CS229T/STATS231: Statistical Learning Theory

Lecturer: Tengyu Ma
Scribe: Jay Whang and Patrick Cho

Lecture #6
October 10, 2018

---

1 Review and Overview

Recall in the last lecture that for any family of scalar functions $\mathcal{F}$, we have

$$\mathop{\mathbb{E}}_{z_1,\ldots,z_n\sim p}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^{n}f(z_i)-\mathop{\mathbb{E}}_{z\sim p}[f(z)]\right)\right]\le 2\mathcal{R}_n(\mathcal{F}) \tag{1}$$

where

$$\mathcal{R}_n(\mathcal{F})\triangleq\mathop{\mathbb{E}}_{z_1,\ldots z_n\sim p}\left[\mathop{\mathbb{E}}_{\substack{\sigma_i\sim\{-1,1\}\\ i=1,\ldots,n}}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(z_i)\right]\right] \tag{2}$$

Let us define a family of functions, $\mathcal{F}$, that depends on loss function ℓ and hypothesis class $\mathcal{H}$ in the following manner

$$\mathcal{F}=\{z=(x,y)\mapsto\ell((x,y),h):h\in\mathcal{H}\} \tag{3}$$

Then, we have

$$\mathop{\mathbb{E}}_{z_1,\ldots,z_n}\left[\sup_{h\in\mathcal{H}}\left|\hat{L}(h)-L(h)\right|\right]\le 4\mathcal{R}_n(\mathcal{F}) \tag{4}$$

where the multiple of 2 comes from applying equation (1) twice - once with the original loss function and another with the negative of the original loss function.

In this lecture, we will try to prove the corresponding high probability bound. In particular, we will see that this bound does not depend on the complexity of hypothesis class $\mathcal{H}$. We will also see this bound applied on two specific loss functions - hinge loss and 0-1 loss. Using Talagrand's lemma, we will see that the bound for these two loss functions no longer depend on the loss function.

2 Preliminaries

Definition 1. *Let $S = \{z_1, ..., z_n\}$ be a set of i.i.d. examples from some distribution p.*

*The **empirical Rademacher complexity** $\mathcal{R}_S(\mathcal{F})$ is defined as:*

$$\mathcal{R}_S(\mathcal{F}) \triangleq \mathop{\mathbb{E}}_{\substack{\sigma_i \sim \{-1,1\} \\ i=1,...,n}} \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^{n} \sigma_i f(z_i) \right) \right] \tag{5}$$

Comparing the definitions of $\mathcal{R}_n(\mathcal{F})$ and $\mathcal{R}_s(\mathcal{F})$, we have

$$\mathcal{R}_n(\mathcal{F}) = \mathop{\mathbb{E}}_{z_1,...,z_n \sim p} [\mathcal{R}_s(\mathcal{F})] \tag{6}$$

Definition 2. *A function $f : \mathbb{R}^n \to \mathbb{R}$ satisfies the **bounded difference condition** if*

$$\exists\ c_1, ...c_n \geq 0 \text{ such that}$$

$$\left| f(x_1, ..., x_i, ..., x_n) - f(x_1, ..., x_i', ..., x_n) \right| \leq c_i \ \ \forall i, x_1, ..., x_n, x_i'$$

Theorem 1 (McDiarmid's Inequality)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ satisfy the bounded difference condition. Let $x_1, ..., x_n$ be independent random variables. Then,*

$$\Pr\left[f(x_1, ...x_n) - \mathbb{E}\left[f(x_1, ..., x_n) \right] \geq \varepsilon \right] \leq \exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^{n} c_i^2} \right) \tag{7}$$

Comparing McDiarmid's Inequality to other concentration inequalities such as Hoeffding's Inequality, we realize that the function is no longer a summation of $x_1, \ldots, x_n$.

3 High Probability Bounds

Lemma 1. *Let $\ell(z, h) \in [-1, 1]$. Then, w.p. $\geq 1 - \delta$,*

$$\mathcal{R}_s(\mathcal{F}) - \mathcal{R}_n(\mathcal{F}) \leq \sqrt{\frac{2 \log \frac{1}{\delta}}{n}} \tag{8}$$

Proof. Let

$$g(z_1, ...z_n) = \mathcal{R}_s(\mathcal{F}) = \mathop{\mathbb{E}}_{\sigma} \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^{n} \sigma_i f(z_i) \right) \right] \tag{9}$$

Then, we have

$$
\begin{aligned}
&\left|g(z_1,...,z_i,...z_n) - g(z_1,...,z_i',...z_n)\right| \\
&= \mathop{\mathbb{E}}_{\sigma}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{j=1}^{n}\sigma_j f(z_j)\right) - \sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{j=1}^{n}\sigma_j f(z_j) - \sigma_i f(z_i) + \sigma_i f(z_i')\right)\right] \\
&\leq \mathop{\mathbb{E}}_{\sigma}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{j=1}^{n}\sigma_j f(z_j)\right) - \sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{j=1}^{n}\sigma_j f(z_j)\right) + \sup_{f\in\mathcal{F}}\left(\frac{1}{n}\left|\sigma_i f(z_i) - \sigma_i f(z_i')\right|\right)\right] \\
&= \mathop{\mathbb{E}}_{\sigma}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\left|\sigma_i f(z_i) - \sigma_i f(z_i')\right|\right)\right] \\
&\leq \frac{2}{n}
\end{aligned}
$$

where we use in order: definition of g, $\sup(A) \leq \sup(A-B)+\sup|B|$, cancellation of terms, $f \in [-1,1]$.

Therefore, setting $c_i = \frac{2}{n}$ and applying McDiarmid's Inequality, we have

$$
\begin{aligned}
\Pr\left[g - \mathop{\mathbb{E}}_{z_1,...z_n\sim p}[g] \geq \varepsilon\right] &= \Pr\left[\mathcal{R}_s(\mathcal{F}) - \mathcal{R}_n(\mathcal{F}) \geq \varepsilon\right] \\
&\leq \exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^{n} c_i^2}\right) \\
&= \exp\left(-\frac{2\varepsilon^2}{\frac{4}{n}}\right) \\
&= \exp\left(-\frac{n\varepsilon^2}{2}\right) \\
&= \delta
\end{aligned}
$$

This is achieved by setting

$$
\varepsilon = \sqrt{\frac{2\log\frac{1}{\delta}}{n}} \qquad \square
$$

So, w.p. $1-\delta$, we have

$$
\mathcal{R}_s(\mathcal{F}) - \mathcal{R}_n(\mathcal{F}) \leq \sqrt{\frac{2\log\frac{1}{\delta}}{n}} \tag{10}
$$

Theorem 2. *Let $\ell(z,h) \in [-1,1]$. Then, w.p. $\geq 1-\delta$,*

$$\left| \mathop{\mathbb{E}}_{z_1,\ldots z_n \sim p} \left[\sup_{h\in\mathcal{H}} \hat{L}(h) - L(h) \right] - \sup_{h\in\mathcal{H}} \hat{L}(h) - L(h) \right| \leq \sqrt{\frac{2\log(2/\delta)}{n}} \tag{11}$$

Remark: Note that in the lecture, we showed a variant of this theorem:

$$\left| \mathop{\mathbb{E}}_{z_1,\ldots z_n \sim p} \left[\sup_{h\in\mathcal{H}} |\hat{L}(h) - L(h)| \right] - \sup_{h\in\mathcal{H}} |\hat{L}(h) - L(h)| \right| \leq \sqrt{\frac{2\log(2/\delta)}{n}}$$

The inequality above is indeed true and can be proved with the same technique as below. However, it's a bit less convenient to deal with the absolute value in $\mathbb{E}_{z_1,\ldots z_n\sim p}\left[\sup_{h\in\mathcal{H}} |\hat{L}(h) - L(h)|\right]$ when we trying to bound it from above.

Proof. Define $g(z_1,\ldots,z_n) = \sup_{h\in\mathcal{H}} \frac{1}{n}\sum \ell(z_i,h) - L(h)$.

For notational simplicity, write $A = \hat{L}(h) - L(h) = \left(\frac{1}{n}\sum \ell(z_i,h)\right) - L(h)$ and $B = \frac{1}{n}(\ell(z_i,h) - \ell(z_i',h))$. Then $g(z_1,\ldots,z_n) = \sup_{h\in\mathcal{H}} A$ and $g(z_1,\ldots,z_i',\ldots,z_n) = \sup_{h\in\mathcal{H}}(A-B)$.

Then,

$$\begin{aligned}
&|g(z_1,\ldots,z_i,\ldots,z_n) - g(z_1,\ldots,z_i',\ldots,z_n)| \\
&\qquad = \left| \sup_{h\in\mathcal{H}}(A) - \sup_{h\in\mathcal{H}}(A-B) \right| \\
&\qquad \leq \sup_{h\in H} |A - (A-B)| && \text{sup is a contraction} \\
&\qquad \leq \sup_{h\in H} |B| \\
&\qquad = \sup_{h\in H} \left| \frac{1}{n}(\ell(z_i,h) - \ell(z_i',h)) \right| \\
&\qquad \leq \frac{2}{n} && \ell \in [-1,1]
\end{aligned}$$

Thus, g satisfies the bounded difference condition. Applying McDiarmid's Inequality on g with $c_i = \frac{2}{n}$, we get:

$$\begin{aligned}
&\Pr\left[\sup_{h\in\mathcal{H}} \hat{L}(h) - L(h) - \mathop{\mathbb{E}}_{z_1,\ldots z_n\sim p} \left[\sup_{h\in\mathcal{H}} \hat{L}(h) - L(h) \right] \right] \\
&\quad = \Pr\left[g(z_1,\ldots,z_n) - \mathbb{E}\, g(z_1,\ldots,z_n) \geq \varepsilon \right] \\
&\quad \leq \exp\left(-\frac{2\varepsilon^2}{\sum c_i^2} \right) = \exp\left(\frac{-n\varepsilon^2}{2} \right)
\end{aligned}$$

Finally, applying this to g and $-g$, and solving for ε gives: w.p $1-\delta$,

$$\left| \sup_{h\in\mathcal{H}} \hat{L}(h) - L(h) - \mathop{\mathbb{E}}_{z_1,\ldots z_n\sim p} \left[\sup_{h\in\mathcal{H}} \hat{L}(h) - L(h) \right] \right| \leq \sqrt{\frac{2\log(2/\delta)}{n}}$$

□

4 Bounding the excess risk

Theorem 2, combined with the symmetrization argument, gives a high probability bound on the excess risk in terms of the Rademacher complexity and the concentration level. We state this formally in the following theorem.

Theorem 3. *Let $\ell(z,h) \in [-1,1]$ for all $h \in H$ and $F = \{z \mapsto \ell(z,h) : h \in H\}$ be the loss class. Then we have with probability at least $1-\delta$ the excess risk bound*

$$L(\hat{h}) - L(h^*) \leq 4R_n(F) + +2\sqrt{\frac{2\log(4/\delta)}{n}}. \tag{12}$$

Remark 1: The constant 4 can be improved if we sacrifice the constant in front of $\sqrt{\frac{\log(4/\delta)}{n}}$ because in the proof below, for h^* we don't have to use a uniform convergence result. We can directly apply Hoeffding inequality on $L(h^*) - \hat{L}(h^*)$. Although in general for the purpose of this class we don't care much about absolute constant dependency.

Remark 2: The proof below is slightly different from the proof given in the lecture. Here we go through $\mathbb{E}_{z_1,\dots z_n \sim p}\left[\sup_{h\in\mathcal{H}} \hat{L}(h) - L(h)\right]$ and $\mathbb{E}_{z_1,\dots z_n \sim p}\left[\sup_{h\in\mathcal{H}} -\hat{L}(h) + L(h)\right]$ instead of $\mathbb{E}_{z_1,\dots z_n \sim p}\left[\sup_{h\in\mathcal{H}} |\hat{L}(h) - L(h)|\right]$ in the lecture. To some extent, the reason why this is easier is that we no longer use the absolute value inside the sup.

Proof. We have

$$\begin{aligned} L(\hat{h}) - L(h^*) &= L(\hat{h}) - \hat{L}(\hat{h}) + \underbrace{\hat{L}(\hat{h}) - \hat{L}(h^*)}_{\leq 0} + \hat{L}(h^*) - L(h^*) \\ &\leq \underbrace{\sup_{h\in H} L(h) - \hat{L}(h)}_{\text{I}} + \underbrace{\sup_{h\in H} \hat{L}(h) - L(h)}_{\text{II}}. \end{aligned}$$

For term II, applying Theorem 2 and standard symmetrization argument, we get that with probability at least $1-\delta/2$,

$$\sup_{h\in H} \hat{L}(h) - L(h) \leq \mathbb{E}\left[\sup_{h\in H} \hat{L}(h) - L(h)\right] + \sqrt{\frac{2\log(4/\delta)}{n}} \leq 2R_n(F) + \sqrt{\frac{2\log(4/\delta)}{n}}.$$

For term I, the same argument with losses $-\ell(z,h)$ gives that with probability at least $1-\delta/2$,

$$\sup_{h\in H} L(h) - \hat{L}(h) \leq \mathbb{E}\left[\sup_{h\in H} L(h) - \hat{L}(h)\right] + \sqrt{\frac{2\log(4/\delta)}{n}} \leq 2R_n(F^-) + \sqrt{\frac{2\log(4/\delta)}{n}},$$

where $F^- = \{-f : f \in F\}$ is the negative loss class. By the union bound, the above two events happen together with probability at least $1-\delta$. Further noticing that $R_n(F^-) = R_n(F)$ (as $\sigma_i f(z_i) \overset{d}{=} \sigma_i(-f(z_i))$, we obtain the desired result. □

5 Geometric Viewpoint of Rademacher Complexity

Let $Q = \{(f(z_1), ..., f(z_n)) : f \in \mathcal{F}\} \subseteq \mathbb{R}^m$. We can think of each element of the set Q as the a vector of losses calculated for a single hypothesis in the hypothesis class $\mathcal{H}$. Then,

$$\mathcal{R}_s(\mathcal{F}) = \mathop{\mathbb{E}}_{\sigma} \sup_{v \in Q} \frac{1}{n} \langle v, \sigma \rangle \tag{13}$$

From this viewpoint, the complexity of the set Q can be viewed as the following process. Firstly, pick a random $\sigma \in \mathbb{R}^n$ where $\sigma_i \in \{-1, 1\}$. Then, we pick the largest point v in the set Q that maximizes the dot product between v and σ. Finally, we take an expectation over σ.

In particular, let us fix a loss function ℓ and a dataset $(z_1, ...z_n)$. Given this loss function, if the hypothesis class $\mathcal{H}$ results in the vector of losses pointing in the same direction, then the Rademacher complexity is low. This is because when taking the sup over elements in the set Q, it is difficult to find one that has a small cosine distance (or large dot product). On the other hand, if the hypothesis class $\mathcal{H}$ allows the vector of losses to point in many different directions, then the Rademacher complexity is high.

6 Talagrand's Lemma

This lemma is also known as *contraction principle* and *Lipschitz composition*.

Lemma 2 (Contraction principle)**.** *Let $\phi : \mathbb{R} \mapsto \mathbb{R}$ be a k-Lipschitz function, then*

$$\mathcal{R}_s(\phi \circ \mathcal{H}) \leq k \cdot \mathcal{R}_s(\mathcal{H}), \tag{14}$$

where $\phi \circ \mathcal{H} = \{z \mapsto \phi(h(z)) : h \in \mathcal{H}\}$.

7 Moving from $\mathcal{R}_s(\mathcal{F})$ to $\mathcal{R}_s(\mathcal{H})$

Consider a family of loss functions $\mathcal{F} = \{z \mapsto \ell((x, y), h) : h \in \mathcal{H}\}$ where $\ell((x, y), h)$ can be expressed as

$$\ell((x, y), h) = \ell(h(x), y)$$

Then, we can use Talagrand's lemma to remove the influence of the loss function. Concretely, instead of the Rademacher complexity of the family of functions $\mathcal{F}$, we can derive a bound that depends on the Rademacher complexity of the hypothesis class $\mathcal{H}$. To be concrete, let us consider binary classification where $y_i \sim \{-1, +1\}$ together with two losses - hinge loss and binary loss.

7.1 Hinge Loss

For hinge loss, we have

$$\ell(z, h) = \max(0, 1 - y_i h(x_i))$$

The hinge loss is 1-lipschitz. Let us define $\phi(t) = \max(0, 1 - t)$. Then, we have

$$\begin{aligned}
\mathcal{R}_s(\mathcal{F}) &= \mathop{\mathbb{E}}_{\sigma} \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \phi(y_i h(x_i)) \\
&\leq \mathop{\mathbb{E}}_{\sigma} \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i y_i h(x_i) \\
&= \mathop{\mathbb{E}}_{\sigma} \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i h(x_i) \\
&= \mathcal{R}(\mathcal{H})
\end{aligned}$$

where the inequality holds true because $\mathcal{F} = \phi \circ \mathcal{H}'$ and by Talagrand, we have

$$\mathcal{R}_s(\mathcal{F}) \leq k \mathcal{R}_s(\mathcal{H}')$$

where $\mathcal{H}' = \{z \mapsto y h(x) : h \in \mathcal{H}\}$. Since hinge loss is 1-lipschitz, $k = 1$.

In the third step of the derivation, we can remove y_i because σ_i and $\sigma_i y_i$ come from the same distribution.

7.2 Binary Loss

Since binary loss is not Lipschitz, we use margin theory to upper bound the binary loss. We define the following γ-margin loss for $\gamma > 0$

$$\ell_\gamma(t) = \begin{cases} 0, & \text{if } t \geq \gamma \\ 1, & \text{if } t \leq 0 \\ 1 - \frac{t}{\gamma}, & \text{otherwise} \end{cases}$$

Also let

$$\ell_\gamma((x, y), h) = \ell_\gamma(y h(x))$$
$$\mathcal{F}_\gamma = \{(x, y) \mapsto \ell_\gamma(y h(x))\}$$

with L_γ and $\hat{L}_\gamma$ defined accordingly.

We desire to give an upper bound on the expected binary loss. We give this bound by first giving bounds for the following three terms: $L(h)$, $\hat{L}_\gamma(h)$ and $|\hat{L}_\gamma(h) - L_\gamma(h)|$.

Firstly, notice that ℓ_γ upper-bounds the binary loss as shown below:

$$L(h) = \mathbb{E}\left[\mathbf{1}\left[yh(x) < 0\right]\right] \leq \mathbb{E}\left[l_\gamma((x,y),h)\right] = L_\gamma(h) \tag{15}$$

Secondly, we bound $\hat{L}_\gamma(h)$ by the training loss w.r.t a margin. Intuitively, the training loss w.r.t. a margin means that a point is considered to be classified wrongly ($\ell = 1$) even if it is on the correct side of the decision boundary if the point is within γ of the margin.

$$\hat{L}_\gamma(h) = \frac{1}{n}\sum_{i=1}^{n} \ell_\gamma(y_i h(x_i)) \leq \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left[y_i h(x_i) \leq \gamma\right] \tag{16}$$

Lastly, we bound the generalization error of γ-margin loss: w.p $\geq 1 - \delta$,

$$\begin{aligned}
\forall h,\ \left|\hat{L}_\gamma(h) - L_\gamma(h)\right| &\lesssim \mathcal{R}_S(\mathcal{F}_\gamma) + \sqrt{\frac{\log(2/\delta)}{n}} \\
&\lesssim \frac{1}{\gamma}\mathcal{R}_S(\mathcal{H}) + \sqrt{\frac{\log(2/\delta)}{n}} \qquad \text{by Talagrand's Lemma}
\end{aligned}$$

In particular, this implies the following:

$$L_\gamma(h) \lesssim \hat{L}_\gamma(h) + \frac{1}{\gamma}\mathcal{R}_S(\mathcal{H}) + \sqrt{\frac{\log(2/\delta)}{n}} \tag{17}$$

Combining inequalities (15), (16) and (17), we obtain:

$$L(h) \leq \hat{L}_\gamma(h) + O\left(\frac{\mathcal{R}_S(\mathcal{H})}{\gamma} + \sqrt{\frac{\log(2/\delta)}{n}}\right) \tag{18}$$

Notice that the above inequality suggests a tradeoff in the choice of γ. For a large γ, the training loss w.r.t. a margin increases but the term on the right decreases and vice versa.