CS 7545: Machine Learning Theory **Fall 2018**

Lecture 21: Growth Function and Massart's Lemma

Lecturer: Jacob Abernethy *Scribes: Kyle Pelton, Justin Yao*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

21.1 Rademacher Complexity and the Symmetrization Lemma

21.1.1 Review

In the last lecture, we introduced the concept of the Rademacher Complexity of a class of functions. The Rademacher Complexity represents the richness of this class of functions, specifically measuring to what extent the functions can fit random noise.

Let G be a class of functions mapping $\mathcal{X} \to \{0,1\}$, let $S = (x_1, \ldots, x_m)$ be a sample of m points from $\mathcal{X}$, and let $\sigma_1, \ldots, \sigma_m$ be i.i.d. Rademacher random variables (i.e., $\sigma_i \in \{-1, 1\}$ and $\sigma_i = 1$ with probability $\frac{1}{2}$ for all $i = 1, \ldots, m$). Using these, we define the following terms:

Definition 21.1 (Empirical Rademacher Complexity) *The* ***Empirical Rademacher Complexity*** *of G is defined as:*

$$\hat{\mathfrak{R}}_S(G) = \mathop{\mathbb{E}}_{\sigma_1,\ldots,\sigma_m}\left[\frac{1}{m}\sup_{g\in G}\sum_{i=1}^{m} g(x_i)\sigma_i\right]$$

Definition 21.2 (Rademacher Complexity) *Given a distribution $D \in \Delta(\mathcal{X})$, the* ***Rademacher Complexity*** *with respect to D is:*

$$\mathfrak{R}_m(G) = \mathop{\mathbb{E}}_{S \overset{i.i.d.}{\sim} D^m}\left[\hat{\mathfrak{R}}_S(G)\right]$$

21.1.2 McDiarmid's Inequality and the Symmetrization Lemma

Using these definitions, we can show the following bound:

Theorem 21.3 (Symmetrization Lemma) *Let $S = (x_1, \ldots, x_m) \overset{i.i.d.}{\sim} D^m$, $g \in G : \mathcal{X} \to \{0,1\}$, $\mathbb{E}\, g = \mathbb{E}_{x\sim D}\left[g(x)\right]$, and $\hat{\mathbb{E}}_S g = \frac{1}{m}\sum_{x_i\in S} g(x_i)$. Then, the following bound holds, with probability $1-\delta$:*

$$\sup_{g\in G}\left(\mathbb{E}\, g - \hat{\mathbb{E}}_S g\right) \le 2\mathfrak{R}_m(G) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}$$

Notice that $2\mathfrak{R}_m(G)$ represents deterministically how much you are going to overfit on this class of functions, while $\sqrt{\frac{\log\frac{1}{\delta}}{2m}}$ represents some deviation term that is independent of S. To prove the above, we need to use the following lemma, also know as McDiarmid's Inequality:

Lemma 21.4 (McDiarmid's Inequality) *Let $f : \mathcal{X} \times \cdots \times \mathcal{X} \to \mathbb{R}$ such that for all $x_i \in \mathcal{X}, 1 \le i \le n$,*

$$|f(x_1, \ldots, x_i, \ldots, x_n) - f(x_1, \ldots, {x'}_i, \ldots, x_n)| \le c_i$$

Then, for any independent random variables $x_1, \ldots, x_n \in \Delta(\mathcal{X})$, we have the following bound:

$$\Pr\left(f(x_1, \ldots, x_n) - \mathbb{E}\left[f(x_1, \ldots, x_n)\right] > t\right) \le \exp\left(\frac{-2t^2}{\sum_{i=1}^{n} c_i^2}\right)$$

Proof: Let $Z_k := \mathbb{E}\left[f(x_1,\ldots,x_n)|x_1,\ldots,x_k\right]$. We first show that Z_k is a martingale[1]. Using the tower property of expectation, we can say the following:

$$\begin{aligned} \mathbb{E}\left[Z_k|x_1,\ldots,x_{k-1}\right] &= \mathbb{E}\left[\mathbb{E}\left[f(x_1,\ldots,x_n)\mid x_1,\ldots,x_k\right]\mid x_1,\ldots,x_{k-1}\right] && (21.1)\\ &= \mathbb{E}\left[f(x_1,\ldots,x_n)\mid x_1,\ldots,x_{k-1}\right] && (21.2)\\ &= Z_{k-1} && (21.3) \end{aligned}$$

This proves that Z_k is a martingale. Now, we can apply Azuma's Inequality to Z_k:

$$\Pr(Z_k - Z_0 > t) \le \exp\left(\frac{-2t^2}{\sum_{i=1}^n c_i^2}\right) \tag{21.4}$$

We know that $Z_0 = \mathbb{E}\left[f(x_1,\ldots,x_n)\right]$ (i.e., the expected value of $f(x_1,\ldots,x_n)$ conditioned on nothing). Additionally, the expected value of $f(x_1,\ldots,x_n)$ conditioned on x_1 through x_n will just equal $f(x_1,\ldots,x_n)$. Based on these facts, we can rewrite the expression above:

$$\begin{aligned} \Pr\left(\mathbb{E}\left[f(x_1,\ldots,x_n)\mid x_1,\ldots,x_n\right] - \mathbb{E}\left[f(x_1,\ldots,x_n)\right] > t\right) &\le \exp\left(\frac{-2t^2}{\sum_{i=1}^n c_i^2}\right) && (21.5)\\ \Pr\left(f(x_1,\ldots,x_n) - \mathbb{E}\left[f(x_1,\ldots,x_n)\right] > t\right) &\le \exp\left(\frac{-2t^2}{\sum_{i=1}^n c_i^2}\right) && (21.6) \end{aligned}$$

This completes the proof. ■

Having proven McDiarmid's Inequality, we are now ready to prove the Symmetrization Lemma

Proof of Theorem 21.3: Let $\Phi(S) = \sup_{g\in G}\left(\mathbb{E}\,g - \hat{\mathbb{E}}_S g\right)$. First, we check that McDiarmid's Inequality can be applied to $\Phi(\cdot)$. To do this, define $S' = S \cup \{x'_i\} \setminus \{x_i\}$. Then, we have the following:

$$\begin{aligned} |\Phi(S) - \Phi(S')| &= \left|\sup_g\left(\mathbb{E}\,g - \hat{\mathbb{E}}_S g\right) - \sup_{g'}\left(\mathbb{E}\,g' - \hat{\mathbb{E}}_{S'} g'\right)\right| && (21.7)\\ &\le \sup_g\left(\hat{\mathbb{E}}_{S'} g - \hat{\mathbb{E}}_S g\right) && (21.8)\\ &= \sup_g \frac{1}{m}\left(\sum_{j\neq i}\left(g(x_j) - g(x_j)\right) + g(x_i) - g(x'_i)\right) && (21.9)\\ &\le \frac{1}{m} && (21.10) \end{aligned}$$

Notice that Inequality (21.10) follows from the fact that g is a binary function. This proves that $\Phi(\cdot)$ satisfies the requirements for McDiarmid's Inequality, with $c_i = \frac{1}{m}$. Using this inequality, we get:

$$\begin{aligned} \Pr\left(\Phi(S) - \mathop{\mathbb{E}}_S\left[\Phi(S)\right] > t\right) &\le \exp\left(\frac{-2t^2}{\sum_{i=1}^m c_i^2}\right) && (21.11)\\ &= \exp\left(\frac{-2t^2}{m\frac{1}{m^2}}\right) && (21.12)\\ &= \exp\left(-2t^2 m\right) && (21.13) \end{aligned}$$

[1] Note that Z_k is known as the Doob martingale.

If we set this last quantity equal to δ and solve for t, we find that $t = \sqrt{\frac{\log \frac{1}{\delta}}{2m}}$. Hence, when we set t to this value, we can say that McDiarmid's Inequality implies the following:

$$\Phi(S) - \mathop{\mathbb{E}}_{S}\left[\Phi(S)\right] \leq \sqrt{\frac{\log \frac{1}{\delta}}{2m}} \tag{21.14}$$

$$\Phi(S) \leq \mathop{\mathbb{E}}_{S}\left[\Phi(S)\right] + \sqrt{\frac{\log \frac{1}{\delta}}{2m}} \tag{21.15}$$

Now, we analyze $\mathbb{E}_S\left[\Phi(S)\right]$ further:

$$\begin{aligned}
\mathbb{E}\left[\Phi(S)\right] &= \mathop{\mathbb{E}}_{S\sim D^m}\left[\sup_{g\in G}\left(\mathbb{E}\, g - \hat{\mathbb{E}}_S g\right)\right] && (21.16)\\
&= \mathop{\mathbb{E}}_{S\sim D^m}\left[\sup_{g\in G}\mathop{\mathbb{E}}_{S'\sim D^m}\left[\hat{\mathbb{E}}_{S'} g - \hat{\mathbb{E}}_S g\right]\right] && (21.17)\\
&\leq \mathop{\mathbb{E}}_{S\sim D^m}\mathop{\mathbb{E}}_{S'\in D^m}\left[\sup_{g\in G}\left(\hat{\mathbb{E}}_{S'} g - \hat{\mathbb{E}}_S g\right)\right] && (21.18)\\
&= \frac{1}{m}\mathop{\mathbb{E}}_{S,S'\sim D^m}\left[\sup_{g\in G}\sum_{i=1}^{m}\left(g({x'}_i) - g(x_i)\right)\right] && (21.19)\\
&= \frac{1}{m}\mathop{\mathbb{E}}_{S,S'\sim D^m}\mathop{\mathbb{E}}_{\sigma_1,\ldots,\sigma_m}\left[\sup_{g\in G}\sum_{i=1}^{m}\left(g({x'}_i) - g(x_i)\right)\sigma_i\right] && (21.20)\\
&\leq \mathop{\mathbb{E}}_{S,S'\sim D^m;\sigma_1,\ldots,\sigma_m}\left[\sup_{g'\in G}\frac{1}{m}\sum_{i=1}^{m} g'(x_i)\sigma_i + \sup_{g\in G}\frac{1}{m}\sum_{i=1}^{m} g(x_i)(-\sigma_i)\right] && (21.21)\\
&= \mathfrak{R}_m(G) + \mathfrak{R}_m(G) && (21.22)\\
&= 2\mathfrak{R}_m(G) && (21.23)
\end{aligned}$$

Equation (21.17) performs a double expectation, which will still result in an average value. Inequality (21.18) relies on the fact that $\sup_\alpha \mathbb{E}_Y[f(\alpha, Y)] \leq \mathbb{E}_Y\left[\sup_\alpha f(\alpha, Y)\right]$. Equation (21.20) introduces i.i.d. Rachemacher random variables. This is still equal to the result in Equation (21.19). When $\sigma_i = 1$, the associated term in the summation remains the same. When $\sigma_i = -1$, the associated term in the summation flips signs. This is the same as swapping this iteration's x_i and ${x'}_i$ between sets S and S'. Because we take the expectation over all S and S', swapping these samples between the sets will not change the total expectation.

Combining this result with Inequality (21.15), we get the desired result:

$$\Phi(S) \leq 2\mathfrak{R}_m(G) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}} \tag{21.24}$$

$$\sup_{g\in G}\left(\mathbb{E}\, g - \hat{E}_S g\right) \leq 2\mathfrak{R}_m(G) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}} \tag{21.25}$$

■

21.2 Growth Function and Massart's Lemma

21.2.1 Growth Function

Definition 21.5 (Growth Function) *Given a binary class of functions H, we define the **growth function** of H as:*

$$\Pi_H(m) = \max_{\{x_1,\ldots,x_m\}\subseteq\mathcal{X}} |\{h(x_1),\ldots,h(x_m) : h \in H\}|$$

This growth function is, like the Rademacher Complexity, a measurement of the richness of a class of functions. Specifically, $\Pi_H(m)$ represents the maximum number of unique ways by which hypotheses in class H can classify m points. The growth function differs from Rademacher Complexity in that it is independent of the distribution of points sampled. Using this definition, we can redefine VC-dimension:

Definition 21.6 (VC-dimension) *For any binary class of functions H, we can define the **VC-dimension** of H as:*

$$VCD(H) = \max\{d : \Pi_H(d) = 2^d\}$$

Using this definition, we know that the following is true:

Fact 21.7 *For any hypothesis class H, if $m \leq VCD(H)$, we know that $\Pi_H(m) = 2^m$.*

However, what happens when $m > VCD(H)$? Ideally, we would like $\Pi_H(m)$ to grow polynomially with respect to m. Otherwise, as we will see after proving Massart's Lemma, the Rademacher Complexity will also grow at an exponential rate.

21.2.2 Massart's Lemma

Theorem 21.8 (Massart's Lemma) *Let A be a finite subset of $\mathbb{R}^m$, and let $\max_{\vec{y}\in A} ||\vec{y}||_2 \leq r$. Then,*

$$\mathop{\mathbb{E}}_{\sigma_1,\ldots,\sigma_m}\left[\sup_{\vec{y}\in A}\sum_{i=1}^m \sigma_i y_i\right] \leq r\sqrt{2\log|A|}$$

Corollary 21.9 *Let H be a binary class of functions. Then,*

$$\mathfrak{R}_m(H) \leq \sqrt{\frac{2\log\Pi_H(m)}{m}}$$

Proof: Given $S = \{x_1,\ldots,x_m\}$, let $A_S = \{(h(x_1),\ldots,h(x_m) : h \in H\}$. Note that $|A_S| \leq \Pi_H(m)$. Using the definition of Empirical Rademacher Complexity, we have that:

$$\hat{\mathfrak{R}}_S(H) = \mathop{\mathbb{E}}_{\sigma_1,\ldots,\sigma_m}\left[\sup_{y\in A_S}\frac{1}{m}\sum_{i=1}^m y_i\sigma_i\right] \tag{21.26}$$

$$= \frac{1}{m}\mathop{\mathbb{E}}_{\sigma_1,\ldots,\sigma_m}\left[\sup_{y\in A_S}\sum_{i=1}^m y_i\sigma_i\right] \tag{21.27}$$

Notice that, since H is a binary classifier, we know that $\max_{\vec{y}\in A_S} ||\vec{y}||_2 \leq \sqrt{m}$. Using this fact, we can now apply Massart's Lemma:

$$\hat{\mathfrak{R}}_S(H) \leq \frac{1}{m}\sqrt{m}\sqrt{2\log|A_S|} \tag{21.28}$$

$$= \sqrt{\frac{2\log|A_S|}{m}} \tag{21.29}$$

$$\leq \sqrt{\frac{2\log\Pi_H(m)}{m}} \tag{21.30}$$

We can now convert from Empirical Rademacher Complexity to Rademacher Complexity:

$$\mathfrak{R}_m(H) = \mathop{\mathbb{E}}_{S}\left[\hat{\mathfrak{R}}_S(H)\right] \leq \mathop{\mathbb{E}}_{S}\left[\sqrt{\frac{2\log \Pi_H(m)}{m}}\right] = \sqrt{\frac{2\log \Pi_H(m)}{m}} \tag{21.31}$$

This completes the proof. ■

21.2.3 Next Time

In the next lecture, we will prove the following:

- If $VCD(H) = d$, then $\Pi_H(m) = O(m^d)$ (Sauer's Lemma)
- Consequently, by Massart's and Sauer's Lemma, $\mathfrak{R}_m(H) \leq \sqrt{\frac{2\log \Pi_H(m)}{m}} \leq \sqrt{\frac{2d\log(m)}{m}}$
- By setting $\epsilon = \sqrt{\frac{2d\log(m)}{m}}$, we see that we need $m \geq \frac{d}{\epsilon^2}$ data points if we want a training error of ϵ.

All of these conclusions imply that the more complex your classifier function is, the more data you will need.