Stat 928: Statistical Learning Theory **Lecture: 10**

Empirical Process Theory and Oracle Inequalities

Instructor: Sham Kakade

1 Risk vs Risk

See Lecture 0 for a discussion on terminology.

2 The Union Bound / Bonferoni

Consider m events $E_1, \dots E_m$, we have $P(E_1 \cup \dots \cup E_m) \leq P(E_1) + \dots + P(E_m)$. In other words, with probability $1 - P(E_1) - \dots - P(E_m)$, none of the events E_i ($i = 1, \dots, m$) occurs.

If we assume the probability $\sum_j P(E_j)$ is small. Union bound is relatively tight when the events E_j are independent.

$$P(E_1 \cup \dots \cup E_m) \geq \sum_j P(E_j) - \sum_{j \neq k} P(E_j \cap E_k) \geq \sum_j P(E_j) - 0.5(\sum_j P(E_j))^2.$$

If E_j are correlated, then it is not tight. For example when they are completely correlated: $E_1 = \dots = E_m$, then

$$P(E_1 \cup \dots \cup E_m) = N^{-1} \sum_j P(E_j).$$

We will come back to this when we discuss chaining.

3 Motivation of Empirical Process

Consider learning problem with observations $Z_i = (X_i, Y_i)$, prediction rule $f(X_i)$ and loss function $L(f(X_i), Y_i)$. Assume further that f is parameterized by $\theta \in \Theta$ as $f_\theta(X_i)$.

Example, $f_\theta(x) = \theta^\top x$ be a linear function, and $L(f_\theta(x), y) = (\theta^\top x - y)^2$ is least squares loss. In the following, we introduce simplified notation $g_\theta(Z_i) = L(f_\theta(X_i), Y_i)$. We are interested in estimating $\hat{\theta}$ from training data. That is, $\hat{\theta}$ depends on Z_i.

Since we are using the training data as a surrogate of the test (true underlying) distribution, we hope training error is similar to test error. In learning theory, we are interested in estimating the following tail quantities for some $\epsilon > 0$:

$$P(n^{-1} \sum_{i=1}^n g_{\hat{\theta}}(Z_i) \geq E g_{\hat{\theta}}(Z) + \epsilon)$$

and

$$P(n^{-1} \sum_{i=1}^n g_{\hat{\theta}}(Z_i) \leq E g_{\hat{\theta}}(Z) - \epsilon).$$

The above two quantities can be bounded using the following two quantities:

$$P(n^{-1}\sum_{i=1}^{n} g_{\hat{\theta}}(Z_i) \geq Eg_{\hat{\theta}}(Z) + \epsilon) \leq P[\sup_{\theta\in\Theta}(n^{-1}\sum_{i=1}^{n} g_\theta(Z_i) - Eg_\theta(Z)) \geq \epsilon]$$

and

$$P(n^{-1}\sum_{i=1}^{n} g_{\hat{\theta}}(Z_i) \leq Eg_{\hat{\theta}}(Z) - \epsilon) \leq P[\sup_{\theta\in\Theta}(Eg_\theta(Z) - n^{-1}\sum_{i=1}^{n} g_\theta(Z_i)) \geq \epsilon].$$

Notation: in the above setting the collection of random variables $n^{-1}\sum_{i=1}^{n} g_\theta(Z_i)$ indexed by $\theta \in \Gamma$ is call an empirical process. We may also call $n^{-1}\sum_{i=1}^{n} g_\theta(Z_i) - Eg_\theta(Z)$ empirical process.

For each fixed θ, $n^{-1}\sum_{i=1}^{n} g_\theta(Z_i) - Eg_\theta(Z) \to 0$ in probability, by LLN. However, in empirical process, we are interested in uniform law of large numbers, that is the following supremum of empirical process defined as

$$\sup_{\theta\in\Theta} |n^{-1}\sum_{i=1}^{n} g_\theta(Z_i) - Eg_\theta(Z)|$$

converges to zero in probability. Given training data $Z_1^n = \{Z_1, \ldots, Z_n\}$, we may let $\hat{\theta}(Z_1^n)$ achieve the supremum above. Then

$$\sup_{\theta\in\Theta} |n^{-1}\sum_{i=1}^{n} g_\theta(Z_i) - Eg_\theta(Z)| = |n^{-1}\sum_{i=1}^{n} g_{\hat{\theta}(Z_1^n)}(Z_i) - Eg_{\hat{\theta}(Z_1^n)}(Z)|,$$

where $\hat{\theta}(Z_1^n)$ depends on the training data. This means that $\sum_{i=1}^{n} g_{\hat{\theta}(Z_1^n)}(Z_i)$ is not sum of independent random variable anymore. Supreme of empirical process is basically the worst case deviation of empirical mean (training error) and true mean (test error) for parameter θ that is chosen based on training data.

Conceptually, as long as you select $\hat{\theta}$ based on training data, you need to use empirical process and uniform law of large numbers. However, if you only consider fixed θ independent of training data, then you can use standard law of large numbers because $g_\theta(Z_i)$ are independent random variable.

4 Oracle Inequality for empirical risk minimization

Consider the empirical risk minimization algorithm:

$$\hat{\theta} = \arg\min_{\theta\in\Theta} \sum_{i=1}^{n} g_\theta(Z_i),$$

and the optimization parameter that minimizes the test error (with inifnite amount of data):

$$\theta_* = \arg\min_{\theta\in\Theta} Eg_\theta(Z).$$

We want to know how much worse is the test error performance of $\hat{\theta}$ compared to that of θ_*. Results of this flavor is referred to as oracle inequality.

We can obtain simple oracle inequality using ULLN of empirical process as follows. Assume that we have the tail bound for the empirical mean of $g_{\theta_*}(Z)$ as:

$$P(n^{-1}\sum_{i=1}^{n} g_{\theta_*}(Z_i) - Eg_{\theta_*}(Z) \geq \epsilon_1) \leq \delta_1(\epsilon_1)$$

Assume that we have the following uniform tail bound for empiricla process for some $\gamma \in [0, 1)$:

$$P(\sup_{\theta}[-n^{-1}\sum_{i=1}^{n} g_\theta(Z_i) + (1-\gamma)Eg_\theta(Z) + \gamma Eg_{\theta_*}(Z)] \geq \epsilon_2) \leq \delta_2(\epsilon_2)$$

Taking the union bound, we obtain with probability $1 - \delta_1(\epsilon_1) - \delta_2(\epsilon_2)$,

$$n^{-1}\sum_{i=1}^{n} g_{\theta_*}(Z_i) - Eg_{\theta_*}(Z) < \epsilon_1, \qquad [-n^{-1}\sum_{i=1}^{n} g_{\hat{\theta}}(Z_i) + (1-\gamma)Eg_{\hat{\theta}}(Z) + \gamma Eg_{\theta_*}(Z)] < \epsilon_2.$$

Since by definition, we have

$$n^{-1}\sum_{i=1}^{n} g_{\hat{\theta}}(Z_i) \leq n^{-1}\sum_{i=1}^{n} g_{\theta_*}(Z_i).$$

Therefore by adding the three inequalities:

$$(1-\gamma)Eg_{\hat{\theta}}(Z) + \gamma Eg_{\theta_*}(Z)] - Eg_{\theta_*}(Z) < \epsilon_1 + \epsilon_2.$$

That is, we have

$$Eg_{\hat{\theta}}(Z) < Eg_{\theta_*}(Z) + (1-\gamma)^{-1}(\epsilon_1 + \epsilon_2).$$

If Θ contains only finite number of functions: $N = |\Theta|$, then we can simply apply the union bound

$$\begin{aligned}
&P(\sup_{\theta}[-n^{-1}\sum_{i=1}^{n} g_\theta(Z_i) + (1-\gamma)Eg_\theta(Z) + \gamma Eg_{\theta_*}(Z)] \geq \epsilon) \\
\leq &\sum_{\theta\in\Theta} P([-n^{-1}\sum_{i=1}^{n} g_\theta(Z_i) + (1-\gamma)Eg_\theta(Z) + \gamma Eg_{\theta_*}(Z)] \geq \epsilon) \\
\leq &|\Theta| \sup_{\theta\in\Theta} P([-n^{-1}\sum_{i=1}^{n} g_\theta(Z_i) + (1-\gamma)Eg_\theta(Z) + \gamma Eg_{\theta_*}(Z) \geq \epsilon).
\end{aligned}$$

5 Recap: Oracle Inequality

Consider the empirical risk minimization algorithm:

$$\hat{\theta} = \arg\min_{\theta\in\Theta} \sum_{i=1}^{n} g_\theta(Z_i),$$

and the optimization parameter that minimizes the test error (with inifnite amount of data):

$$\theta_* = \arg\min_{\theta\in\Theta} Eg_\theta(Z).$$

If

$$P(n^{-1}\sum_{i=1}^{n} g_{\theta_*}(Z_i) - Eg_{\theta_*}(Z) \geq \epsilon_1) \leq \delta_1(\epsilon_1),$$

which means that the training error of the optimal parameter isn't much larger than test error.

Assume also that we have the following uniform tail bound for empiricla process for some $\gamma \in [0, 1)$:

$$P(\sup_{\theta}[-n^{-1}\sum_{i=1}^{n} g_\theta(Z_i) + (1-\gamma)Eg_\theta(Z) + \gamma Eg_{\theta_*}(Z)] \geq \epsilon_2) \leq \delta_2(\epsilon_2),$$

which means that the training error of an arbitrary inferior parametr isn't much smaller than its test error.

Then we have oracle inequality with probability $1 - \delta_1(\epsilon_1) - \delta_2(\epsilon_2)$,

$$Eg_{\hat{\theta}}(Z) < Eg_{\theta_*}(Z) + (1-\gamma)^{-1}(\epsilon_1 + \epsilon_2).$$

This means that the generalization performance of ERM isn't much worst than that of the optimal parameter.

6 Lower bracketing covering number

If Θ is infinite, then we can use the idea of covering number. There are different definitions. Let $G = \{g_\theta : \theta \in \Theta\}$ be the function class of the empirical process. $G_N = \{g_1(z), \ldots, g_N(z)\}$ is a ϵ-lower bracketing cover of G if for all $\theta \in \Theta$, there exists $j = j(\theta)$ such that

$$\sup_z [g_j(z) - g_\theta(z)] \leq 0 \qquad Eg_j(z) \geq Eg_\theta(z) - \epsilon.$$

The smallest cardinality $N_{LB}(G, \epsilon)$ of such G_N is called ϵ-lower bracketing covering number. Similarly one can define upper bracketing covering number. The logarithm of covering number is called entropy. We shall mention that the functions $g_j(z)$ may not necessarily be a function $g_\theta(z)$ for $\theta \in \Theta$.

Let $G(\epsilon/2)$ be an $\epsilon/2$ lower bracketing cover of G, then pick $j = j(\theta)$

$$\begin{aligned}
&\sup_\theta [-n^{-1}\sum_{i=1}^n g_\theta(Z_i) + (1-\gamma)Eg_\theta(Z) + \gamma Eg_{\theta_*}(Z)] \\
=&\sup_\theta [-n^{-1}[\sum_{i=1}^n g_\theta(Z_i) - \sum_{i=1}^n g_j(Z_i)] \\
&- n^{-1}\sum_{i=1}^n g_j(Z_i) + (1-\gamma)Eg_j(Z) + \gamma Eg_{\theta_*}(Z) + (1-\gamma)[-Eg_j(Z) + Eg_\theta(Z)]]|_{j=j(\theta)} \\
\leq &\sup_{j \in G(\epsilon/2)} [-n^{-1}\sum_{i=1}^n g_j(Z_i) + (1-\gamma)Eg_j(Z) + \gamma Eg_{\theta_*}(Z) + (1-\gamma)\epsilon/2].
\end{aligned}$$

Thus,

$$\begin{aligned}
&P(\sup_\theta [-n^{-1}\sum_{i=1}^n g_\theta(Z_i) + (1-\gamma)Eg_\theta(Z) + \gamma Eg_{\theta_*}(Z)] \geq \epsilon) \\
\leq &P(\sup_{j \in G(\epsilon/2)} [-n^{-1}\sum_{i=1}^n g_j(Z_i) + (1-\gamma)Eg_j(Z) + \gamma Eg_{\theta_*}(Z) + (1-\gamma)\epsilon/2] \geq \epsilon) \\
\leq &\sum_{j \in G(\epsilon/2)} P(-n^{-1}\sum_{i=1}^n g_j(Z_i) + Eg_j(Z) \geq \gamma(Eg_j(Z) - Eg_{\theta_*}(Z)) + 0.5(1+\gamma)\epsilon) \\
\leq &|G(\epsilon/2)| \sup_{j \in G(\epsilon/2)} P(-n^{-1}\sum_{i=1}^n g_j(Z_i) + Eg_j(Z) \geq \gamma(Eg_j(Z) - Eg_{\theta_*}(Z)) + 0.5(1+\gamma)\epsilon).
\end{aligned}$$

The summation bound with $\gamma > 0$ is a form of an idea in empirical referred to as peeling, and some times also called shell bounds. We will present a simple example below to illustrate the basic concepts.

7 A Simple Example

This example is to get you familiar with the intuitions and notations. We will consider more complex examples in future lectures, but the basic idea resembles this example.

Consider one dimensional classification problem, with $x \in [-1, 1]$ and $y \in \{\pm 1\}$. Assume that conditioned on x, the class label is given by $y = \epsilon(2I(x \geq \theta_*) - 1)$ for some unknown θ_*, with independent random noise $\epsilon \in \{-1, 1\}$, and $p = P(\epsilon = 1) > 0.5$. This means that the optimal Bayes classifier is $f_*(x) = 1$ when $x \geq \theta$ and $f_*(x) = -1$ when $x < \theta$, and the Bayes error is $1 - p$.

Since we don't know the true threshold θ_*, we can consider a family of classifiers $f_\theta(x) = 2I(x \geq \theta) - 1$, with θ to be learned from training data. Given sample $Z = (X, Y)$, the classifier error function for this classifier is

$$g_\theta(Z) = I(f_\theta(X) \neq Y).$$

Given training data $Z_1^n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, we can learn a threshold $\hat{\theta}$ using empirical risk minimization that finds θ by minimizing the training error:

$$\hat{\theta} = \arg\min_\theta \sum_{i=1}^n g_\theta(Z_i).$$

We want to know the generalization performance of $\hat{\theta}$ compared to the Bayes error. That is, to give an upper bound of

$$Eg_\theta(Z) - (1 - p).$$

We will examine the following few issues in order to understand what is going on:

- $1/\sqrt{n}$ convergence (using Chernoff bound) versus $1/n$ convergence (using refined Chernoff bound or Bennet).
- The role of peeling.

7.1 Bracketing cover of the function class

Given ϵ, and let $\theta_j = -1 + j\epsilon$ for $j = 1, \ldots, \lceil 2/\epsilon \rceil$. Let

$$g_j(z) = \begin{cases} 0 & \text{if } x \in [\theta_j - \epsilon, \theta_j] \\ g_{\theta_j}(z) & \text{otherwise,} \end{cases}$$

where $z = (x, y)$.

It follows that for any $\theta \in [-1, 1]$, if we let θ_j be the smallest j such that $\theta_j \geq \theta$, then we have $g_j(z) = 0 \leq g_\theta(z)$ when $x \in [\theta, \theta_j]$, and $g_j(z) = g_\theta(z)$ when $x \notin [\theta, \theta_j]$, where $z = (x, y)$. Moreover,

$$Eg_j(z) - Eg_\theta(z) = E_{x \in [\theta, \theta_j]} - g_\theta(z) \geq -\epsilon.$$

Note that since only the analysis depends on covering number, generally we can deisgn a covering number that depends on the truth θ_*, and may cover the space non-uniformly. This is not considered here.

7.2 Using Standard Chernoff bound without peeling

At θ_*, we have from Chernoff bound:

$$P(n^{-1}\sum_{i=1}^{n} g_{\theta_*}(Z_i) - Eg_{\theta_*}(Z) \geq \epsilon) \leq \exp(-2n\epsilon^2).$$

Alternatively, we say that with probability $1-\delta_1$:

$$\sum_{i=1}^{n} g_{\theta_*}(Z_i) - Eg_{\theta_*}(Z) < \epsilon_1 = \sqrt{\ln(1/\delta_1)/2n}.$$

Now we want to evaluate using lower bracking cover $G(\epsilon/2)$ as:

$$\begin{aligned}
&P(\sup_{\theta}[-n^{-1}\sum_{i=1}^{n} g_{\theta}(Z_i) + Eg_{\theta}(Z)] \geq \epsilon) \\
\leq &|G(\epsilon/2)| \sup_{j \in G(\epsilon/2)} P(-n^{-1}\sum_{i=1}^{n} g_j(Z_i) + Eg_j(Z) \geq 0.5\epsilon) \\
\leq &\lceil 4/\epsilon \rceil e^{-n\epsilon^2/2}.
\end{aligned}$$

We used $|G(\epsilon/2)| \leq \lceil 4/\epsilon \rceil$. Alternatively, we say that with probability $1-\delta_2$ (and note that $\epsilon_2 \geq \sqrt{2/n}$):

$$\sup_{\theta}[-n^{-1}\sum_{i=1}^{n} g_{\theta}(Z_i) + Eg_{\theta}(Z)] < \epsilon_2 = \sqrt{2(\ln\lceil 4/\epsilon_2 \rceil - \ln\delta_2)/n}. \leq \sqrt{2(\ln\lceil 4\sqrt{n/2} \rceil - \ln\delta_2)/n}.$$

Let $\delta = 2\delta_1 = 2\delta_2$, we have with probability at least $1-\delta$:

$$Eg_{\hat{\theta}}(Z) - (1-p) < \sqrt{\ln(2/\delta)/2n} + \sqrt{2(\ln\lceil 4\sqrt{n/2} \rceil + \ln(2/\delta))/n} < \sqrt{2\ln\lceil 4\sqrt{n/2} \rceil/n} + 3\sqrt{\ln(2/\delta)/2n}.$$