EECS 598: Statistical Learning Theory, Winter 2014 **Topic 15**

Rademacher Complexity of Kernel Classes

Lecturer: Clayton Scott *Scribe: Brian Segal*

1 Introduction

In the last lecture we discussed surrogate losses, such as the hinge loss that gives rise to SVMs. We frequently use a surrogate loss instead of the 0-1 loss because surrogate losses lead to more tractable optimization probelms. However, our underlying goal is usually to minimize the 0-1 loss. We previously showed that if L is classification calibrated, then consistency w.r.t R_L implies consistency w.r.t. R. The next step is to establish the consistency of kernel methods with respect to R_L. Moving in this direction, in this lecture we use Rademacher complexity to derive a uniform deviation bound, over a ball in a RKHS, for risks based on surrogate losses that are Lipschitz continuous.

The bounds in this lecture do not require that $\mathcal{Y}$ be finite, so here $\mathcal{Y}$ refers to the response space of a classification or regression problem.

2 Lipschitz Composition Property of Rademacher Complexity

Lemma 1. *(Zhang and Meir, 2003) Suppose $\{\phi_i\}, \{\psi_i\}, i = 1, \ldots, n,$ are two sets of functions on Θ such that for each i and $\theta, \theta' \in \Theta, |\phi_i(\theta) - \phi_i(\theta')| \le |\psi_i(\theta) - \psi_i(\theta')|$. Then for all functions $c : \Theta \to \mathbb{R}$,*

$$\mathbb{E}\left[\sup_\theta \left\{c(\theta) + \sum_{i=1}^n \sigma_i \phi_i(\theta)\right\}\right] \le \mathbb{E}\left[\sup_\theta \left\{c(\theta) + \sum_{i=1}^n \sigma_i \psi_i(\theta)\right\}\right].$$

Before proving Lemma 1, we state some useful corollaries.

Corollary 1. *Let $\mathcal{G} \subseteq [a,b]^{\mathcal{X}}$ and suppose $\tau : \mathbb{R} \to \mathbb{R}$ is C-Lipschitz continuous. Then for any $S = (Z_1, \ldots, Z_n)$,*

$$\widehat{\mathfrak{R}}_S(\tau \circ \mathcal{G}) \le C \widehat{\mathfrak{R}}_S(\mathcal{G})$$

where $\tau \circ \mathcal{G} = \{x \mapsto \tau(g(x)) \,|\, g \in \mathcal{G}\}$.

Proof. Apply lemma 1 with $\Theta = \mathcal{G}, \theta = g, \phi_i(g) = \tau(g(Z_i)), \psi_i(g) = Cg(Z_i)$, and $c(\theta) = 0$. Since τ is C-Lipschitz continuous, $|\tau(g(Z_i)) - \tau(g(Z_i'))| \le C|g(Z_i) - g(Z_i')|$, so the conditions of the lemma hold. Then dividing both sides of the inequality by n, the LHS becomes $\widehat{\mathfrak{R}}_S(\tau \circ \mathcal{G})$ and the RHS becomes $\widehat{\mathfrak{R}}_S(\mathcal{G})$. □

Corollary 2. *Suppose $\mathcal{F} \subseteq [a,b]^{\mathcal{X}}$ and $L : \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ is a loss such that $L(y, \cdot)$ is C-Lipschitz $\forall y \in \mathcal{Y}$. Then for all $S = ((X_1, Y_1), \ldots, (X_n, Y_n))$,*

$$\widehat{\mathfrak{R}}_S(L \circ \mathcal{F}) \le C \widehat{\mathfrak{R}}_S(\mathcal{F})$$

where $L \circ \mathcal{F} = \{(x, y) \mapsto L(y, f(x)) \,|\, f \in \mathcal{F}\}$.

Proof. Let $\Theta = \mathcal{F}, \theta = f, \phi_i(f) = L(Y_i, f(X_i)), \psi_i(f) = Cf(X_i)$, and $c(\theta) = 0$. Now argue as in the proof of Corollary 1. □

Note. Corollaries 1 and 2 are similar except for the domains of the function classes; whereas the domains of the functions classes are the same in Corollary 1, they are different in Corollary 2.

Proof of Lemma 1. By induction. The lemma holds for $n = 0$, because in that case we only have $c(\theta)$ on both sides of the inequality. Now suppose the lemma holds for some $n > 0$ and let σ be a vector of Rademacher variables. Then

$$
\begin{aligned}
&\mathbb{E}_\sigma \left[\sup_\theta \left\{ c(\theta) + \sum_{i=1}^{n+1} \sigma_i \phi_i(\theta) \right\} \right] \\
&\quad = \mathbb{E}_{\sigma_1,\dots,\sigma_n} \mathbb{E}_{\sigma_{n+1}} \left[\sup_\theta \left\{ c(\theta) + \sum_{i=1}^{n+1} \sigma_i \phi_i(\theta) \right\} \right] \\
&\quad = \mathbb{E}_{\sigma_1,\dots,\sigma_n} \left[\frac{1}{2} \sup_\theta \left\{ c(\theta) + \sum_{i=1}^{n} \sigma_i \phi_i(\theta) + \phi_{n+1}(\theta) \right\} + \frac{1}{2} \sup_{\theta'} \left\{ c(\theta') + \sum_{i=1}^{n} \sigma_i \phi_i(\theta') - \phi_{n+1}(\theta') \right\} \right] && (1) \\
&\quad = \mathbb{E}_{\sigma_1,\dots,\sigma_n} \left[\sup_{\theta,\theta'} \left\{ \frac{c(\theta) + c(\theta')}{2} + \sum_{i=1}^{n} \sigma_i \frac{\phi_i(\theta) + \phi_i(\theta')}{2} + \frac{\phi_{n+1}(\theta) - \phi_{n+1}(\theta')}{2} \right\} \right] \\
&\quad = \mathbb{E}_{\sigma_1,\dots,\sigma_n} \left[\sup_{\theta,\theta'} \left\{ \frac{c(\theta) + c(\theta')}{2} + \sum_{i=1}^{n} \sigma_i \frac{\phi_i(\theta) + \phi_i(\theta')}{2} + \frac{|\phi_{n+1}(\theta) - \phi_{n+1}(\theta')|}{2} \right\} \right] && (2) \\
&\quad \le \mathbb{E}_{\sigma_1,\dots,\sigma_n} \left[\sup_{\theta,\theta'} \left\{ \frac{c(\theta) + c(\theta')}{2} + \sum_{i=1}^{n} \sigma_i \frac{\phi_i(\theta) + \phi_i(\theta')}{2} + \frac{|\psi_{n+1}(\theta) - \psi_{n+1}(\theta')|}{2} \right\} \right] && \text{(Lemma 1 conditions)} \\
&\quad = \mathbb{E}_{\sigma_1,\dots,\sigma_n} \left[\sup_{\theta,\theta'} \left\{ \frac{c(\theta) + c(\theta')}{2} + \sum_{i=1}^{n} \sigma_i \frac{\phi_i(\theta) + \phi_i(\theta')}{2} + \frac{\psi_{n+1}(\theta) - \psi_{n+1}(\theta')}{2} \right\} \right] && \text{(same as (2) above)} \\
&\quad = \mathbb{E}_{\sigma_1,\dots,\sigma_n} \mathbb{E}_{\sigma_{n+1}} \left[\sup_\theta \left\{ c(\theta) + \sigma_{n+1} \psi_{n+1}(\theta) + \sum_{i=1}^{n} \sigma_i \phi_i(\theta) \right\} \right] && \text{(reversing above steps, applied to } \psi_{n+1}\text{)} \\
&\quad = \mathbb{E}_{\sigma_{n+1}} \mathbb{E}_{\sigma_1,\dots,\sigma_n} \left[\sup_\theta \left\{ c(\theta) + \sigma_{n+1} \psi_{n+1}(\theta) + \sum_{i=1}^{n} \sigma_i \phi_i(\theta) \right\} \right] \\
&\quad \le \mathbb{E}_{\sigma_{n+1}} \mathbb{E}_{\sigma_1,\dots,\sigma_n} \left[\sup_\theta \left\{ c(\theta) + \sigma_{n+1} \psi_{n+1}(\theta) + \sum_{i=1}^{n} \sigma_i \psi_i(\theta) \right\} \right] && \text{(induction hypothesis)} \\
&\quad = \mathbb{E}_\sigma \left[\sup_\theta \left\{ c(\theta) + \sum_{i=1}^{n+1} \sigma_i \psi_i(\theta) \right\} \right]
\end{aligned}
$$

Where (1) follows because σ_{n+1} is uniformly distributed on $\{-1, +1\}$. To see that (2) holds, note that if $\phi_{n+1}(\theta) < \phi_{n+1}(\theta')$, then swapping θ and θ' increases the last term while leaving the others fixed.

□

3 Kernel Classes

Theorem 1. *Suppose k is a bounded kernel with $\sup_x \sqrt{k(x,x)} = B < \infty$ and let $\mathcal{F}$ be its RKHS. Let $M > 0$ be fixed. Then for any $S = (X_1, ..., X_n)$,*

$$\widehat{\mathfrak{R}}_S(B_k(M)) \le \frac{MB}{\sqrt{n}}$$

where $B_k(M) = \{f \in \mathcal{F} \mid \|f\|_{\mathcal{F}} \le M\}$

Proof. Fix $S = (X_1, ..., X_n)$. Then

$$
\begin{aligned}
\widehat{\mathfrak{R}}_S(B_k(M)) &= \mathbb{E}_\sigma \left[\sup_{f \in B_k(M)} \frac{1}{n} \sum_{i=1}^n \sigma_i f(X_i) \right] \\
&= \frac{1}{n} \mathbb{E}_\sigma \left[\sup_{f \in B_k(M)} \sum_{i=1}^n \sigma_i \langle f, k(\cdot, X_i) \rangle \right] \\
&= \frac{1}{n} \mathbb{E}_\sigma \left[\sup_{f \in B_k(M)} \left\langle f, \sum_{i=1}^n \sigma_i k(\cdot, X_i) \right\rangle \right] && \text{(linearity of inner product)} \\
&= \frac{1}{n} \mathbb{E} \left[\left\langle M \frac{\sum_{i=1}^n \sigma_i k(\cdot, X_i)}{\|\sum_{i=1}^n \sigma_i k(\cdot, X_i)\|}, \sum_{i=1}^n \sigma_i k(\cdot, X_i) \right\rangle \right] && \text{(Cauchy-Schwartz condition for equality)} \\
&= \frac{M}{n} \mathbb{E}_\sigma \left[\left\| \sum_{i=1}^n \sigma_i k(\cdot, X_i) \right\| \right] \\
&= \frac{M}{n} \mathbb{E}_\sigma \left[\sqrt{ \left\| \sum_{i=1}^n \sigma_i k(\cdot, X_i) \right\|^2 } \right] \\
&\le \frac{M}{n} \sqrt{ \mathbb{E}_\sigma \left\| \sum_{i=1}^n \sigma_i k(\cdot, X_i) \right\|^2 } && \text{(Jensen's inequality)} \\
&= \frac{M}{n} \sqrt{ \sum_{i=1}^n \|k(\cdot, X_i)\|^2 } && (\mathbb{E}_\sigma [\sigma_i \sigma_j] = 0, i \neq j) \\
&= \frac{M}{n} \sqrt{ \sum_{i=1}^n k(X_i, X_i) } && \text{(reproducing property)} \\
&\le \frac{M}{n} \sqrt{nB^2} \\
&= \frac{MB}{\sqrt{n}}.
\end{aligned}
$$

$\square$

Note. $\frac{M}{n}\sqrt{\sum_{i=1}^n k(X_i, X_i)} = \frac{M}{n}\sqrt{\mathrm{tr}(K)}$, where K is the kernel matrix. Also, any kernel on a compact set satisfies $\sup_x \sqrt{k(x,x)} < \infty$, provided k is continuous.

Note. By the Kintchine-Kahane inequality, the first inequality above holds in the opposite direction with an additional factor of $\sqrt{2}$.

Now we can derive a uniform deviation bound on balls $B_k(M)$ in the RKHS of kernel k. Recall the two-sided Rademacher complexity bound:

Suppose $\mathcal{G} \subseteq [a, b]^{\mathcal{Z}}$ and $Z_1, ..., Z_n$ are iid. Then $\forall \delta > 0$ w.p. $\ge 1 - \delta$ w.r.t. $(Z_1, \dots, Z_n)$,

$$
\sup_{g \in \mathcal{G}} \left| \mathbb{E} g(Z) - \frac{1}{n} \sum_{i=1}^n g(Z_i) \right| \le 2\mathfrak{R}_n(\mathcal{G}) + (b - a) \sqrt{\frac{\ln(2/\delta)}{2n}}. \tag{3}
$$

We can apply this bound with $\mathcal{G} = L \circ B_k(M)$ and $Z = (X, Y)$. Then $g(Z) = L(Y, f(X))$ for $f \in B_k(M)$, and

$$\mathbb{E}g(Z) = \mathbb{E}L(Y, f(X)) = R_L(f)$$

Similarly, we have

$$\frac{1}{n}\sum g(Z_i) = \frac{1}{n}\sum L(Y_i, f(X_i)) =: \widehat{R}_L(f)$$

This gives us $R_L(f) - \widehat{R}_L(f)$ on the LHS of (3) and $\mathfrak{R}_n(L \circ B_k(M))$ in the first term on the RHS. We can use our previous results to bound the RHS, which leads to the following theorem:

Theorem 2. *Let k be a bounded kernel, $\sup_x \sqrt{k(x,x)} = B < \infty$. Suppose $L(y, \cdot)$ is C-Lipschitz continuous for all $y \in \mathcal{Y}$, and that $L_0 := \sup_{y \in \mathcal{Y}} L(y, 0) < \infty$. Fix $M > 0$ and $\delta > 0$. Then w.p. $\geq 1 - \delta$ w.r.t. the iid sample $((X_1, Y_1), \ldots, (X_n, Y_n))$,*

$$\sup_{f \in B_k(M)} \left| R_L(f) - \widehat{R}_L(f) \right| \leq \frac{2CMB}{\sqrt{n}} + (L_0 + CMB)\sqrt{\frac{\ln(2/\delta)}{2n}}.$$

Proof. The first term on the RHS comes from Corollary 2 and Theorem 1, noting that $\mathfrak{R}_n(B_k(M)) = \mathbb{E}\widehat{\mathfrak{R}}_S(B_k(M)) \leq MB/\sqrt{n}$. The term $b - a$ from (3) results from the observation that for $f \in B_k(M)$,

$$\begin{aligned}
\|f\|_\infty &= \sup_{x \in \mathcal{X}} |f(x)| \\
&= \sup_{x \in \mathcal{X}} |\langle f, k(\cdot, x)\rangle_{\mathcal{F}}| && \text{(reproducing property)} \\
&\leq \sup_{x \in \mathcal{X}} \|f\|_{\mathcal{F}} \|k(\cdot, x)\|_{\mathcal{F}} && \text{(Cauchy-Schwarz)} \\
&\leq MB.
\end{aligned}$$

Then $a := 0 \leq g(Z) = L(Y, f(X)) \leq L_0 + CMB =: b$. □

Note. The assumption $L_0 < \infty$ always holds for classification since $\mathcal{Y}$ is finite. The bound also applies to some regression settings.

Note. Some common loss functions, such as the hinge and logistic losses, satisfy the Lipschitz assumption, but others, such as the exponential and squared error losses, do not.

References

[1] R. Meir and T. Zhang, "Generalization Error Bounds for Bayesian Mixture Algorithms" *Journal of Machine Learning Research*, vol. 4, pp. 839-860, 2003.