}\right) = 2d\exp\left(-\frac{2nt^2}{dc^2}\right) \tag{9}$$

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

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

Bounding 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 $1 - \delta$,

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

Direct Bounding Approach

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

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

For $f \in \mathcal{F}$ fixed, 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 $1-\delta$,

$$\mathcal{R}(\hat{f}) - \inf_{f\in\mathcal{F}}\mathcal{R}(f) \leq 2l_\infty\sqrt{\frac{\log(2|\mathcal{F}|/\delta)}{n}} \tag{10}$$

Bounding the Expectation I

Theorem (Expectation of the Maximum)

If $Z_1, \ldots, Z_n$ are (potentially dependent) random variables which are σ-sub-Gaussian, then

$$\mathbb{E}[\max\{Z_1 - \mathbb{E}[Z_1], \ldots, Z_n - \mathbb{E}[Z_n]\}] \leq \sqrt{2\sigma^2 \log n}$$

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

$$\mathbb{E}\left[\mathcal{R}(\hat{f}) - \inf_{f \in \mathcal{F}} \mathcal{R}(f)\right] \leq 2\mathbb{E}\left[\sup_{f \in \mathcal{F}} \left|\hat{R}(f) - \mathcal{R}(f)\right|\right] \leq 2l_\infty \sqrt{\frac{2 \log |\mathcal{F}|}{n}} \tag{11}$$

Bounding the Expectation II

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

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

Bounding the Expectation III

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

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

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

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

I Minimizer over t, we get $t = \frac{\sqrt{2n\log|\mathcal{F}|}}{l_\infty}$, and therefore $\mathbb{E}\left[\sup_{f\in\mathcal{F}} |\hat{\mathcal{R}}(f) - \mathcal{R}(f)|\right] \leq l_\infty\sqrt{\frac{2\log|\mathcal{F}|}{n}}$.

Comparison of Two Approaches

- Directly bounding approach

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

- Bounding via unifrom deviation from expectation

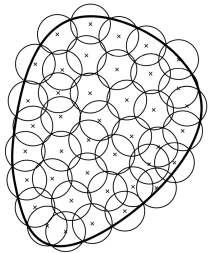
$$\mathcal{R}(\hat{f}) - \inf_{f \in \mathcal{F}} \mathcal{R}(f) \leq 2 \sup_{f \in \mathcal{F}} \left\{ \hat{\mathcal{R}}(f) - \mathcal{R}(f) \right\} \leq 2 l_\infty \left(\sqrt{\frac{2\log(1/\delta)}{n}} + \sqrt{\frac{2\log|\mathcal{F}|}{n}} \right)$$

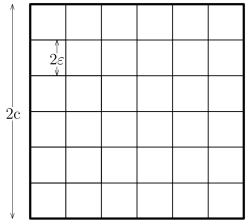
Covering Numbers

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".

Definition (Covering Numbers)

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





Bounding the Expectation I

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 $(f_i)_{i\in\{1,\dots,m_\epsilon\}}$ the associated cover elements

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

Bounding the Expectation. Using bounds 11 on the expectation of the maximum (bounded random variables are sub-Gaussian), we have

$$\begin{aligned}\mathbb{E}\left[\sup_{f\in\mathcal{F}} \left|\hat{\mathcal{R}}(f) - \mathcal{R}(f)\right|\right] &\leq 2G\epsilon + \mathbb{E}\left[\sup_{i\in\{1,\dots,m(\epsilon)\}} \left|\hat{\mathcal{R}}(f_i) - \mathcal{R}(f_i)\right|\right] \\ &\leq 2G\epsilon + l_\infty \sqrt{\frac{2\log m(\epsilon)}{n}}\end{aligned} \tag{12}$$

Bounding the Expectation II

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

In addition, since $m(\epsilon) \sim \epsilon^{-d}$, ignoring constants, we need to balance $\epsilon + \sqrt{\frac{d \log(1/\epsilon)}{n}}$, which leads to, with a choice of ϵ proportional to $1/\sqrt{n}$, to a rate proportional

$$\sqrt{\frac{d \log n}{n}} \Rightarrow \sqrt{d/n}$$

Unfortunately, this often leads to a non-optimal dependence on dimension. One very powerful tool that avoids these undesired dependences on dimension is Rademacher complexities or Gaussian complexities.

Acknowledgment

Thank You!

- Here I would like to thank Professor Mao for his generous help and valuable advice during my prepartion.
- Meanwhile, I would also like to thank Tingyu Wang for pointing out my mistakes and typos in the slides.