Rademacher Averages of Kernel Classes

Lecturer: Peter Bartlett *Scribe: Todd Templeton*

Consider $l(y,\hat{y}) = 1[\text{sign}(\hat{y}) \neq y]$, $\phi(yf(x))$, and $\tilde{\phi}(\alpha) = \min(1, \phi(\alpha))$. Suppose $\phi(\alpha) \geq 1[\alpha \leq 0]$; then

$$L(f) \leq \mathbb{E}\left[\tilde{\phi}(Yf(X))\right] \leq \hat{\mathbb{E}}\left[\tilde{\phi}(Yf(X))\right] + \epsilon \leq \hat{\mathbb{E}}\left[\phi(Yf(X))\right] + \epsilon \ .$$

Theorem 1.1. For $\phi : \mathbb{R} \to \mathbb{R}^+$ such that $\phi(\alpha) \geq 1[\alpha \leq 0]$, with probability at least $\geq 1 - \delta$, we have

$$\begin{aligned}\sup_{f\in\mathcal{F}}\left(L(f) - \hat{\mathbb{E}}\left[\phi(Yf(X))\right]\right) &\leq 2R_n\left(\left\{(x,y) \to \tilde{\phi}(yf(x)) : f \in \mathcal{F}\right\}\right) + \sqrt{\frac{\log 1/\delta}{2n}} \\ &\leq cR_n(\mathcal{F}) + \sqrt{\frac{\log 1/\delta}{2n}} \ ,\end{aligned}$$

provided that $\tilde{\phi}$ is Lipschitz.

Example. Consider $\mathcal{F} = \lambda\text{co}(\mathcal{G})$ with $\mathcal{G} \subseteq \{\pm 1\}^{\mathcal{X}}$ and $d_{VC}(\mathcal{G}\} \leq d$. We have $R_n(\mathcal{F}) = R_n(\lambda\text{co}(\mathcal{G})) = \lambda R_n(\text{co}(\mathcal{G})) = \lambda R_n(\mathcal{G}) \leq c\lambda\sqrt{\frac{d}{n}}$. Therefore, with high probability,

$$\forall f \in \mathcal{F}, L(f) \leq \hat{\mathbb{E}}\left[\phi(Yf(X))\right] + c\lambda\sqrt{\frac{d}{n}} + c\sqrt{\frac{\log 1/\delta}{n}} \ .$$

The second term is a regularization factor; as λ (a measure of the complexity of the class) becomes large, we expect the first term to become small (at the expense of the regularization term becoming large).

Example. Consider $\mathcal{H}$ = kernel class (RKHS). We wish to find $\min_{f\in\mathcal{H}}\left(||f||^2_{\mathcal{H}} + C\hat{\mathbb{E}}\left[\phi(Yf(X))\right]\right)$. Consider $\min_{f\in\mathcal{H}} \hat{\mathbb{E}}\left[\phi(Yf(X))\right]$ s.t. $||f||_{\mathcal{H}} \leq \lambda$, and define $\mathcal{F} = \{f \in \mathcal{H} : ||f||_{\mathcal{H}} \leq \lambda\}$. Recall that

$$R_n(\mathcal{F}) = \mathbb{E}\left[\hat{R}_n(\mathcal{F})\right] = \mathbb{E}\left[\mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i) \,|X_1,\ldots,X_n\right]\right] \ .$$

We assert (and prove below) that $\hat{R}_n(\mathcal{F}) \leq \frac{\lambda}{\sqrt{n}}\sqrt{\frac{\text{trace}(\mathsf{K})}{n}}$ with $\mathsf{K}_{ij} = k(x_i,x_j)$. Note that $\text{trace}(\mathsf{K}) = \sum_{i=1}^n k(x_i,x_i)$, so the sum of square norms in feature space is equal to the sum of the eigenvalues of K. We also assert (and prove below) that

$$R_n(\mathcal{F}) \leq \frac{\lambda}{\sqrt{n}}\sqrt{\sum_{j=1}^{\infty}\mu_j} \ ,$$

where μ_j are eigenvalues of the integral operator

$$T_k : f \to \int_{\mathcal{X}} k(X,\cdot)\, f(X)\, dP(X) \ .$$

Therefore, from above,

$$L(f) \leq \hat{\mathbb{E}}\left[\phi(Yf(X))\right] + c\lambda\sqrt{\frac{R^2}{n}} + \sqrt{\frac{\log 1/\delta}{2n}} \ ,$$

where R is the radius of the ball in feature space containing all the feature vectors.

PROOF. We begin by showing that $\hat{R}_n(\mathcal{F}) \leq \frac{\lambda}{\sqrt{n}}\sqrt{\frac{\mathrm{trace}(\mathsf{K})}{n}}$:

$$\begin{aligned}
\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(X_i) &= \sup_{\|f\|_{\mathcal{H}}\leq\lambda}\left\langle \frac{1}{n}\sum_{i=1}^{n}\epsilon_i k(X_i,\cdot), f\right\rangle \\
&= \|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i k(X_i,\cdot)\| \times \lambda \quad \text{(Cauchy-Schwarz, } \|f\|\leq\lambda) \\
&= \langle \frac{1}{n}\sum_{i=1}^{n}\epsilon_i k(X_i,\cdot), \frac{1}{n}\sum_{j=1}^{n}\epsilon_j k(X_j,\cdot)\rangle^{1/2}\lambda \\
&= \sqrt{\frac{1}{n^2}\sum_{i,j}\epsilon_i\epsilon_j k(X_i,X_j)\lambda} \\
&= \frac{\lambda}{n}\sqrt{\epsilon'\mathsf{K}\epsilon},
\end{aligned}$$

where ϵ is the vector of ϵ_i. Therefore,

$$\begin{aligned}
\hat{R}_n(\mathcal{F}) &= \mathbb{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(X_i)\,|X_1,\ldots,X_n\right] \\
&= \frac{\lambda}{n}\mathbb{E}\left[\sqrt{\epsilon'\mathsf{K}\epsilon}|X_1,\ldots,X_n\right] \\
&\leq \frac{\lambda}{n}\sqrt{\mathbb{E}\left[\epsilon'\mathsf{K}\epsilon|X_1,\ldots,X_n\right]} \quad \text{(Jensen's inequality)} \\
&= \frac{\lambda}{n}\sqrt{\sum_{i=1}^{n}\mathsf{K}_{ii}} \\
&= \frac{\lambda}{\sqrt{n}}\sqrt{\frac{\mathrm{trace}(\mathsf{K})}{n}}.
\end{aligned}$$

We now show that $R_n(\mathcal{F}) \leq \frac{\lambda}{\sqrt{n}}\sqrt{\sum_{j=1}^{\infty}\mu_j}$:

$$\begin{aligned}
R_n(\mathcal{F}) &= \mathbb{E}\left[\hat{R}_n(\mathcal{F})\right] \\
&\leq \lambda\sqrt{\frac{\mathbb{E}\left[k(X_1,X_1)\right]}{n}} \\
&= \lambda\sqrt{\frac{\mathbb{E}\left[\sum_{j=1}^{\infty}\mu_j\psi_j^2(X_1)\right]}{n}} \\
&= \lambda\sqrt{\frac{\mathbb{E}\left[\sum_{j=1}^{\infty}\mu_j\right]}{n}},
\end{aligned}$$

i.e. $R_n(\mathcal{F}) \leq \frac{\lambda}{\sqrt{n}}\sqrt{\mathrm{trace}(T_k)}$, where the final step above uses the fact that $k(x,y) = \sum_{j=1}^{\infty}\mu_j\psi_j(x)\psi_j(y)$. □