EECS 598-005: Theoretical Foundations of Machine Learning **Fall 2015**

Lecture 19: Fast Rates 1

Lecturer: Matus Telgarsky, Scribes: Cupjin Huang, Editors: Biaoshuai Tai, Wei Lee, Shang-En Huang

Notations Notations in this lecture are pretty convoluted. In general, quantities with $\hat{}$ indicates the empirical version of the corresponding quantity without $\hat{}$. Given a loss function ℓ, R_ℓ denote the expected loss under a certain distribution. For example, for 0-1 loss z, R_z is the expected 0-1 loss while $\hat{R}_z$ is the empirical loss with n i.i.d. samples. Sometimes the number of samples n will be implicit and may appear in the formula without mentioning. d stands for the VC-dimension of an implicit function class $\mathcal{F}$.

In the latter part of the lecture we are going to use $\mathfrak{R}$ to denote empirical Rademacher complexity, be careful not to confuse this with the expected loss R.

19.1 Overview and Definitions

19.1.1 Intuition of Fast Rate

In PAC learning setting where there exists a perfect classifier f^*, We know that with probability at least $1-\delta$,

$$R_z(f_{\text{ERM}}) - \hat{R}_z(f_{\text{ERM}}) = O\left(\frac{d\ln(n/\delta)}{n}\right) = \tilde{O}\left(\frac{1}{n}\right), \tag{19.1}$$

where $f_{\text{ERM}} = \operatorname{argmin}_{f\in\mathcal{F}} \hat{R}_z(f)$ is a function minimizing the empirical risk, or equivalently, consistent on the given samples.

On the other hand, however, the upper bound of the difference might become much looser when no perfect classifier exists. Recall that in Rademacher analysis, the result we get is with probability at least $1-\delta$,

$$\sup_{f\in\mathcal{F}} |R_z(f) - \hat{R}_z(f)| = O\left(\sqrt{\frac{d\ln(n/\delta)}{n}}\right) = \tilde{O}\left(\frac{1}{\sqrt{n}}\right). \tag{19.2}$$

There are two main differences of these two bounds: Firstly, Eqn. (19.1) only holds for the empirical risk minimizers, but Eqn. (19.2) is a bound for the worst case so it holds for all functions in the class $\mathcal{F}$; Secondly, Eqn. (19.1) provides a much faster "convergence" speed when the number of samples increases, thus might be more feasible for practical use. The phenomenon that for certain learning problems, the difference between the risk and the empirical risk has a much faster convergence rate than $\tilde{O}(1/\sqrt{n})$, is often referred to as "fast rate".

19.1.2 What is going on?

To get an intuition of why we can get "fast" rate in the first case, consider the following problem where X is a random variable in $[0,1]$. Then for $p := \mathbb{E}[X]$ and $\hat{p} := \frac{1}{n}\sum_{i=1}^n X_i = \hat{\mathbb{E}}[X]$, we can get two different deviation bounds for $|p-\hat{p}|$. First, by Hoeffding's Inequality, the following holds with probability at least $1-\delta$:

$$|p-\hat{p}| \le \sqrt{\frac{2\ln(1/\delta)}{n}} = \tilde{O}\left(\frac{1}{\sqrt{n}}\right).$$

Note that this inequality doesn't depend on the probability p. If we know more about p, then the following bound by Bernstein might be tighter:

$$|p - \hat{p}| \leq \sqrt{\frac{2p(1-p)\ln(n/\delta)}{n}} + \frac{\ln(n/\delta)}{n}.$$

In particular, when p is sufficiently small (or equivalently, sufficiently large) such that $p(1-p) \leq \frac{1}{n}$, then we can get an $\tilde{O}\left(1/n\right)$ bound. In fact, here's a somewhat general principle, which holds for almost all cases: "easy" problems usually have fast rates.

Remark One common proof technique of fast rate is to incorporate the variance term, such as $p(1-p)$ term here.

Another approach to look at this problem is to approximate binomial distribution. For a random variable $Y \in \{0, 1\}$, we have

$$\begin{aligned} \Pr\left[\hat{\mathbb{E}}[Y] \geq \epsilon\right] &= \Pr[\text{Binomial}(n, \mathbb{E}[Y]) \geq n\epsilon] \\ &= \sum_{k \geq n\epsilon} \binom{n}{k} \left(\mathbb{E}[Y]\right)^k \left(1 - \mathbb{E}[Y]\right)^{n-k} \end{aligned}$$

The rule of thumb to approximate binomial distribution is as follows:

- If $n\,\mathbb{E}[Y] \leq 1$, it can be approximated using Poisson random variable with parameter $n\,\mathbb{E}[Y]$;
- Otherwise, it can be approximated using Gaussian distribution.

Intuitively, fast rates come when approximation with Poisson random variables is available.

19.1.3 Relative VC bound

A tighter version of VC bound takes in the error rate into consideration, which allows us to get a Bernstein-like inequality. More specifically, with probability $\geq 1 - \delta$, for all $f \in \mathcal{F}$,

$$R_z(f) - \hat{R}_z(f) = O\left(\sqrt{\hat{R}_z(f)\frac{d\ln(n/\delta)}{n}} + \frac{d\ln(n/\delta)}{n}\right).$$

Remarks

- This equation is asymmetric with respect to $R_z(f)$ and $\hat{R}_z(f)$.
- If we know nothing about $\hat{R}_z(f)$, then this bound becomes $\tilde{O}(1/\sqrt{n})$. However, if $\hat{R}_z(f)$ is sufficiently small, say, if $\hat{R}_z(f) = 0$, then we can directly recover the rate for PAC learning.
- Note that the value $\hat{R}_z(f)$ appearing in the right hand side is still a random variable, which means that sometimes we are not done yet.

We shall not go through the proof in this class, however it is worth mentioning that the main trick applied in that proof is to control the term

$$\Pr\left[\sup_{f \in \mathcal{F}} \frac{R_z(f) - \hat{R}_z(f)}{\sqrt{\hat{R}_z(f)}} \leq \epsilon\right]$$

with some fuzz for the corner case $\hat{R}_z(f) = 0$. The corner case is easy to deal with as we can add an arbitrarily small constant on the denominator to avoid the problem.

In the next class, we are going to prove the following theorem:

Theorem 19.1. *Given*

- Linear predictors $\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle$ *with* $\|\mathbf{w}\|_2 \leq B$ *for some radius* B*.*
- *Univariate convex differentiable* loss function $\ell : \mathbb{R} \to \mathbb{R}$
- Excess risk $\epsilon_\ell(\mathbf{w}) := R_\ell(\mathbf{w}) - \inf_{\|\mathbf{w}'\| \leq B} R_\ell(\mathbf{w}')$ *where* $R_\ell(\mathbf{w}) := \mathbb{E}[\ell(y\langle \mathbf{w}, \mathbf{x} \rangle)]$ *satisfies*

$$\epsilon_\ell(\mathbf{w}) \geq \frac{\lambda}{2}\|\mathbf{w} - \overline{\mathbf{w}}\|_2^2$$

for some $\lambda > 0$*, where* $\overline{\mathbf{w}} \in \operatorname{argmin}_{\mathbf{w}} R_\ell(\mathbf{w})$ *is the minimizer and* y *is what nature reveals.*

Then we have with probability at least $1 - \delta$

$$\epsilon_\ell(\mathbf{w}) \leq 2\hat{\epsilon}_\ell(\mathbf{w}) + O\left(\frac{\ell'(B)\ln(n/\delta)}{\lambda n}\right),$$

where $\hat{\epsilon}_\ell(\mathbf{w}) := \hat{R}_\ell(\mathbf{w}) - \inf_{\|\mathbf{w}'\| \leq B} \hat{R}_\ell(\mathbf{w}')$ *and* ℓ' *is the derivative of* ℓ*.*

Remarks

- One may wonder why is there a coefficient 2 in front of $\hat{\epsilon}_\ell(\mathbf{w})$: In fact this doesn't matter. The goal of a learning algorithm is to find $\mathbf{w}$ such that $\hat{\epsilon}_\ell(\mathbf{w})$ goes to 0. With this in mind, the bound here is essentially a fast rate bound.
- One research topic that might be interesting is that it is still not fully understood when fast rates are possible.
- Recall strong convexity: we call that a function f is λ-strongly convex if for all $\mathbf{w}, \mathbf{w}'$ we have

  $$f(\mathbf{w}) \geq f(\mathbf{w}') + \langle \nabla f(\mathbf{w}'), \mathbf{w} - \mathbf{w}' \rangle + \frac{\lambda}{2}\|\mathbf{w} - \mathbf{w}'\|_2^2.$$

  The difference between the property of $\epsilon_\ell(\cdot)$ here and strong convexity is that 1) this only holds for the single point $\overline{\mathbf{w}}$ while strong convexity is a global property and 2) there is no linear term.

  However, it is worth mentioning that the fast rate result here is somehow related to the *stochastic gradient descent* (SGD) algorithm. Using SGD on a general convex function will have a convergence rate about $\tilde{O}(1/\sqrt{t})$, while if the convex function is required to be strongly convex then the convergence rate becomes $\tilde{O}(1/t)$.

19.2 Lemmas for the Proof of the Theorem

We will not prove this theorem in this course. We are going to go through some lemmas which are used in the proof. It is also a good chance to get a review of Rademacher complexity.

We use $\mathfrak{R}(V) := \hat{\mathfrak{R}}_S(V)$ here to denote the empirical Rademacher complexity of a certain function class V on an implicitly given sample S. Denote $n := |S|$, then each function is represented by only the n values with input chosen from S, so we can regard V as a subset of $\mathbb{R}^n$. In fact, we are not going to care about what the function class is or what the sample is, just think of V as a subset of $\mathbb{R}^n$. Note that, the following lemmas still hold when we take expectation for each empirical Rademacher complexity term, so they automatically hold for the case where $\mathfrak{R}$ stands for the Rademacher complexity.

19.2.1 Lemmas

Lemma 19.2. *Given $c \geq 0$. For all $V \subseteq \mathbb{R}^n$ and $\mathbf{v}_0 \in \mathbb{R}^n$*

$$\mathfrak{R}(cV + \{\mathbf{v}_0\}) = c \cdot \mathfrak{R}(V).$$

Lemma 19.3. *Given countably many sets $(V_j)_{j \geq 1}$ with $\mathbf{0} \in V_j$ for all j,*

$$\mathfrak{R}\Big(\bigcup_{j\geq 1} V_j\Big) \leq \sum_{j \geq 1} \mathfrak{R}(V_j).$$

Lemma 19.4. *Given $W := \{\mathbf{z} \mapsto \langle \mathbf{w}, \mathbf{z} \rangle, \|\mathbf{w}\|_2 \leq B\}$ and inputs $\|\mathbf{z}_i\|_2 \leq X$ with some constant X, we have*

$$\mathfrak{R}(W) \leq \frac{BX}{\sqrt{n}}.$$

Lemma 19.5. *Given $g : \mathbb{R} \to \mathbb{R}$ L-Lipschitz, then*

$$\mathfrak{R}(g \circ V) \leq L\mathfrak{R}(V).$$

19.2.2 Proofs of Lemmas

Proof of Lemma 19.2:

$$\begin{aligned}
\mathfrak{R}(cV + \{\mathbf{v}_0\}) &= \mathop{\mathbb{E}}_{\sigma}\left[\sup_{\mathbf{v}\in V} \frac{c}{n}\sum_i \sigma_i(\mathbf{v})_i + \frac{1}{n}\sum_i \sigma_i(\mathbf{v}_0)_i\right] \\
&= \mathop{\mathbb{E}}_{\sigma}\left[\frac{c}{n}\sup_{\mathbf{v}\in V}\sum_i \sigma_i(\mathbf{v})_i\right] + \mathop{\mathbb{E}}_{\sigma}\left[\frac{1}{n}\sum_i \sigma_i(\mathbf{v}_0)_i\right] \\
&= c\mathfrak{R}(V), \qquad \text{(observing that } \mathbb{E}_\sigma\left[\tfrac{1}{n}\textstyle\sum_i \sigma_i(\mathbf{v}_0)_i\right] = 0.)
\end{aligned}$$

where $(\mathbf{v})_i$ is the i-th coordinate of $\mathbf{v}$.

■

Historical Remark The original definition of Rademacher complexity was $\mathfrak{R}(V) := \mathbb{E}_\sigma[\sup_{\mathbf{v}\in V} |\frac{1}{n}\sigma_i(\mathbf{v})_i|]$. If we are using this definition of Rademacher complexity, then Lemma 19.2 does not hold anymore.

The big problem of this definition is $\mathfrak{R}(\{\mathbf{v}_0\})$ may not be necessarily equal to 0 in general, so it fails to capture the "complexity" of a function class in some sense. However, there are still papers using this definition, so be careful not to refer to results using different definitions for Rademacher complexity.

Proof of Lemma 19.4: Given samples $(\mathbf{z}_i)_{i=1}^n$, set $\mathbf{Z} = \begin{bmatrix}\mathbf{z}_1 & \mathbf{z}_2 & \cdots & \mathbf{z}_n\end{bmatrix} \in \mathbb{R}^{n\times n}$. Then

$$\begin{aligned}
n\mathfrak{R}(W) &= \mathop{\mathbb{E}}_{\sigma}\left[\sup_{\mathbf{w}\in W} \mathbf{w}^\top \mathbf{Z}\sigma\right] \\
&\leq \mathop{\mathbb{E}}_{\sigma}\left[\sup_{\mathbf{w}\in W} \|\mathbf{w}\|_2 \|\mathbf{Z}\sigma\|_2\right] \\
&\leq B\mathop{\mathbb{E}}_{\sigma}\left[\|\mathbf{Z}\sigma\|_2\right] \\
&= B\mathop{\mathbb{E}}_{\sigma}\left[\sqrt{\|\mathbf{Z}\sigma\|_2^2}\right] \\
&\leq B\sqrt{\mathop{\mathbb{E}}_{\sigma}\left[\|\mathbf{Z}\sigma\|_2^2\right]},
\end{aligned}$$

where the second line is due to Cauchy-Schwarz, the third line is due to that $\|\mathbf{w}\| \le B$ for all $\mathbf{w} \in W$ and the fifth line is due to Jensen's Inequality. To bound $\mathbb{E}_\sigma[\|\mathbf{Z}\sigma\|_2^2]$, observe that

$$\begin{aligned}\mathop{\mathbb{E}}_{\sigma}\left[\|\mathbf{Z}\sigma\|_2^2\right] &= \mathop{\mathbb{E}}_{\sigma}\left[\sum_{i\neq j}\sigma_i\sigma_j\langle\mathbf{z}_i,\mathbf{z}_j\rangle + \sum_i \sigma_i^2\langle\mathbf{z}_i,\mathbf{z}_i\rangle\right]\\ &= \sum_{i=1}^n \langle\mathbf{z}_i,\mathbf{z}_i\rangle^2\\ &\le nX^2.\end{aligned}$$

Plugging this back to the previous inequality leads to the result. ■

Remark In general, Hölder's Inequality tells us

$$\mathbf{w}^\top\mathbf{Z}\sigma \le \|\mathbf{w}\|\|\mathbf{Z}\sigma\|_*$$

for any norm $\|\cdot\|$. However, the Jensen Inequality trick cannot apply to general norms, so we may need another family of tricks when we are not only considering l_2-norm. In particular, norms with the form $\|\mathbf{x}\| = \sqrt{\mathbf{x}^\top\mathbf{M}\mathbf{x}}$ for some positive definite $\mathbf{M}$ is easy to deal with.

Remarks Before mentioning the proof of Lemma 19.5, let's just talk about some remarks about the proof.

- We are not going to give the whole proof of this lemma. Instead, we are going to show that

$$\mathop{\mathbb{E}}_{\sigma}\left[\sup_{\mathbf{v}\in V}\frac{1}{n}\sum_i g(v_i)\sigma_i\right] \le \mathop{\mathbb{E}}_{\sigma}\left[\sup_{\mathbf{v}\in V}\frac{1}{n}\left(\sigma_1 L v_1 + \sum_{i\ge 2}\sigma_i g(v_i)\right)\right],$$

  where we denote $\mathbf{v} = (v_1, v_2, \cdots, v_n)$. Note that the proof here can be easily turned into an inductive step in the proof of Lemma 19.5. Also, one may wonder that the left hand side does not look exactly like $\mathfrak{R}(g\circ V)$, but it can be proved easily that

$$\mathfrak{R}(g\circ V) = \mathop{\mathbb{E}}_{\sigma}\left[\sup_{\mathbf{v}'\in g\circ V}\frac{1}{n}\sum_i\sigma_i v_i'\right] = \mathop{\mathbb{E}}_{\sigma}\left[\sup_{\mathbf{v}\in V}\sum_i\frac{1}{n}\sigma_i g(v_i)\right].$$

- The main idea we are going to use here is **symmetrization**. We have already seen this technique a lot of times.

Proof of Lemma 19.5: In the following proof we denote σ_{-1} the string by truncating the first entry of σ, $\mathbf{v} = (v_1, \dots, v_n)$ and $\mathbf{u} = (u_1, \dots, u_n)$. Then

$$\begin{aligned}n\mathfrak{R}(g\circ V) &= \mathop{\mathbb{E}}_{\sigma}\left[\sup_{\mathbf{v}\in V}\sum_i g(v_i)\sigma_i\right]\\ &= \frac{1}{2}\mathop{\mathbb{E}}_{\sigma_{-1}}\left[\sup_{\mathbf{v}\in V}\left(g(v_1) + \sum_{i\ge 2}g(v_i)\sigma_i\right) + \sup_{\mathbf{u}\in V}\left(-g(u_1) + \sum_{i\ge 2}g(u_i)\sigma_i\right)\right]\\ &= \frac{1}{2}\mathop{\mathbb{E}}_{\sigma_{-1}}\left[\sup_{\mathbf{u},\mathbf{v}\in V} g(v_1) - g(u_1) + \sum_{i\ge 2}\sigma_i g(v_i) + \sum_{i\ge 2}\sigma_i g(u_i)\right]\end{aligned}$$

$$
\begin{aligned}
&\le \frac{1}{2} \mathop{\mathbb{E}}_{\sigma_{-1}} \left[\sup_{\mathbf{u},\mathbf{v} \in V} L(v_1 - u_1) + \sum_{i \ge 2} \sigma_i g(v_i) + \sum_{i \ge 2} \sigma_i g(u_i) \right] \\
&= \frac{1}{2} \mathop{\mathbb{E}}_{\sigma_{-1}} \left[\sup_{\mathbf{v} \in V} \left(L v_1 + \sum_{i \ge 2} \sigma_i g(v_i) \right) + \sup_{\mathbf{u} \in V} \left(-L u_1 + \sum_{i \ge 2} \sigma_i g(u_i) \right) \right] \\
&= \mathop{\mathbb{E}}_{\sigma} \left[\sup_{\mathbf{v} \in V} \sigma_1 L v_1 + \sum_{i \ge 2} \sigma_i g(v_i) \right].
\end{aligned}
$$

The second line is just expanding the expectation up to the first term, and the last line is just folding it back. The only thing worth explaining is the inequality in the fourth line. By Lipschitz condition we have $g(v_1) - g(u_1) \le L|v_1 - u_1|$. The reason that we can replace the absolute value by a pair of parenthesis is that if $v_1 < u_1$, then apparently the whole value is going to increase when we switch v and u. Therefore we can assume $v_1 \ge u_1$ without loss of generality. ■

Remarks

- We haven't got time to prove Lemma 19.3. Note that the condition $\mathbf{0} \in V_j$ is essential; otherwise for arbitrary V with finite cardinality we may get

  $$\mathfrak{R}(V) \le \sum_i \mathfrak{R}(\{\mathbf{v}_i\}) = 0,$$

  which is apparently not the case. In fact, if one tries to prove the lemma without $\mathbf{0} \in V_j$, the approach could be like

  $$\mathfrak{R}\left(\bigcup_j V_j \right) = \mathop{\mathbb{E}}_{\sigma} \left[\sup_j \sup_{\mathbf{v} \in V_j} \frac{1}{n} \sum \sigma_i (\mathbf{v})_i \right] \le \sum_j \mathop{\mathbb{E}}_{\sigma} \left[\sup_{\mathbf{v} \in V_j} \frac{1}{n} \sum \sigma_i (\mathbf{v})_i \right] = \sum_j \mathfrak{R}(V_j).$$

  However, this only makes sense when each single $\sum_i \frac{1}{n} \sigma_i (\mathbf{v})_i$ is non-negative. On the other hand, if we use the original definition of Rademacher complexity where everything is with the absolute value, then Lemma 19.3 indeed holds without the condition $\mathbf{0} \in V_j, \forall j$.

- Another way to prove Lemma 19.4 is to use Fenchel duality. Suppose for some convex function $F(\cdot)$ it holds that $F(\mathbf{w}) \le B$ for all $\mathbf{w} \in W$. Then for all $\lambda > 0$ we have

  $$\mathbf{w}^\top \mathbf{Z} \sigma \le F(\mathbf{w}/\lambda) + F^*(\lambda \mathbf{Z} \sigma).$$

  This can be handy when proving Lemma 19.4. (Exercise: try this).