CS 4540/CS 8803: Algorithmic Theory of Intelligence Fall 2023

Lecture 13: Rademacher Complexity

Instructor: Xinyuan Cao *Lecture date: 11/13*

Scribed by: Xinyuan Cao

1 Rademacher Complexity

1.1 Motivation for a new complexity measure

Given a function class $\mathcal{H}$, we have studied how to derive the sample complexity to learn a hypothesis $h \in \mathcal{H}$ using $\mathcal{H}(m)$, the maximum number of distinct ways to label m points. We note that $\mathcal{H}(m)$ is distribution independent. In some cases, the distribution independence is nice because it works for any data distribution. On the other hand, the bound we get may not be tight for some distributions that are benign than the worst case. Furthermore, $\mathcal{H}(m)$ only applies to binary classification, but we are often interested in generalization bounds for multi-class classification and regression as well. Here we introduce Rademacher complexity, which is a more modern notion of complexity that is distribution dependent and defined for any class real-valued functions.

1.2 Definitions

Given a space Z and a distribution $D|_Z$. Let $S = \{z_1, \cdots, z_m\}$ be a set of samples drawn iid from $D|_Z$. Let $\mathcal{H}$ be a class of functions $h : Z \to \mathbb{R}$.

Definition 1 (Empirical Rademacher Complexity)**.** *The empirical Rademacher Complexity of $\mathcal{H}$ is defined as*

$$\hat{R}_m(\mathcal{H}) = \mathop{\mathbb{E}}_{\sigma \overset{iid}{\sim} \{\pm 1\}} \left[\sup_{h \in \mathcal{H}} \left(\frac{1}{m} \sum_{i=1}^{m} \sigma_i h(z_i) \right) \right]$$

where $\sigma_1, \cdots, \sigma_m$ are independent random variables uniformly chosen from $\{-1, 1\}$. We call such random variables as Rademacher variables.

Definition 2 (Rademacher Complexity)**.** *The Rademacher Complexity of $\mathcal{H}$ is defined as*

$$R_m(\mathcal{H}) = \mathbb{E}_D[\hat{R}_m(\mathcal{H})] = \mathop{\mathbb{E}}_{z \overset{iid}{\sim} D} \left[\mathop{\mathbb{E}}_{\sigma \overset{iid}{\sim} \{\pm 1\}} \left[\sup_{h \in \mathcal{H}} \left(\frac{1}{m} \sum_{i=1}^{m} \sigma_i h(z_i) \right) \right] \right]$$

Intuitively, the supreme measures the maximum correlation between $h(z_i)$ and σ_i over $h \in \mathcal{H}$. By taking expectation over the Rademacher variable σ, the empirical Rademacher complexity of $\mathcal{H}$ quantifies the capacity of functions from $\mathcal{H}$ to fit random noise on a fixed sample set S. The Rademacher Complexity then evaluates the extent to which $\mathcal{H}$ can fit noise on a dataset drawn from the distribution D.

1.3 Dependence on Distribution

To give the intuition that Rademacher complexity depends on the distribution, we consider an extreme example P as a point mass. That is $z = z_0$ almost surely. Assume that $-1 \leq h(z_0) \leq 1$ for

all $h \in \mathcal{H}$. Then

$$\begin{aligned}
R_m(\mathcal{H}) &= \mathop{\mathbb{E}}_{z \overset{iid}{\sim} D}\left[\mathop{\mathbb{E}}_{\sigma \overset{iid}{\sim}\{\pm 1\}}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i h(z_i)\right)\right]\right] \\
&= \mathop{\mathbb{E}}_{\sigma \overset{iid}{\sim}\{\pm 1\}}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{m}h(z_0)\sum_{i=1}^{m}\sigma_i\right)\right] \\
&\leq \mathop{\mathbb{E}}_{\sigma \overset{iid}{\sim}\{\pm 1\}}|\frac{1}{m}\sum_{i=1}^{m}\sigma_i| \\
&\leq \mathop{\mathbb{E}}_{\sigma \overset{iid}{\sim}\{\pm 1\}}\sqrt{\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i\right)^2} \\
&\leq \frac{1}{m}\sqrt{\mathbb{E}_\sigma\left(\sum_{i=1}^{m}\sigma_i\right)^2} && \text{Jensen's Inequality} \\
&= \frac{1}{m}\sqrt{\mathbb{E}_{\sigma_i,\sigma_j}\sum_{i,j=1}^{m}\sigma_i\sigma_j} \\
&= \frac{1}{m}\sqrt{\mathbb{E}_{\sigma_i}\sum_{i=1}^{m}\sigma_i^2} \\
&= \frac{1}{m}\sqrt{m} = \frac{1}{\sqrt{m}}
\end{aligned}$$

This bound does not depend on $\mathcal{H}$. This example illustrates that a bound on the Rademacher complexity can sometimes depend only on the (known) distribution of the Rademacher random variables.

2 Generalization Bounds

2.1 Uniform Convergence

A central objective of learning theory is to achieve strong generalization, meaning that a hypothesis that performs well on the training data should also perform well on unseen data. We now present a uniform convergence result utilizing Rademacher complexity, which allows us to establish a more comprehensive generalization bound.

For $h \in \mathcal{H}$, we denote the empirical mean over a sample S as $\hat{\mathbb{E}}_S[f(z)] := \frac{1}{|S|}\sum_{z\in S} f(z)$. Then we have the following theorem.

Theorem 3. *For a distribution $D|_Z$ and parameter $\delta \in (0,1)$. If $\mathcal{H} \subseteq \{h : Z \to [a, a+1]\}$ and $S = \{z_1, \cdots, z_m\}$ is drawn i.i.d. from $D|_Z$ then with probability $\geq 1-\delta$ over the draw of S, for any $h \in \mathcal{H}$,*

$$\mathbb{E}_D[h(z)] \leq \hat{\mathbb{E}}_S[h(z)] + 2R_m(\mathcal{H}) + \sqrt{\frac{\ln(1/\delta)}{2m}}.$$

Furthermore, with probability $1-\delta$, for any $h \in \mathcal{H}$,

$$\mathbb{E}_D[h(z)] \leq \hat{\mathbb{E}}_S[h(z)] + 2\hat{R}_m(\mathcal{H}) + 3\sqrt{\frac{\ln(2/\delta)}{2m}}.$$

2.2 Generalization

Consider a binary classification, where $X = \mathbb{R}^d, Y = \{\pm 1\}$. For any $h : X \to Y$, we consider the 0-1 loss function $l_h : Z \to R$. That is,

$$l_h(z) = l_h(x, y) = \mathbb{1}_{h(x)\neq y} = \frac{1 - yh(x)}{2}.$$

Let $L(\mathcal{H}) = \{l_h : h \in \mathcal{H}\}$, then we can compute its Rademacher complexity as

$$\begin{aligned} R_m(L(\mathcal{H})) &= \mathop{\mathbb{E}}_{x_i,y_i,\sigma_i}\left[\sup_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i \frac{1 - y_i h(x_i)}{2}\right] \\ &= \mathop{\mathbb{E}}_{x_i,y_i,\sigma_i}\left[\frac{1}{m}\sum_{i=1}^{m}\sigma_i + \frac{1}{2}\sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m} -\sigma_i y_i h(x_i)\right] \\ &= \frac{1}{2}\mathop{\mathbb{E}}_{x_i,y_i,\sigma_i}\left[\sup_{h\in\mathcal{H}}\sigma_i h(x_i)\right] \\ &= \frac{1}{2}R_m(\mathcal{H}) \end{aligned}$$

Denote $\hat{\text{err}}_S(h) := \hat{\mathbb{E}}_S[l_h(z)] = \frac{1}{m}\sum_{i=1}^{m} l_h(z_i)$ as the training error and and $\text{err}_D(h) := \mathbb{E}_D[l_h(z)]$ as the generalization error. By Theorem 3, we get can bound the generalization error as follows.

Corollary 4. *For $\mathcal{H} \subset \{h : X \to Y\}$ and $0-1$ loss, we have*

$$err_D(h) \leq \hat{err}_S(h) + R_m(\mathcal{H}) + \sqrt{\frac{\ln(1/\delta)}{2m}}$$

2.3 Proof

To prove the generalization bounds using Rademacher complexity, we will use the following concentration bound.

Theorem 5 (McDiarmid Inequality)**.** *Let $x_1, \cdots, x_n$ be independent random variables taking on values in a set A, and let $c_1, \cdots, c_n$ be positive real constants. If $f : A^n \to \mathbb{R}$ satisfies*

$$\sup_{x_1,\cdots,x_n,x_i'\in A} |f(x_1, \cdots, x_i, \cdots, x_n) - f(x_1, \cdots, x_i', \cdots, x_n)| \leq c_i,$$

for $1 \leq i \leq n$, then

$$\Pr[f(x_1, \cdots, x_n) \geq \mathbb{E}[f(x_1, \cdots, x_n)] + \epsilon] \leq e^{\frac{-2\epsilon^2}{\sum_{i=1}^{n} c_i^2}}$$

Proof of Theorem 3. For a fixed function $h \in \mathcal{H}$, by the definition of supremum,

$$\mathbb{E}_D[h(z)] \leq \hat{\mathbb{E}}_S[h(z)] + \sup_{g\in\mathcal{H}}\Big(\mathbb{E}_D[g(z)] - \hat{\mathbb{E}}_S[g(z)]\Big)$$

Denote

$$\phi(S) = \sup_{g\in\mathcal{H}}\Big(\mathbb{E}_D[g(z)] - \hat{\mathbb{E}}_S[g(z)]\Big)$$

We would like to bound ϕ in terms of its expectation by McDiarmid Inequality. Specifically we claim that

$$\sup_{z_1,\cdots,z_m,z_m'} |\phi(z_1, \cdots, z_i, \cdots, z_m) - \phi(z_1, \cdots, z_i', \cdots, z_m)| \leq \frac{1}{m}.$$

We leave the proof in Lemma 6. Then we apply McDiarmid Inequality and we have

$$\Pr[\phi(S) \geq \mathbb{E}[\phi(S)] + t] \leq e^{\frac{-2t^2}{\sum_{i=1}^{m}\frac{1}{m^2}}} = e^{-2t^2 m}$$

So for $t \geq \sqrt{\frac{\ln(1/\delta)}{2m}}$, the probability is less than δ. That is with probability at least $1-\delta$,

$$\mathbb{E}_D[h(z)] \leq \hat{\mathbb{E}}_S[h(z)] + \mathbb{E}_S\left[\sup_{g\in\mathcal{H}}\left(\mathbb{E}_D[g(z)] - \hat{\mathbb{E}}_S[g(z)]\right)\right] + \sqrt{\frac{\ln(1/\delta)}{2m}}. \tag{1}$$

Next we will bound the expectation of $\phi(S)$ in terms of the Rademacher complexity of $\mathcal{H}$. Let $\tilde{S} = \{\tilde{z}_1, \cdots, \tilde{z}_m\}$ be a set independently drawn from S. Since

$$\mathbb{E}_{\tilde{S}}[\hat{\mathbb{E}}_{\tilde{S}}[g(z)]] = \mathbb{E}_D[g(z)], \quad \mathbb{E}_{\tilde{S}}[\hat{\mathbb{E}}_S[g(z)]] = \hat{\mathbb{E}}_S[g(z)]$$

we can rewrite the expectation

$$\begin{aligned}
\mathbb{E}_S\left[\sup_{g\in\mathcal{H}}\left(\mathbb{E}_D[g(z)] - \hat{\mathbb{E}}_S[g(z)]\right)\right] =& \mathbb{E}_S\left[\sup_{g\in\mathcal{H}}\mathbb{E}_{\tilde{S}}\left[\hat{\mathbb{E}}_{\tilde{S}}[g(z)] - \hat{\mathbb{E}}_S[g(z)]\right]\right] \\
=& \mathbb{E}_S\left[\sup_{g\in\mathcal{H}}\mathbb{E}_{\tilde{S}}\left[\frac{1}{m}\sum_{i=1}^m g(\tilde{z}_i) - \frac{1}{m}\sum_{i=1}^m g(z_i)\right]\right] \\
=& \mathbb{E}_S\left[\sup_{g\in\mathcal{H}}\mathbb{E}_{\tilde{S}}\left[\frac{1}{m}\sum_{i=1}^m (g(\tilde{z}_i) - g(z_i))\right]\right]
\end{aligned}$$

Since sup is convex, we use Jensen's Inequality to move sup inside the expectation.

$$\mathbb{E}_S\left[\sup_{g\in\mathcal{H}}\mathbb{E}_{\tilde{S}}\left[\frac{1}{m}\sum_{i=1}^m (g(\tilde{z}_i) - g(z_i))\right]\right] \leq \mathbb{E}_{S,\tilde{S}}\left[\sup_{g\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^m (g(\tilde{z}_i) - g(z_i))\right]$$

Since S and S' are iid, multiplying each term in the summation by a Rademacher variable σ_i does not change the distribution. That gives us

$$\begin{aligned}
\mathbb{E}_{S,\tilde{S}}\left[\sup_{g\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^m (g(\tilde{z}_i) - g(z_i))\right] =& \mathbb{E}_{S,\tilde{S},\sigma}\left[\sup_{g\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^m \sigma_i(g(\tilde{z}_i) - g(z_i))\right] \\
\leq& \mathbb{E}_{S,\tilde{S},\sigma}\left[\sup_{g\in\mathcal{H}}\left(\frac{1}{m}\sum_{i=1}^m \sigma_i g(\tilde{z}_i)\right) + \sup_{g\in\mathcal{H}}\left(\frac{1}{m}\sum_{i=1}^m \sigma_i g(\tilde{z}_i)\right)\right] \\
=& \mathbb{E}_{\sigma,S}\left[\sup_{g\in\mathcal{H}}\left(\frac{1}{m}\sum_{i=1}^m \sigma_i g(z_i)\right)\right] + \mathbb{E}_{\sigma,\tilde{S}}\left[\sup_{g\in\mathcal{H}}\left(\frac{1}{m}\sum_{i=1}^m \sigma_i g(\tilde{z}_i)\right)\right] \\
=& 2R_m(\mathcal{H})
\end{aligned}$$

Combining with Equation (1), we get

$$\mathbb{E}_D[h(z)] \leq \hat{\mathbb{E}}_S[h(z)] + 2R_m(\mathcal{H}) + \sqrt{\frac{\ln(1/\delta)}{2m}}.$$

To show the second half, we note that $\hat{R}_m(\mathcal{H})$ satisfies McDiarmid's Inequality with constant $1/m$. So we can apply McDiarmid's Inequality to bound $\hat{R}_m(\mathcal{H})$ using $R_m(\mathcal{H})$. □

Lemma 6. *Denote*

$$\phi(S) = \sup_{g\in\mathcal{H}}\left(\mathbb{E}_D[g(z)] - \hat{\mathbb{E}}_S[g(z)]\right)$$

Then we have

$$\sup_{z_1,\cdots,z_m,z'_m} |\phi(z_1,\cdots,z_i,\cdots,z_m) - \phi(z_1,\cdots,z'_i,\cdots,z_m)| \leq \frac{1}{m}.$$

Proof. Let $S = z_1, \cdots, z_m, S' = z_1, \cdots, z'_j, \cdots, z_m$. By definition,

$$|\phi(S) - \phi(S')| = \left| \sup_{h \in \mathcal{H}} \left(\mathbb{E}_D[h(z)] - \hat{\mathbb{E}}_S[h(z)] \right) - \sup_{h \in \mathcal{H}} \left(\mathbb{E}_D[h(z)] - \hat{\mathbb{E}}_{S'}[h(z)] \right) \right|$$

Let $h^* \in \mathcal{H}$ be the maximizing function for the supremum in $\phi(S)$. Then by definition of supremum, h^* can at best maximize the $\phi(S')$ term as well. Then we have

$$\begin{aligned}
|\phi(S) - \phi(S')| &= \left| \mathbb{E}_D[h(z)] - \hat{\mathbb{E}}_S[h(z)] - \sup_{h \in \mathcal{H}} \left(\mathbb{E}_D[h(z)] - \hat{\mathbb{E}}_{S'}[h(z)] \right) \right| \\
&\leq \left| \mathbb{E}_D[h(z)] - \hat{\mathbb{E}}_S[h(z)] - \mathbb{E}_D[h^*(z)] + \hat{\mathbb{E}}_{S'}[h^*(z)] \right| \\
&= \left| \hat{\mathbb{E}}_{S'}[h^*(z)] - \hat{\mathbb{E}}_S[h^*(z)] \right| \\
&= \left| \frac{1}{m} \sum_{z \in S} h^*(z) - \frac{1}{m} \sum_{z \in S'} h^*(z_i) \right|
\end{aligned}$$

Since S and S' differ in only one element,

$$\begin{aligned}
|\phi(S) - \phi(S')| \leq & \frac{1}{m} \left| \sum_{i \neq j} h^*(z_i) - \sum_{i \neq j} h^*(z_i) + h^*(z_j) - h^*(z'_j) \right| \\
= & \frac{1}{m} \left| h^*(z_j) - h^*(z'_j) \right| \\
\leq & \frac{1}{m}
\end{aligned}$$

□