CMSC 35900 (Spring 2008) Learning Theory **Lecture: 17**

Rademacher Composition and Linear Prediction

Instructors: Sham Kakade and Ambuj Tewari

1 Rademacher Composition

In order to use our Rademacher bound, we need to find Rademacher complexities of loss classes. It is usually more conveniently to directly work with our hypothesis class, than a loss class. To do this, we need to understand how Rademacher complexities compose with loss classes. To this end, the follow lemma is useful.

Lemma 1.1. *(Composition Lemma) Assume that $\phi : \mathbb{R} \to \mathbb{R}$ is a L_ϕ-Lipschitz continuous function, i.e. $|\phi(t)-\phi(s)| \leq L_\phi|t-s|$. Let $\mathcal{F} \subset \mathbb{R}^m$. Then*

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^{m}\epsilon_i\phi(f_i)\right] \leq L_\phi\mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^{m}\epsilon_i\phi(f_i)\right]$$

In other words:

$$\mathfrak{R}(\phi(\mathcal{F})) \leq L_\phi\mathfrak{R}(\mathcal{F})$$

Proof. We prove the case where $L_\phi = 1$. The general proof follows from this case.

Let us consider a class of vector valued functions $\Psi = \{\psi = (\psi_1, \psi_2, \dots \psi_m)\}$ where each $\psi \in \Psi$ is a function where each ψ_i is either ϕ or the identity function I. We will prove that:

$$\sup_{\psi\in\Psi} \mathfrak{R}(\psi(\mathcal{F})) \leq \mathfrak{R}(\mathcal{F})$$

The claim follows from this.

Let ψ be some function in which at least one component is not the identity function. Without loss of generality, assume this is the first component, i.e.

$$\psi = (\phi, \psi_2, \dots, \psi_m)$$

Define

$$\psi' = (I, \psi_2, \dots, \psi_m)$$

We will now prove that:

$$\mathfrak{R}(\psi(\mathcal{F})) \leq \mathfrak{R}(\psi'(\mathcal{F}))$$

The previous claim follows from this since we can flip any component that is ϕ to the identity function, without decreasing the Rademacher complexity.

To prove this, we start by making the expectation explicit in the first Rademacher number ϵ_1:

$$\begin{aligned}
& \mathfrak{R}(\psi(\mathcal{F})) \\
= \quad & \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^{m}\epsilon_i\psi_i(f_i)\right] \\
= \quad & \frac{1}{2m}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(\phi(f_1)+\sum_{i=2}^{m}\epsilon_i\psi_i(f_i)\right)+\sup_{f\in\mathcal{F}}\left(-\phi(f_1)+\sum_{i=2}^{m}\epsilon_i\psi_i(f_i)\right)\right] \\
= \quad & \frac{1}{2m}\mathbb{E}\left[\sup_{f,f'\in\mathcal{F}}\left(\phi(f_1)+\sum_{i=2}^{m}\epsilon_i\psi_i(f_i)-\phi(f'_1)+\sum_{i=2}^{m}\epsilon_i\psi_i(f'_i)\right)\right] \\
\leq \quad & \frac{1}{2m}\mathbb{E}\left[\sup_{f,f'\in\mathcal{F}}\left(|f_1-f'_1|+\sum_{i=2}^{m}\epsilon_i\psi_i(f_i)+\sum_{i=2}^{m}\epsilon_i\psi_i(f'_i)\right)\right] \\
= \quad & \frac{1}{2m}\mathbb{E}\left[\sup_{f,f'\in\mathcal{F}}\left(f_1-f'_1+\sum_{i=2}^{m}\epsilon_i\psi_i(f_i)+\sum_{i=2}^{m}\epsilon_i\psi_i(f'_i)\right)\right] \\
= \quad & \frac{1}{2m}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(f_1+\sum_{i=2}^{m}\epsilon_i\psi_i(f_i)\right)+\sup_{f\in\mathcal{F}}\left(-f_1+\sum_{i=2}^{m}\epsilon_i\psi_i(f_i)\right)\right] \\
= \quad & \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^{m}\epsilon_i\psi'_i(f_i)\right] \\
= \quad & \mathfrak{R}(\psi'(\mathcal{F}))
\end{aligned}$$

We are able to drop the absolute value (in the step after the inequality), since it is clear that the sup will be achieved when this function is positive. This completes the proof. □

2 Linear Prediction

Let us assume that our loss function is of the form $\phi(w \cdot x, y)$. Let us also consider the empirical risk minimization algorithms:

$$\hat{w}_2 = \text{argmin}_{w:||w||_2\leq W_2}\sum_{i=1}^{m}\ell(w\cdot x_i, y_i)$$
$$\hat{w}_1 = \text{argmin}_{w:||w||_1\leq W_1}\sum_{i=1}^{m}\ell(w\cdot x_i, y_i)$$

These problems are closely related to the L_1 and L_2 regularization (these are essentially the dual problems).

Let us now understand the generalization ability of these algorithms.

2.1 Rademacher Bounds for Linear Classes

Let $\mathcal{F}$ be the class of linear predictors.

Lemma 2.1. *Let $\mathcal{F}$ be the class of linear predictors, with the L_2-norm of the weights bounded by W_2. Also assume that with probability one that $||x||_2 \leq X_2$. Then*

$$\mathfrak{R}(\mathcal{F}) \leq \frac{X_2 W_2}{\sqrt{m}}$$

Proof. Let $\mathcal{F}_{x_1,x_2,\ldots,x_m}$ be the class:

$$\{(w \cdot x_1, w \cdot x_2, \ldots w \cdot x_m) : ||w||_2 \leq W_2\}$$

We now bound this empirical Rademacher complexity:

$$\begin{aligned}
\mathfrak{R}(\mathcal{F}) &= \frac{1}{m}\mathbb{E}\left[\sup_{w:||w||_2 \leq W_2} \sum_{i=1}^{m} \epsilon_i w \cdot x_i\right] \\
&= \frac{1}{m}\mathbb{E}\left[\sup_{w:||w||_2 \leq W_2} w \cdot \sum_{i=1}^{m} \epsilon_i x_i\right] \\
&= \frac{W_2}{m}\mathbb{E}\left[||\sum_{i=1}^{m} \epsilon_i x_i||_2\right] \\
&\leq \frac{W_2}{m}\sqrt{\mathbb{E}\left[\sum_{i=1}^{m} ||\epsilon_i x_i||_2^2\right]} \\
&= \frac{W_2}{m}\sqrt{\mathbb{E}\left[\sum_{i=1}^{m} ||x_i||_2^2\right]} \\
&= \frac{X_2 W_2}{\sqrt{m}}
\end{aligned}$$

where we have used Jensen's inequality. □

Lemma 2.2. *Let $\mathcal{F}$ be the class of linear predictors, with the L_1-norm of the weights bounded by W_1. Also assume that with probability one that $||x||_\infty \leq X_\infty$. Then*

$$\mathfrak{R}(\mathcal{F}) \leq X_\infty W_1 \sqrt{\frac{2 \log d}{m}}$$

where d is the dimensionality of x.

Proof. Let $\mathcal{F}_{x_1,x_2,\ldots,x_m}$ be the class:

$$\{(w \cdot x_1, w \cdot x_2, \ldots w \cdot x_m) : ||w||_1 \leq W_1\}$$

Using the definition of the dual norms, we now bound this empirical Rademacher complexity:

$$\begin{aligned}
\mathfrak{R}(\mathcal{F}) &= \frac{1}{m}\mathbb{E}\left[\sup_{w:||w||_1 \leq W_1} \sum_{i=1}^{m} \epsilon_i w \cdot x_i\right] \\
&= \frac{1}{m}\mathbb{E}\left[\sup_{w:||w||_1 \leq W_1} w \cdot \sum_{i=1}^{m} \epsilon_i x_i\right] \\
&= \frac{W_1}{m}\mathbb{E}\left[||\sum_{i=1}^{m} \epsilon_i x_i||_\infty\right] \\
&= \frac{W_1}{m}\mathbb{E}\left[\sup_j \sum_{i=1}^{m} \epsilon_i [x_i]_j\right] \\
&\leq \frac{W_1\sqrt{2 \log d}}{m} \sup_j \sqrt{\sum_{i=1}^{m} [x_i]_j^2} \\
&\leq X_\infty W_1 \sqrt{\frac{2 \log d}{m}}
\end{aligned}$$

where we have used Massart's finite lemma. □

2.2 Generalization

Corollary 2.3. *Under the assumptions above, for the $L2$ case, we have:*

$$\mathcal{L}(\hat{w}_2) - \text{argmin}_{w:||w||_2 \leq W_2} \mathcal{L}(w) \leq 2L_\phi \frac{X_2 W_2}{\sqrt{m}} + 2\sqrt{\frac{\log 2/\delta}{2m}}$$

and for the $L1$ case, we have:

$$\mathcal{L}(\hat{w}_1) - \text{argmin}_{w:||w||_1 \leq W_1} \mathcal{L}(w) \leq 2L_\phi X_\infty W_1 \sqrt{\frac{2\log d}{m}} + 2\sqrt{\frac{\log 2/\delta}{2m}}$$

The proof just follow from the previous lemmas, along with our Rademacher bound for loss classes.

3 Comparison to Online to Batch Conversion

If we assume that $\phi(w \cdot x, y)$ is convex, as a function of w, then we can run an online algorithm (with the constrained decision space, either W_1 or W_2) and then do an online to batch conversion. Here, it is easy to see that $G_2 \leq L_\phi X_2$ and $G_\infty \leq L_\phi X_\infty$. Using this in our previous bounds, we find that the online to batch conversions are just as sharp as the previous bounds (the constants are slightly better).