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

Lecture 20: Fast Rates, Part II

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

20.1 Setup

Assume we are given the following:

- *Parameter set* W
- *Distribution* on $\mathcal{X} \times \mathcal{Y}$, with $\|\mathbf{x}_2\| \leq 1$ and $y \in \{-1, 1\}$
- *Sample Set* $S = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$
- *Loss* $\ell : \mathbb{R} \to \mathbb{R}$. Define the *risk* $R_\ell(\mathbf{w}) = \mathbb{E}[\ell(y\langle \mathbf{w}, \mathbf{x}\rangle)]$. Note that we said we needed convexity of our loss function last time, but it turns out we don't necessarily need it.

We also make the following important assumptions:

- $\exists\, \overline{\mathbf{w}} \in W$ with $R_\ell(\overline{\mathbf{w}}) = \inf_{\mathbf{v} \in W} R_\ell(\mathbf{v})$. That is, we assume there exists a minimizer, though this need not always be the case in general.
- $\exists\, \lambda > 0$ with $\epsilon_\ell(\mathbf{w}) \geq \frac{\lambda}{2}\|\mathbf{w} - \overline{\mathbf{w}}\|_2^2 \;\; \forall\, \mathbf{w} \in W$. That is, the excess risk is bounded from below by the distance from the optimum times a scaling factor. This is a form of strong convexity but only at the optimum.
- $\exists\, \ell > 0$ where ℓ is L-Lipschitz on the interval $[\inf_{\mathbf{w} \in W, \mathbf{z} \in S} \langle \mathbf{w}, \mathbf{z}\rangle, \sup_{\mathbf{w} \in W, \mathbf{z} \in S} \langle \mathbf{w}, \mathbf{z}\rangle]$. This is the Lipschitz condition on the loss that we need. ℓ being L-Lipschitz everywhere might be too strong an assumption, but at least we want it to be Lipschitz on the range we care about. Indeed, many papers say "let a strongly convex Lipschitz function exist", but they actually mean "strongly convex and Lipschitz over some bounded set".

The main result that we will prove this lecture is the following:

Theorem 20.1. *With probability* $1 - \delta$*, for any* $\mathbf{w} \in W$*,*

$$\epsilon_\ell(\mathbf{w}) \leq 2\hat{\epsilon}_\ell(\mathbf{w}) + \frac{1024 L^2 \ln(4e/\delta)}{\lambda n}.$$

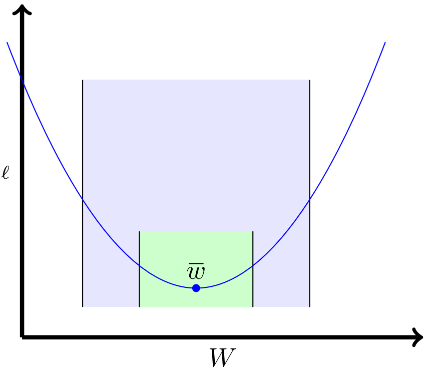
Here $\epsilon_\ell(\mathbf{w})$ is the empirical risk and $\hat{\epsilon}_\ell(\mathbf{w})$ is the excess risk of $\mathbf{w}$ according to the loss function ℓ. Instead of $1/\sqrt{n}$, here we have $1/n$, which makes this a fast rate.

20.2 Basic Proof Structure

We will start with two basic operations that happen all the time in local Rademacher complexity analysis.

- "Centering operation" by $\overline{\mathbf{w}}$: we care about how $\mathbf{w}$ is related to $\overline{\mathbf{w}}$, not $\mathbf{w}$ by itself. Concretely, the new loss function maps $\mathbf{z}$ to $\ell(\langle \mathbf{w}, \mathbf{z}\rangle) - \ell(\langle \overline{\mathbf{w}}, \mathbf{z}\rangle)$.
- "Units corrections": scaled by its variance.

Figure 20.1: A simple shelling diagram



To foster intuition, note that these two operations are reminiscent of the Central Limit Theorem where we subtract by the mean and scale by the standard deviation.

Now we will control the scaling according to various shells, where each shell is over a wider interval of $\overline{\mathbf{w}}$ than the previous ones in sequence. (This is another technical detail that is common in local Rademacher complexity analysis.) All elements in a shell get scaled a certain way, so elements are scaled differently according to which shell they are in. As an example, let the vertical lines mark off shells in the figure. (Note that by strong convexity we assume our function is bounded below by a quadratic, so it will look something like the one plotted.)

20.3 Parameter definitions

Let's define the following parameters for our analysis:

- $r > 0$, a real number, is a parameter that basically controls how big the first shell is. We will optimize this at the end.
- $k_{\mathbf{w}} = \inf\{k \in \mathbb{Z}_{\geq 0} : \epsilon_\ell \leq r(4^k)\}$. This is the smallest k such that the excess risk is bounded by $r(4^k)$. It's a parameter that is how we'll choose the shell we're in, and thereby control the scaling.
- $f_{\mathbf{w}} = \ell(\langle \mathbf{w}, \mathbf{z}\rangle) - \ell(\langle \overline{\mathbf{w}}, \mathbf{z}\rangle)$. Note that f_w can be negative on some points, but in expectation must be nonnegative for $\overline{\mathbf{w}}$ to be a minimizer.
- $g_{\mathbf{w}}(\mathbf{z}) = 4^{-k_{\mathbf{w}}} f_{\mathbf{w}}(\mathbf{z})$. Here the $4^{-k_{\mathbf{w}}}$ is just a scaling term.
- $G = \{g_{\mathbf{w}} : \mathbf{w} \in W\}$ is the function class that we'll be applying local Rademacher complexity (over a sample S) to.

Further clarification on the meaning of $k_{\mathbf{w}}$:

- If $k_{\mathbf{w}} = 0$, then $\epsilon_\ell(\mathbf{w}) \leq r = r4^0$.
- If $k_{\mathbf{w}} \geq 1$, then
$$r4^{k_{\mathbf{w}}-1} \leq \epsilon_\ell(\mathbf{w}) \leq r4^{k_{\mathbf{w}}} \iff \frac{\epsilon_\ell(\mathbf{w})}{r} \leq 4^{k_{\mathbf{w}}} < \frac{4\epsilon_\ell(\mathbf{w})}{r}. \tag{20.1}$$
It will simplify the proof to have the shells grow in a doubling way as they do here.

20.4 The Proof

We will use the following uniform deviation bound:

Lemma 20.2 (Uniform deviation bound)**.** *With probability $1-\delta$, for any $\mathbf{w} \in W$, we have*

$$\mathop{\mathbb{E}}_{S}(g_{\mathbf{w}}) - \mathop{\hat{\mathbb{E}}}_{S}(g_{\mathbf{w}}) \leq \Upsilon, \tag{20.2}$$

where S is the sample, and

$$\Upsilon := 2\mathfrak{R}(G) + 4 \sup_{\mathbf{z}\in S, \mathbf{w}\in W} |g_{\mathbf{w}}(\mathbf{z})| \sqrt{\frac{2\ln(4/\delta)}{n}}.$$

■

For the ease of notation, denote

$$\begin{aligned} \text{❶} &= \mathop{\mathbb{E}}_{S}(g_{\mathbf{w}}) - \mathop{\hat{\mathbb{E}}}_{S}(g_{\mathbf{w}}) \\ \text{❷} &= \mathfrak{R}(G) \\ \text{❸} &= 4 \sup_{\mathbf{z}\in S, \mathbf{w}\in W} |g_{\mathbf{w}}(\mathbf{z})| \end{aligned}$$

Proof of Theorem 20.1. (20.2) is not exactly what we want, so let's decode it.

- First, we rewrite ❶ $= 4^{-k_{\mathbf{w}}}(\epsilon_\ell(\mathbf{w}) - (\hat{R}_\ell(\mathbf{w}) - \hat{R}_\ell(\overline{\mathbf{w}})))$, and note that $\hat{R}_\ell(\mathbf{w}) - \hat{R}_\ell(\overline{\mathbf{w}}) \leq \hat{\epsilon}_\ell(\mathbf{w})$, so we have that $4^{-k_{\mathbf{w}}}(\epsilon_\ell(\mathbf{w}) - (\hat{R}_\ell(\mathbf{w}) - \hat{R}_\ell(\overline{\mathbf{w}}))) \geq 4^{-k_{\mathbf{w}}}(\epsilon_\ell(\mathbf{w}) - \hat{\epsilon}_\ell(\mathbf{w}))$. Therefore, we have $\epsilon_\ell(\mathbf{w}) + \hat{\epsilon}_\ell(\mathbf{w}) \leq 4^{k_{\mathbf{w}}}\Upsilon$.

$$\begin{cases} \text{If } k_{\mathbf{w}} = 0 \text{ (we're in that central shell), then } \epsilon_\ell(\mathbf{w}) \leq \hat{\epsilon}_\ell(\mathbf{w}) + \Upsilon. \\ \text{If } k_{\mathbf{w}} \geq 1, \text{ then } \epsilon_\ell(\mathbf{w}) - \hat{\epsilon}_\ell(\mathbf{w}) \leq \frac{4\epsilon_\ell(\mathbf{w})}{r}\Upsilon \text{ by Equation 20.1.} \\ \text{If } \Upsilon \leq \frac{r}{8},\ \epsilon_\ell(\mathbf{w}) \leq 2\hat{\epsilon}_\ell(\mathbf{w}) \end{cases}$$

- We deal with ❸ first. Given given $\mathbf{z} \in S$, $\mathbf{w} \in W$, the following are all "forcing moves" (like in chess!):

$$\begin{aligned} |g_{\mathbf{w}}(\mathbf{z})| &= 4^{-k_{\mathbf{w}}} |\ell(\langle \mathbf{w}, \mathbf{z}\rangle) - \ell(\langle \overline{\mathbf{w}}, \mathbf{z}\rangle)| \\ &\leq 4^{-k_{\mathbf{w}}} L |\langle \mathbf{w} - \overline{\mathbf{w}}, \mathbf{z}\rangle| && \text{(Lipschitz condition)} \\ &\leq 4^{-k_{\mathbf{w}}} L \|\mathbf{w} - \overline{\mathbf{w}}\|_2 \|\mathbf{z}\|_2 && \text{(Cauchy-Schwarz Inequality)} \\ &\leq 4^{-k_{\mathbf{w}}} L \|\mathbf{w} - \overline{\mathbf{w}}\|_2 && \text{(by assumption, } \|\mathbf{z}\|_2 \leq 1) \\ &\leq 4^{-k_{\mathbf{w}}} L \sqrt{\frac{2\epsilon_\ell(\mathbf{w})}{\lambda}} && \text{(local strong convexity assumption)} \\ &\leq 4^{-k_{\mathbf{w}}} L \sqrt{2r\frac{4^{k_{\mathbf{w}}}}{\lambda}} \\ &\leq L\sqrt{\frac{2r}{\lambda}}. \end{aligned}$$

- Next, we'll use an important trick called *peeling* for ❷. We define a countable sequence $(\mathcal{F}_k)_{k\geq 0}$ such that $G \subseteq \bigcup_k (\mathcal{F}_k), \mathbf{0} \in \mathcal{F}_k$. Hence $\mathfrak{R}(G) \leq \mathfrak{R}\left(\bigcup_k \mathcal{F}_k\right) \leq \sum\limits_k \mathfrak{R}(\mathcal{F}_k)$.

  Define $\mathcal{F}_k = \{4^{-k} f_{\mathbf{w}} : \mathbf{w} \in W, k_{\mathbf{w}} \leq k\}$. Note that we automatically have $\mathbf{0} \in \mathcal{F}_k$, since $\overline{\mathbf{w}} \in W$. We know that $\epsilon_\ell(\mathbf{w}) \leq 4^{k_{\mathbf{w}}} r \leq 4^k r$.

So

$$\begin{aligned}
\mathfrak{R}(\mathcal{F}_k) &\le 4^{-k}\mathfrak{R}(\{f_{\mathbf{w}} : \mathbf{w} \in W, \epsilon_\ell(\mathbf{w}) \le r4^k\}) \\
&\le 4^{-k}\mathfrak{R}\left(\left\{\mathbf{z} \to \ell(\langle \mathbf{w}, \mathbf{z}\rangle) - \ell(\langle \overline{\mathbf{w}}, \mathbf{z}\rangle)) : \mathbf{w} \in W, \|\mathbf{w} - \overline{\mathbf{w}}\| \le \sqrt{\frac{2r4^k}{\lambda}}\right\}\right) \\
&\le 4^{-k}L\mathfrak{R}\left(\left\{\mathbf{z} \to \langle \mathbf{w} - \overline{\mathbf{w}}, \mathbf{z}\rangle : \mathbf{w} \in W, \|\mathbf{w} - \overline{\mathbf{w}}\|_2 \le \sqrt{\frac{2r4^k}{\lambda}}\right\}\right) && (L\text{-Lipschitz}) \\
&\le 4^{-k}L\mathfrak{R}\left(\left\{\mathbf{z} \to \langle \mathbf{w}, \mathbf{z}\rangle : \mathbf{w} \in W + \{\overline{\mathbf{w}}\}, \|\mathbf{w}\|_2 \le \sqrt{\frac{2r4^k}{\lambda}}\right\}\right) && (\text{shifted by } \overline{\mathbf{w}}) \\
&\le 4^{-k}L\sqrt{\frac{2r4^k}{\lambda n}} = 2^{-k}L\sqrt{\frac{2r}{\lambda n}}.
\end{aligned}$$

Therefore,

$$❷ \le \sum_k \mathfrak{R}(\mathcal{F}_k) \le L\sqrt{\frac{2r}{\lambda n}}\sum_{k\ge 0} 2^{-k} = L\sqrt{\frac{2r}{\lambda n}}.$$

The important thing about this final summation is that it is convergent. Also note that when we do this union, we will be double-counting a lot of things (e.g. when we're in the first shell we'll also be in all the bigger shells), but we'll also have our variance smashed down, so we pay only a constant factor cost.

Finally, let's figure out what Υ is. We have that

$$\begin{aligned}
\Upsilon &\le 4L\sqrt{\frac{2r}{\lambda n}} + 4L\sqrt{\frac{2r}{\lambda}}\sqrt{\frac{2\ln 4/\delta}{n}} \\
&\le 4L\sqrt{\frac{8r(1 + \ln 4/\delta)}{\lambda n}} \\
&= 4L\sqrt{\frac{8r\ln(4e/\delta)}{\lambda n}} \\
&= \sqrt{r}L\sqrt{\frac{128\ln(4e/\delta)}{\lambda n}}
\end{aligned}$$

Now we'll try to satisfy $\Upsilon = r/8$, or equivalently $8\Upsilon^2 = r^2$.
In this case, we choose $\sqrt{r} = 8L\sqrt{\frac{128\ln(4e/\delta)}{\lambda n}}$, or equivalently

$$r = 8L^2\frac{128\ln(4e/\delta)}{\lambda n} = \frac{1024L^2\ln(4e/\delta)}{\lambda n}$$

giving us what we set out to show at the beginning. ■

20.5 More Information

More information about local Rademacher complexity can be found in the original paper on the topic [1]. There is also a recent paper [2] that tries to characterize when fast rates are possible, which may be germane to the lectures this week.

References

[1] Peter L Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *Annals of Statistics*, pages 1497–1537, 2005.

[2] Tim van Erven, Peter D. Grünwald, Nishant A. Mehta, Mark D. Reid, and Robert C. Williamson. Fast rates in statistical and online learning. *CoRR*, abs/1507.02592, 2015.