ORIE 7790 High Dimensional Probability and Statistics Lecture 24–25 - 05/07,12/2020

Lecture 24–25: Uniform Laws and Localization

Lecturer: Yudong Chen *Scribe: Lijun Ding*

In this lecture, we will establish certain uniform laws via the localization technique. These bounds enable us to get sharper bounds on the testing error based on bounds for the training error.

Relevant readings are the following:

- Wainwright's book: 14.1, 14.2 (optional: 14.3)
- Seminal paper by Bartlett et al: `https://arxiv.org/pdf/math/0508275.pdf`
- Recent paper using localization to analyze SDP methods: `https://arxiv.org/pdf/2004.01869.pdf` ("fixed point" there means "critical radius" in our terminology)
- For a more systematic treatment: Vladimir Koltchinskii, "Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems"

1 From Training Error to Test Error

Recall the setup for non-parametric regression (Lecture 14 -15):

$$y_i = f^*(x_i) + \sigma w_i.$$

Here $x_i \overset{\text{iid}}{\sim} \mathbb{P}$, $i = 1, \dots, n$ are the feature vectors, and $w_i \overset{\text{iid}}{\sim} N(0,1)$, $i = 1, \dots, n$ are the noise. The notation $\mathbb{P}$ denotes a certain probability distribution on the space $\mathcal{X}$ that x lives in. We assume that the function f^* is in some function class $\mathcal{F}$. For example, $\mathcal{F}$ may be the set of all real valued 1-Lipschitz functions on $[0,1]$ with function value 0 at 0.

We can obtain an estimate $\hat{f}$ of f^* via non-parametric least-squares:

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - f(x_i)\right)^2.$$

As explained in Lectures 14–15, this optimization problem is often tractable, e.g., when $\mathcal{F}$ is a certain parametric family, the Lipschitz function class, or the convex Lipschitz function class. We also derived bounds on the empirical error

$$\|\hat{f} - f^*\|_n^2 \triangleq \frac{1}{n} \sum_{i=1}^{n} \left(\hat{f}(x_i) - f^*(x_i)\right)^2 = \int_{\mathcal{X}} \left(\hat{f} - f^*\right)^2 \mathbb{P}_n(dx), \tag{1}$$

where the distribution $\mathbb{P}_n(x) \triangleq \sum_{i=1}^n \delta_{x_i}(x)$ is the empirical distribution of $\{x_i\}_{i=1}^n$.

The quantity in (1) is (a denoised version of) the training error. What we are really interested in is the test/population error:

$$\|\hat{f} - f^*\|_2^2 \triangleq \mathbb{E}\left(\hat{f}(X) - f^*(X)\right)^2 = \int_{\mathcal{X}} \left(\hat{f} - f^*(x)\right)^2 \mathbb{P}(dx). \tag{2}$$

Here X is considered as a fresh sample from $\mathbb{P}$ that is independent of the data (x_i, y_i), $i = 1, \dots, n$. We would like to see how to convert bounds on training error to test error. This requires some uniform laws over the function class $\mathcal{F}$ as $\hat{f}$ here correlates with the data (x_i, y_i), $i = 1, \dots, n$.

2 Uniform Law via Localization

We first explain why we want a uniform law of large numbers. If the $\hat{f}$ in (1) were a fixed $f \in \mathcal{F}$, letting $g = f - f^*$ we can use the strong law of large number to conclude that the training error $\|g\|_n$ in (1) converges to the testing error $\|g\|_2$ in (2) almost surely and in probability; that is,

$$\|g\|_n^2 \xrightarrow{\text{a.s. and p}} \|g\|_2^2.$$

Using Hoeffding's inequality we can further establish non-asymptotic high probability bounds on the difference between the two quantities. However, since $\hat{f}$ is not fixed but random and depends on the data (y_i, x_i), we need some uniform bound for all $f \in \mathcal{F}$. By recentering $\mathcal{F}$, we may assume $0 \in \mathcal{F}$. In this case, the results from Lec 12–13 established such a uniform bound using Rademacher complexity:

$$\sup_{g \in \mathcal{F}} \left| \|g\|_n^2 - \|g\|_2^2 \right| \lesssim R_n(\mathcal{F}) \triangleq \mathbb{E}_{\{x_i\}_{i=1}^n, \epsilon} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i f(x_i) \right|. \tag{3}$$

This time, we derive tighter bounds using localized Rademarcher complexity:

$$R_n(\delta; \mathcal{F}) \triangleq \mathbb{E}_{\{x_i\}_{i=1}^n, \epsilon} \sup_{f \in \mathcal{F}, \|f\|_2 \le \delta} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i f(x_i) \right|. \tag{4}$$

Note the extra constraint $\|f\|_2 \le \delta$ under the supremum.

2.1 Main Theorem

We shall assume the following two conditions.

A1. The function class $\mathcal{F}$ is star-shaped: $f \in \mathcal{F} \Rightarrow \alpha f \in \mathcal{F}, \forall \alpha \in [0, 1]$.

A2. The function class $\mathcal{F}$ is b-uniformly bounded: $\|f\|_\infty \le b$, $\forall f \in \mathcal{F}$.

Theorem 1. *Suppose that A1 and A2 hold. Let $\delta_n > 0$ be any solution to $R_n(\delta; \mathcal{F}) \le \frac{\delta^2}{b}$. Then*

$$\mathbb{P}\left(\forall f \in \mathcal{F} : \left| \|f\|_n^2 - \|f\|_2^2 \right| \le \frac{1}{2}\|f\|_2^2 + \frac{\delta_n}{2} \right) \ge 1 - c\exp(-cn\delta_n^2/b^2). \tag{5}$$

If in addition $n\delta_n^2 \ge \frac{2}{c} \log\log \frac{1}{\delta_n}$, then

$$\mathbb{P}\left(\forall f \in \mathcal{F} : \left| \|f\|_n^2 - \|f\|_2^2 \right| \le c_0\delta_n \right) \ge 1 - c'\exp(-c'n\delta_n^2/b). \tag{6}$$

The constants c, c' in the theorem above are universal constants. We defer the proof to Section 4.

We remark that same bound holds with the localized empirical Redemacher complexity:

$$\hat{R}_n(\delta, \mathcal{F}) \triangleq \mathbb{E}_{\{\epsilon_i\}_{i=1}^n} \sup_{f \in \mathcal{F}, \|f\|_n \le \delta} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i f(x_i) \right|. \tag{7}$$

The reason is that the minimal solutions $\hat{\delta}, \delta_n$ to $\hat{R}_n(\delta, \mathcal{F}) \le \frac{\delta^2}{b}$ and $R_n(\delta, \mathcal{F}) \le \frac{\delta^2}{b}$, respectively, satisfy $\hat{\delta} \in [c\delta_n, C\delta_n]$ with probability at least $1 - \exp(-c_1 n\delta_n^2/b)$ for some universal $c, c_1, c_2, C > 0$. See the appendix of Chapter 14 of Wainwright's book.

The significance of Theorem 1 might be understood via the following examples.

3 Examples

We shall illustrate the use of Theorem 1 in the examples of quadratic function class and convex regression.

3.1 Example 1: Quadratic Function Class

Consider the function class

$$\mathcal{F}_2 \triangleq \left\{ f_\theta : [-1,1] \to \mathbb{R} \text{ with } x \mapsto \theta_0 + \theta_1 x + \theta_2 x^2,\, \theta \in \mathbb{R}^3,\, \|f_\theta\|_\infty \le 1 \right\}.$$

We assume the feature X follows the uniform distribution on $[-1,1]$ denoted as $\mathbb{P}$.

Without localization: First, we derive a bound without localization. By Dudley's Entropy Integral bound, we have

$$\hat{R}_n(\delta, \mathcal{F}) \le \hat{R}_n(\mathcal{F}) \lesssim \frac{1}{\sqrt{n}}, \quad \forall \{x_i\},$$

where $\hat{R}_n(\mathcal{F})$ is the empirical Radamacher complexity. The detailed proof is left as an exercise. Hence, using Theorem 1, w.h.p.:

$$\left| \|f\|_n^2 - \|f\|_2^2 \right| \lesssim \|f\|_2^2 + \frac{1}{\sqrt{n}}, \quad \forall f \in \mathcal{F}_2.$$

We get a $1/\sqrt{n}$ rate, often known as a "slow rate".

With localization: Next, we use the localization technique. We first reparametrize $\mathcal{F}_2$ using orthonormal (Legendre) basis:

$$\phi_0(x) = \frac{1}{\sqrt{2}}, \quad \phi_1(x) = \sqrt{\frac{3}{2}} x, \quad \phi_2(x) = \sqrt{\frac{5}{8}} \left(3x^2 - 1\right),$$

which satisfies $\langle \phi_j, \phi_k \rangle \triangleq \int_{-1}^1 \phi_j(x)\phi_k(x)dx = \begin{cases} 1 & j = k, \\ 0 & j \neq k. \end{cases}$ Any function in $\mathcal{F}_2$ has expansion $f_\gamma(x) = \gamma_0\phi_0(x) + \gamma_1\phi_1(x) + \gamma_2(x)$ for some $\gamma = (\gamma_0, \gamma_1, \gamma_2) \in \mathbb{R}^3$, with $\|f_\gamma\|_2 = \|\gamma\|_2$.

Define the (random) feature matrix $M = [\phi_j(x_i)] \in \mathbb{R}^{n \times 3}$. Let us compute the localized Rademacher complexity:

$$\begin{aligned} R_n\left(\delta; \mathcal{F}\right) &= \mathbb{E}_{\{x_i, \epsilon_i\}_{i=1}^n} \sup_{f_\gamma \in \mathcal{F}, \|f\|_2 \le \delta} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i f_\gamma(x_i) \right| \\ &\overset{(a)}{\le} \mathbb{E} \sup_{\|\gamma\| \le \delta} \left| \frac{1}{n} \epsilon^\top M \gamma \right| \\ &\overset{(b)}{\le} \frac{\delta}{n} \mathbb{E} \|\epsilon^2 M\|_2 \\ &\overset{(c)}{\le} \frac{\delta}{n} \sqrt{\mathbb{E}\|\epsilon^\top M\|_2^2}. \end{aligned}$$

The first step (a) is due to orthogonality of γ and we actually enlarge the function class a bit as we don't require $\|f_\gamma\|_\infty \le 1$ in this step. The second step (b) is due to Cauchy-Schwarz. The last step (c) is due to Jensen's inequality.

But

$$\mathbb{E}\|\epsilon^\top M\|_2^2 = \mathbb{E}_{\{x_i\}_{i=1}^n} \mathrm{Tr}\left(M^\top \mathbb{E}_\epsilon \epsilon\epsilon^\top M\right) \overset{\mathbb{E}\epsilon\epsilon^\top = I}{=} \mathbb{E}_{\{x_i\}_{i=1}^n} \mathrm{Tr}\left(M^\top M\right) \overset{\text{orthonormality}}{=} 3n.$$

So $R_n(\delta; \mathcal{F}_2) \lesssim \frac{\delta}{\sqrt{n}}$ and the critical radius δ_n can be chosen as $\delta_n \asymp \frac{1}{\sqrt{n}}$. Hence, using Theorem 1, we find that with high probability:

$$\left| \|f\|_n^2 - \|f\|_2^2 \right| \lesssim \|f\|_2^2 + \frac{1}{n}.$$

We get a "fast rate" of $1/n$.

3.2 Example 2: Convex Regression

Recall the setup of convex regression:

$$y_i = f^*(x_i) + \sigma w_i,$$

where $x_i \overset{\text{iid}}{\sim} \mathbb{P}$, $w_i \overset{\text{iid}}{\sim} N(0,1)$, and the function f^* is in the set of all 1-Lipschitz convex functions $\mathcal{F}_{\text{conv}}$:

$$\mathcal{F}_{\text{conv}} = \{f : [0,1] \to \mathbb{R},\ f(0) = 0,\ f \text{ is convex and 1-Lipschitz}\}.$$

In Lecture 14-15, we showed $\|\hat{f} - f^*\|_n^2 \leq \left(\frac{1}{n}\right)^{4/5}$. Here we show that $\|\hat{f} - f^*\|_n$ is close to $\|\hat{f} - f^*\|_2$. To apply our theorem, we shall recenter $\mathcal{F}$ and consider

$$\mathcal{F}^*_{\text{conv}} \triangleq \mathcal{F}_{\text{conv}} - f^*.$$

The centered function class $\mathcal{F}^*_{\text{conv}}$ is a star-shaped and 2-uniformly bounded. By chaining and Dudley (see Lemma 3 in Lecture 14-15), we have

$$\hat{R}\left(\delta, \mathcal{F}^*_{\text{conv}}\right) \leq \frac{64}{\sqrt{n}} \int_{\frac{\delta^2}{4}}^{\delta} \sqrt{\log \underbrace{N\left(t, \mathbb{B}_n(\delta, \mathcal{F}^*_{\text{conv}}), \|\cdot\|_n\right)}_{N_\delta(t)}} dt + \frac{\delta^2}{4}. \tag{8}$$

Using the fact that $\|\cdot\|_n \leq \|\cdot\|_\infty$, we have

$$\log N_\delta(t) \leq \log N\left(t, \mathbb{B}_n(\delta, \mathcal{F}^*_{\text{conv}}), \|\cdot\|_\infty\right) \overset{(a)}{\lesssim} \sqrt{\frac{1}{t}}. \tag{9}$$

The step (a) can be found in Lecture 14-15, Section 6.4. Thus by combining (8) and (9), we find that

$$\hat{R}\left(\delta, \mathcal{F}^*_{\text{conv}}\right) \leq \frac{64}{\sqrt{n}} \int_{\frac{\delta^2}{4}}^{\delta} \left(\frac{1}{t}\right)^{\frac{1}{4}} dt + \frac{\delta^2}{4} = \frac{64 \times 4}{3\sqrt{n}} \delta^{\frac{3}{4}} + \frac{\delta^2}{4}.$$

Hence, we conclude that $\hat{\delta_n} \asymp \left(\frac{1}{n}\right)^{\frac{2}{5}}$ solves $\hat{R}_n\left(\delta; \mathcal{F}^*_{\text{conv}}\right) \leq \frac{\delta^2}{2}$.

By Theorem 1 applied to $\hat{f} - f^* \in \mathcal{F}^*_{\text{conv}}$, we have w.h.p.

$$\left| \|\hat{f} - f^*\|^2 - \|\hat{f} - f^*\|_n^2 \right| \leq \frac{1}{2}\|\hat{f} - f^*\|_2^2 + c\left(\frac{1}{n}\right)^{\frac{4}{5}}$$

$$\Rightarrow \|\hat{f} - f^*\|_2^2 \leq 2\|\hat{f} - f^*\|_n^2 + 2c\left(\frac{1}{n}\right)^{\frac{4}{5}}.$$

Recall that the training error satisfies $\|\hat{f} - f^*\|_n^2 \leq \left(\frac{1}{n}\right)^{4/5}$. Hence training error and test error are of the same order:

$$\|\hat{f} - f^*\|_2^2 \lesssim \left(\frac{1}{n}\right)^{\frac{4}{5}}.$$

This bound is in fact minimax optimal.

4 Proof of Theorem 1

By re-scaling of the function f and the function class, we may assume without loss of generality that $b = 1$, and that δ_n solves $R_n(\delta; \mathcal{F}) \leq \frac{\delta}{16}$.

For each $r \in [0,1]$, define the ball

$$\mathbb{B}_2(r; \mathcal{F}) \triangleq \{f \in \mathcal{F} : \|f\|_2 \leq r\},$$

and the random variable

$$Z(r) \triangleq \sup_{f \in \mathbb{B}_2(r,\mathcal{F})} \left| \|f\|_2^2 - \|f\|_n^2 \right|.$$

Define the "bad event":

$$\mathcal{E} \triangleq \left\{ \exists f \in \mathcal{F} : \left| \|f\|_2^2 - \|f\|_n^2 \right| > \frac{1}{2}\|f\|_2^2 + \frac{\delta_n^2}{2} \right\},$$

and the auxiliary event:

$$A(r) \triangleq \left\{ Z(r) \geq \frac{r^2}{2} \right\}.$$

Claim 1. $\mathcal{E} \subset A(\delta_n)$.

Proof Indeed, since event $\mathcal{E}$ happens, we know there is some $f \in \mathcal{F}$ such that $\| \left| \|f\|_2^2 - \|f\|_n^2 \right| \geq \frac{1}{2}\|f\|_2^2 + \frac{\delta_n^2}{2} \|$. Consider the following two cases based on $\|f\|_2$ and δ_n:

(i) This f satisfies $\|f\|_2 \leq \delta_n$, and so $f \in \mathbb{B}_2(\delta_n, \mathcal{F})$. Since f also satisfies that $\| \left| \|f\|_2^2 - \|f\|_n^2 \right| \geq \frac{1}{2}\|f\|_2^2 + \frac{\delta_n^2}{2} \|$, we know that the event $A(\delta_n)$ happens.

(ii) Otherwise, we should have $\|f\|_2 > \delta_n$. Now we may scale f and consider

$$\tilde{f} \triangleq \frac{\delta_n}{\|f\|_2} f \in \mathcal{F} \quad \text{as } \mathcal{F} \text{ is star-shaped.}$$

This $\tilde{f}$ satisfies $\|\tilde{f}\| = \delta_n$ by definition. Combining the properties of $\tilde{f}$ with $\| \left| \|f\|_2^2 - \|f\|_n^2 \right| \geq \frac{1}{2}\|f\|_2^2 + \frac{\delta_n^2}{2}$. We find that $\left| \|\tilde{f}\|_2^2 - \|\tilde{f}\|_n^2 \right| > \frac{1}{2}\delta_n^2$. Hence $A(\delta_n)$ happens again in this case.

In both cases, the event $A(\delta_n)$ happens as claimed. □

Thus, the theorem is proved if we can establish the following bound on the probability of event $A(\delta_n)$:

$$\mathbb{P}\left(A(\delta_n)\right) = \mathbb{P}\left(Z(\delta_n) \geq \frac{\delta_n^2}{2} \right) \leq 2\exp(-cn\delta_n^2).$$

The proof follows a standard procedure: first bound the expectation of $Z(r)$ and then show concentration.

Expectation We bound the expectation via symmetrization:

$$\begin{aligned} \mathbb{E}\left[Z_n(\delta_n)\right] &\leq 2\mathbb{E}\left[\sup_{f \in \mathbb{B}_2(\delta_n,\mathcal{F})} \left| \frac{1}{n}\sum_{i=1}^n f^2(x_i) \right| \right] \\ &\overset{(a)}{\leq} \mathbb{E}\left[\sup_{f \in \mathbb{B}_2(\delta_n,\mathcal{F})} \left| \frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i) \right| \right] \\ &= 4R_n(\delta_n) \\ &\overset{(b)}{\leq} \frac{\delta_n^2}{4}. \end{aligned} \tag{10}$$

Here in step (a), we uses the fact that $\mathcal{F}$ is 1-uniformly bounded as well as the Ledoux-Talagrand contraction principle. We use the definition of δ_n in the last step (b).

Concentration Take an arbitrary $f \in \mathbb{B}_2(\delta_n, \mathcal{F})$. We have $\|f\|_\infty \leq 1$, and so

$$\| \underbrace{f^2 - \mathbb{E}[f^2(X)]}_{\triangleq g} \|_\infty \leq 1.$$

Here X follows the distribution of the features x_i. The variance of g is bounded by

$$\mathrm{Var}(g) \leq \mathbb{E}(f^4(X)) \overset{(a)}{\leq} \mathbb{E}[f^2] = \|f\|_2^2 \leq \delta_n^2.$$

The step (a) holds because $\|f\|_\infty \leq 1$. We shall make use of Talagrand's Functional Bernstein inequality. This is a refinement of Functional Hoeffding introduced in Lecture 5-6 Theorem 5.

Theorem 2 (Talagrand's Functional Bernstein)**.** *Suppose $\mathcal{G}$ is b-uniformly bounded and $Z \triangleq \sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^n g(X_i)$ for some iid X_i. Then*

$$\mathbb{P}\left(Z \geq \mathbb{E}[Z] + t\right) \leq 2\exp\left(\frac{-nt^2}{8e\mathbb{E}[\Sigma^2] + 4bt}\right),$$

where $\Sigma^2 = \sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^n g^2(X_i)$. Moreover,

$$\mathbb{E}(\Sigma^2) \leq \sup_{g\in\mathcal{G}} \mathbb{E}(g^2) + 2b\mathbb{E}(Z).$$

Applying Functional Bernstein with $t = \frac{\delta_n^2}{4}$ yields

$$\mathbb{P}\left(Z(\delta_n) \geq \mathbb{E}[Z(\delta_n)] + \frac{\delta_n^2}{4}\right) \leq 2\exp\left(\frac{-n\delta_n^4}{c(\delta_n^2 + \delta_n^2 + \delta_n^2)}\right) = 2e^{-cn\delta_n^2}.$$

Combining expectation and concentration bounds proves the first inequality (5) in Theorem 1.

See Wainwright's book for the second inequality (6) in Theorem 1.