CS 839 Probability and Learning in High Dimension Lecture 15 - 03/28/2022

Lecture 15: Statistical Learning

Lecturer: Yudong Chen *Scribe: Jenny Wei*

In this lecture, We will derive the estimation error in the learning task and introduce a Rademacher complexity based technique to upper bound it. We will also see one example as the application of this bound.[1]

1 Problem Set Up

Consider the following learning task. Let f^*: $\mathfrak{X} \to [0,1]$ being the unknown true regression function. We observe n data points $(x_1, f^*(x_1)), \ldots, (x_n, f^*(x_n))$, where the feature vectors x_i's are sampled i.i.d. from some unknown distribution μ. The goal is to estimate f^* given the data.

For a given function $f : \mathfrak{X} \to [0,1]$, define the population risk (a.k.a. test error):

$$\mathcal{L}(f) = \mathbb{E}_{x\sim\mu}\Big(f(x) - f^*(x)\Big)^2$$

and the empirical risk (a.k.a. training error):

$$\mathcal{L}_n(f) = \frac{1}{n}\sum_{i=1}^{n}\Big(f(x_i) - f^*(x_i)\Big)^2.$$

Ideally, we want to find the population risk minimizer $f_o = \arg\min_{f\in\mathcal{F}} \mathcal{L}(f)$, which is however not computable since f^* and μ. As a surrogate, we consider the empirical risk minimizer (ERM)

$$\hat{f} = \arg\min_{f\in\mathcal{F}} \mathcal{L}_n(f).$$

Our goal is to control the population risk of the ERM $\hat{f}$.

Risk decomposition: We can decompose the population risk of $\hat{f}$ as follows

$$\begin{aligned}
\underbrace{\mathcal{L}(\hat{f})}_{\text{test error}} &= \underbrace{\Big(\mathcal{L}(\hat{f}) - \mathcal{L}_n(\hat{f})\Big)}_{\text{generalization gap}} + \underbrace{\mathcal{L}_n(\hat{f})}_{\text{training error}} \\
&\le \Big(\mathcal{L}(\hat{f}) - \mathcal{L}_n(\hat{f})\Big) + \mathcal{L}_n(f_o) \\
&= \underbrace{\Big(\mathcal{L}(\hat{f}) - \mathcal{L}_n(\hat{f})\Big) + (\mathcal{L}_n(f_o) - \mathcal{L}(f_o))}_{\text{statistical error}} + \underbrace{\mathcal{L}(f_o)}_{\text{approximation error}},
\end{aligned}$$

where the inequality above holds since $\hat{f}$ minimizes $\mathcal{L}_n$. We have mentioned that $\mathcal{L}(\hat{f}) - \mathcal{L}_n(\hat{f})$ is called the **generalization gap/error**, which is the gap between the test and training error of $\hat{f}$. The first two terms

[1] *Reading:*

- Section 4.1 and 4.2 in [Wainwright, 2019]
- Section 8.4 in [Vershynin, 2018]
- Section 3.3 in [Duchi, 2021]

in the last line above are the differences between an empirical quantity (defined by $\mathcal{L}_n$) and its population counterpart (defined by $\mathcal{L}$). These two terms represent the **statistical/estimation error** due to having a finite number of data points. The last term $\mathcal{L}(f_o)$ measures how well the function class $\mathcal{F}$ can approximate the true function f^* under the real data distribution μ; this term represents the **approximation error**.

We can upper bound the statistical error error by the supremum of the difference:

$$\left(\mathcal{L}(\hat{f}) - \mathcal{L}_n(\hat{f})\right) + (\mathcal{L}_n(f_o) - \mathcal{L}(f_o)) \le 2 \sup_{f\in\mathcal{F}} |\mathcal{L}_n(f) - \mathcal{L}(f)|,$$

which leads to the bound

$$\underbrace{\mathcal{L}(\hat{f}) - \mathcal{L}(f_o)}_{\text{excess risk}} \le 2 \sup_{f\in\mathcal{F}} |\mathcal{L}_n(f) - \mathcal{L}(f)|.$$

In what follows, we establish upper bound on the above supremum using Rademacher complexity.

2 Upper Bound Using Rademacher Complexity

Assume $\mathcal{F}$ is $[0, 1]$-bounded, i.e. $\forall f \in \mathcal{F}$, $\forall x \in \mathcal{X}$: $f(x) \in [0, 1]$. Also assume $f^* \in \mathcal{F}$. Recall that we consider the mean square loss. The supremum above can be written as

$$\begin{aligned}\sup_{f\in\mathcal{F}} |\mathcal{L}_n(f) - \mathcal{L}(f)| &= \sup_{f\in\mathcal{F}} \Big|\frac{1}{n}\sum_{i=1}^{n} \underbrace{\Big(f(x_i) - f^*(x_i)\Big)^2}_{g(x_i)} - \mathbb{E}\Big(f(x) - f^*(x)\Big)^2\Big| \\ &= \sup_{g\in\mathcal{G}} \Big|\frac{1}{n}\sum_{i=1}^{n} g(x_i) - \mathbb{E}[g(x_i)]\Big|,\end{aligned}$$

where $\mathcal{G} \triangleq \left\{x \mapsto \Big(f(x) - f^*(x)\Big)^2\right\}$. Note that the quantity of the form $\sup - \mathbb{E}\sup$ can be bounded by concentration inequalities. Below we focus on bounding the expectation $\mathbb{E}\sup$.

2.1 Symmetrization

Let $(\epsilon_1, \ldots, \epsilon_n)$ be i.i.d Rademacher random variables. By the symmetrization argument in Theorem 8 of lecture 14, we have

$$\mathbb{E}\sup_{g\in\mathcal{G}} \Big|\frac{1}{n}\sum_{i=1}^{n}(g(x_i) - \mathbb{E}g(x_i))\Big| \le 2\mathbb{E}_x\mathbb{E}_\epsilon \sup_{g\in\mathcal{G}} \Big|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i g(x_i)\Big|.$$

We introduce some definitions. Define **empirical Rademacher complexity** of $\mathcal{G}$ as

$$\mathcal{R}_n(\mathcal{G}|x) \triangleq \mathbb{E}_\epsilon \sup_{g\in\mathcal{G}} \Big|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i g(x_i)\Big|,$$

and the **Rademacher complexity** of $\mathcal{G}$ as

$$\mathcal{R}_n(\mathcal{G}) \triangleq \mathbb{E}_x[\mathcal{R}_n(\mathcal{G}|x)].$$

Using these notations, we have established the following:

Theorem 1 (Symmetrization Bound)**.**

$$\mathbb{E}\sup_{g\in\mathcal{G}} \Big|\frac{1}{n}\sum_{i=1}^{n}\big(g(x_i) - \mathbb{E}g(x_i)\big)\Big| \le 2\mathcal{R}_n(\mathcal{G}) = 2\,\mathbb{E}_x[\mathcal{R}_n(\mathcal{G}|x)].$$

2.2 Contraction

Recall that $g(x_i) := \Big(f(x_i) - f^*(x_i)\Big)^2$, and that f and f^* are $[0,1]$-bounded. Also note that the square function $\phi : \theta \mapsto \theta^2$ is 2-Lipschitz over $[-1.1]$. By the Rademacher contraction principle in Theorem 7 from Lecture 14, we have

$$\begin{aligned}
\mathcal{R}_n(\mathcal{G}|x) &= \mathbb{E}_\epsilon \sup_{f\in\mathcal{F}} \Big| \frac{1}{n}\sum_{i=1}^n \epsilon_i \Big(f(x_i) - f^*(x_i)\Big)^2 \Big| \\
&= \mathbb{E}_\epsilon \sup_{f\in\mathcal{F}} \Big| \frac{1}{n}\sum_{i=1}^n \epsilon_i \phi\big(f(x_i) - f^*(x_i)\big) \Big| \\
&\leq 2\,\mathbb{E}_\epsilon \sup_{f\in\mathcal{F}} \Big| \frac{1}{n}\sum_{i=1}^n \epsilon_i \big(f(x_i) - f^*(x_i)\big) \Big| && \text{contraction principle} \\
&\leq 4\,\mathbb{E}_\epsilon \sup_{f\in\mathcal{F}} \Big| \frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i) \Big| && f^* \in \mathcal{F} \\
&= 4\mathcal{R}_n(\mathcal{F}|x).
\end{aligned}$$

It follows that $\mathcal{R}_n(\mathcal{G}) \leq 4\mathcal{R}_n(\mathcal{F})$.

2.3 Putting Together

Recapping the arguments above, we have

$$\begin{aligned}
\mathcal{L}_n(\hat{f}) - \mathcal{L}(f_o) &\lesssim \mathbb{E} \sup_{f\in\mathcal{F}} \Big| \mathcal{L}_n(f) - \mathcal{L}(f) \Big| && \text{risk decomposition} \\
&= \sup_{f\in\mathcal{F}} \Big| \frac{1}{n}\sum_{i=1}^n \Big(f(x_i) - f^*(x_i)\Big)^2 - \mathbb{E}\Big(f(x) - f^*(x)\Big)^2 \Big| \\
&= \sup_{g\in\mathcal{G}} \Big| \frac{1}{n}\sum_{i=1}^n \Big[g(x_i) - \mathbb{E}\, g(x_i)\Big] \Big| \\
&\lesssim \mathcal{R}_n(\mathcal{G}) && \text{symmetrization} \\
&\lesssim \mathcal{R}_n(\mathcal{F}) && \text{contraction principle} \\
&= \mathbb{E}_x \mathbb{E}_\epsilon \sup_{f\in\mathcal{F}} \Big| \frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i) \Big|.
\end{aligned}$$

We have upper bounded the supremum of one empirical process by that of another, and both processes are indexed by $f \in \mathcal{F}$. We are doing this because $\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_x[\mathcal{R}_n(\mathcal{F}|x)]$ is often easier to control. In particular, we can bound $\mathcal{R}_n(\mathcal{F}|x)$ conditioned on the data x. For fixed x, the quantity $\mathcal{R}_n(\mathcal{F}|x) = \mathbb{E}_\epsilon \sup_{f\in\mathcal{F}} |\frac{1}{n}\langle \epsilon, f(X)\rangle|$ is the supremum of a (canonical) Rademacher process. To control this supremum, we may use the following techniques:

- Union bound
- Dudley integral bound (e.g., when $\mathcal{F}$ is the set of Lipschitz functions; see Lecture 14 for details.)
- VC-dimension (usually used for binary functions; not covered in this course)
- Talagrand comparison: $\mathbb{E}_\epsilon \sup_{f\in\mathcal{F}} |\langle \epsilon, f(X)\rangle| \lesssim \mathbb{E}_{g\sim\mathcal{N}(0,I)} \sup_{f\in\mathcal{F}} |\langle g, f(X)\rangle|$. Then we can use any techniques for Gaussian process to bound the RHS.

In the following section, we give an example for bounding $\mathcal{R}_n(\mathcal{F})$ and $\mathcal{R}_n(\mathcal{F}|x)$ using union bound.

3 Example: Glivenk-Cantelli Uniform Law of Large Number (ULLN)

Let $x_1, \ldots, x_n$ be i.i.d random variables with distribution μ and Cumulative Distribution Function (CDF) $F(\theta) = \Pr[x_1 \le \theta] = \mathbb{E}[\mathbb{1}\{x_1 \le \theta\}]$.

We can estimate the true CDF F using empirical CDF, defined as

$$\hat{F}(\theta) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{x_1 \le \theta\} = \frac{1}{n}\sum_{i=1}^{n} g_\theta(x_i),$$

where we have Introduced the short hand $g_\theta(x) := \mathbb{1}\{x \le \theta\}$. Denote the set of such indicator functions by $\mathcal{G} \triangleq \{g_\theta : \theta \in \mathbb{R}\}$. Note that the functions g_θ are **not** Liptschitz.

We have

$$\begin{aligned}
\mathbb{E}\sup_{\theta\in\mathbb{R}}\left|\hat{F}(\theta) - F(\theta)\right| &= \mathbb{E}\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^{n} g(x_i) - \mathbb{E}\, g(x_1)\right| \\
&\le \mathcal{R}_n(\mathcal{G}) && \text{(Theorem 1)} \\
&= \mathbb{E}_x\, \mathcal{R}_n(\mathcal{G}|x) \\
&= \frac{1}{n}\,\mathbb{E}_x\,\mathbb{E}_\epsilon \sup_{\theta\in\mathbb{R}}\left|\sum_{i=1}^{n}\epsilon_i g_\theta(x_i)\right|.
\end{aligned}$$

Let us condition on fixed $x_1, \ldots, x_n$; assume w.l.o.g. that $x_1 \le x_2 \le \cdots \le x_n$. Note that the n-dimensional vector $\big(g_\theta(x_1), \ldots, g_\theta(x_n)\big) \in \{0,1\}^n$ can take on at most $n+1$ values:

$$\begin{gathered}
(0, 0, \ldots, 0) \\
(1, 0, \ldots, 0) \\
(1, 1, \ldots, 0) \\
\vdots \\
(1, 1, \ldots, 1)
\end{gathered}$$

Therefore, $\sup_{\theta\in\mathbb{R}}\left|\sum_i \epsilon_i g_\theta(x_i)\right|$ is the supremum of at most $(n+1)$ random variables. Moreover, for each $\theta \in \mathbb{R}$, the random variable $\epsilon_i g_\theta(x_i)$ is zero-mean and lies in the interval $\in [-1, 1]$. It follows that the sum $\sum_i \epsilon_i g_\theta(x_i)$ is a zero-mean, $O(n)$-sub-Gaussian random variable by Hoeffding inequality.

Using bound on the maximum of a finite number of sub-Gaussian random variables (Lecture 13, Lemma 3), we obtain

$$\mathbb{E}\sup_{\theta\in\mathbb{R}}\left|\sum_{i=1}^{n}\epsilon_i g_\theta(x_i)\right| \lesssim \sqrt{n\log(n+1)}.$$

Combining pieces, we obtain the following upper bound on the expectation:

$$\mathbb{E}\sup_{\theta\in\mathbb{R}}\left|\hat{F}(\theta) - F(\theta)\right| \le \sqrt{\frac{\log n}{n}}.$$

We can further use the Bounded Difference Inequality (Lecture 14, Theorem 4) to prove concentration around the expectation. Together, we obtain the following theorem:

Theorem 2. *With probability at least* $1 - e^{-n\delta^2}$,

$$\sup_{\theta\in\mathbb{R}}\left|\hat{F}(\theta) - F(\theta)\right| \le \sqrt{\frac{\log n}{n}} + \delta.$$

Hence $\sup_{\theta\in\mathbb{R}}\left|\hat{F}(\theta) - F(\theta)\right| \xrightarrow{a.s.} 0.$

Remark: We can remove the $\sqrt{\log n}$ factor using Dudley's integral bound and VC-dimension.

References

[Duchi, 2021] Duchi, J. (2021). Lecture notes in statistics 311/electrical engineering 377: Information theory and statistics. `http://web.stanford.edu/class/stats311/lecture-notes.pdf`.

[Vershynin, 2018] Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

[Wainwright, 2019] Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.