Machine Learning Theory (CS 6783)

Lecture 7: Massart's Lemma, Rademacher Complexity Properties

1 Recap

1. For any statistical learning problem we have,

$$\mathbb{E}_S\left[L_D(\hat{y}_{\text{erm}}) - \inf_{f\in\mathcal{F}} L_D(f)\right] \le \frac{2}{n}\mathbb{E}_S\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^n \epsilon_t \ell(f(x_t), y_t)\right]$$

2. For binary classification loss we saw that

$$\frac{2}{n}\mathbb{E}_S\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^n \epsilon_t \ell(f(x_t), y_t)\right] = \frac{1}{n}\mathbb{E}_S\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^n \epsilon_t f(x_t)\right]$$

3. Effective size of function class on data

$$\mathbb{E}_S\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^n \epsilon_t \ell(f(x_t), y_t)\right] = \mathbb{E}_S\mathbb{E}_\epsilon\left[\sup_{\mathbf{f}\in\mathcal{F}_{|x_1,\ldots,x_n}}\frac{1}{n}\sum_{t=1}^n \epsilon_t \ell(\mathbf{f}[t], y_t)\right] \le \mathbb{E}_S\left[\sqrt{\frac{\log|\mathcal{F}_{|x_1,\ldots,x_n}|}{n}}\right]$$

where $\mathcal{F}_{|x_1,\ldots,x_n} = \{f(x_1),\ldots,f(x_n) : f\in\mathcal{F}\}$.

4. By VC Lemma, $\left|\mathcal{F}_{|x_1,\ldots,x_n}\right| \le \sum_{i=0}^{VC(\mathcal{F})}\binom{n}{i} \le n^{VC(\mathcal{F})}$

5. If $\mathrm{VC}(\mathcal{F}) = \infty$, then $\mathcal{V}_n^{PAC}(\mathcal{F}) \ge 1/4$

2 Massart's Finite Lemma

Lemma 1. *For any set $V \subset \mathbb{R}^n$:*

$$\frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{\mathbf{v}\in V}\sum_{t=1}^n \epsilon_t \mathbf{v}[t]\right] \le \frac{1}{n}\sqrt{2\left(\sup_{\mathbf{v}\in V}\sum_{t=1}^n \mathbf{v}^2[t]\right)\log|V|}$$

Proof.

$$\begin{aligned}
\sup_{\mathbf{v}\in V}\sum_{t=1}^n \epsilon_t\mathbf{v}[t] &= \frac{1}{\lambda}\log\left(\sup_{\mathbf{v}\in V}\exp\left(\lambda\sum_{t=1}^n \epsilon_t\mathbf{v}[t]\right)\right)\\
&\le \frac{1}{\lambda}\log\left(\sum_{\mathbf{v}\in V}\exp\left(\lambda\sum_{t=1}^n \epsilon_t\mathbf{v}[t]\right)\right)\\
&= \log\left(\sum_{\mathbf{v}\in V}\prod_{t=1}^n \exp\left(\lambda\epsilon_t\mathbf{v}[t]\right)\right)
\end{aligned}$$

Taking expectation w.r.t. Rademacher random variables,

$$\begin{aligned}
\mathbb{E}_\epsilon\left[\sup_{\mathbf{v}\in V}\sum_{t=1}^n \epsilon_t \mathbf{v}[t]\right] &\le \frac{1}{\lambda}\mathbb{E}_\epsilon\left[\log\left(\sum_{\mathbf{v}\in V}\prod_{t=1}^n \exp\left(\lambda\epsilon_t\mathbf{v}[t]\right)\right)\right] \\
&\le \frac{1}{\lambda}\log\left(\sum_{\mathbf{v}\in V}\prod_{t=1}^n \mathbb{E}_{\epsilon_t}\left[\exp\left(\lambda\epsilon_t\mathbf{v}[t]\right)\right]\right) \\
&= \frac{1}{\lambda}\log\left(\sum_{\mathbf{v}\in V}\prod_{t=1}^n \frac{e^{\lambda\mathbf{v}[t]}+e^{-\lambda\mathbf{v}[t]}}{2}\right)
\end{aligned}$$

For any x, $\frac{e^x+e^{-x}}{2} \le e^{x^2/2}$

$$\begin{aligned}
&\le \frac{1}{\lambda}\log\left(\sum_{\mathbf{v}\in V} e^{\lambda^2\sum_{t=1}^n \mathbf{v}^2[t]/2}\right) \\
&\le \frac{1}{\lambda}\log\left(|V|e^{\lambda^2\sup_{\mathbf{v}\in V}\left(\sum_{t=1}^n \mathbf{v}^2[t]\right)/2}\right) \\
&= \frac{\log|V|}{\lambda} + \frac{\lambda\sup_{\mathbf{v}\in V}\left(\sum_{t=1}^n \mathbf{v}^2[t]\right)}{2}
\end{aligned}$$

Choosing $\lambda = \sqrt{\frac{2\log|V|}{\sup_{\mathbf{v}\in V}\left(\sum_{t=1}^n \mathbf{v}^2[t]\right)}}$ completes the proof. $\square$

3 The Magic of Rademacher Complexity

Define empirical Rademacher complexity of a class $\mathcal{G}$, a set of functions on $\mathcal{Z}$, on a sample $S = \{z_1, \ldots, z_n\}$ as

$$\hat{\mathcal{R}}_S(\mathcal{G}) := \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{g\in\mathcal{G}}\sum_{t=1}^n \epsilon_t g(z_t)\right]$$

In class we showed that

$$\mathbb{E}_S\left[L_D(\hat{y}_{erm}) - \inf_{f\in\mathcal{F}} L_D(f)\right] \le 2\ \mathbb{E}_S\left[\hat{\mathcal{R}}_S(\ell\circ\mathcal{F})\right]$$

where $\ell\circ\mathcal{F} = \{(x,y)\mapsto \ell(f(x),y) : f\in\mathcal{F}\}$

Proposition 2. *For any sample $S = \{z_1, \ldots, z_n\}$ and any classes $\mathcal{G}$, $\mathcal{H}$ mapping instances in $\mathcal{Z}$ to reals :*

1. *If $\mathcal{H}\subset\mathcal{G}$, then $\hat{\mathcal{R}}_S(\mathcal{H}) \le \hat{\mathcal{R}}_S(\mathcal{G})$*
2. *For any fixed function $h : \mathcal{Z}\mapsto\mathbb{R}$, $\hat{\mathcal{R}}_S(\mathcal{G}+h) = \hat{\mathcal{R}}_S(\mathcal{G})$*
3. *$\hat{\mathcal{R}}_S(\mathrm{cvx}(\mathcal{G})) = \hat{\mathcal{R}}_S(\mathcal{G})$*
4. *$\hat{\mathcal{R}}_S(\mathcal{G}+\mathcal{H}) = \hat{\mathcal{R}}_S(\mathcal{G}) + \hat{\mathcal{R}}_S(\mathcal{H})$*

Proof.

1. $\hat{\mathcal{R}}_S(\mathcal{H}) = \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{g\in\mathcal{H}}\sum_{t=1}^n \epsilon_t g(z_t)\right] \le \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{g\in\mathcal{G}}\sum_{t=1}^n \epsilon_t g(z_t)\right] \le \hat{\mathcal{R}}_S(\mathcal{G})$.

2.
$$\begin{aligned}
\hat{\mathcal{R}}_S(\mathcal{G}+h) &= \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{g\in\mathcal{G}}\sum_{t=1}^n \epsilon_t(g(z_t)+h(z_t))\right] \\
&= \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{g\in\mathcal{G}}\left\{\sum_{t=1}^n \epsilon_t g(z_t)\right\} + \sum_{t=1}^n \epsilon_t h(z_t))\right] \\
&= \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{g\in\mathcal{G}}\sum_{t=1}^n \epsilon_t g(z_t)\right] + 0 = \hat{\mathcal{R}}_S(\mathcal{G})
\end{aligned}$$

3. $\mathrm{cvx}(\mathcal{G}) = \{\mathbf{z} \mapsto \mathbb{E}_{g\sim\pi}[g(z)] : \pi \in \Delta(\mathcal{G})\}$
$$\begin{aligned}
\hat{\mathcal{R}}_S(\mathrm{cvx}(\mathcal{G})) &= \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{\pi\in\Delta(\mathcal{G})}\sum_{t=1}^n \epsilon_t \mathbb{E}_{g\in\pi}[g(z_t)]\right] \\
&= \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{\pi\in\Delta(\mathcal{G})}\mathbb{E}_{g\in\pi}\left[\sum_{t=1}^n \epsilon_t g(z_t)\right]\right] \\
&= \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{g\in\mathcal{G}}\sum_{t=1}^n \epsilon_t g(z_t)\right] \\
&= \hat{\mathcal{R}}_S(\mathcal{G})
\end{aligned}$$

4.
$$\begin{aligned}
\hat{\mathcal{R}}_S(\mathcal{G}+\mathcal{H}) &= \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{g\in\mathcal{G},h\in\mathcal{H}}\sum_{t=1}^n \epsilon_t\left(g(z_t)+h(z_t)\right)\right] \\
&= \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{g\in\mathcal{G}}\sum_{t=1}^n \epsilon_t g(z_t) + \sup_{h\in\mathcal{H}}\sum_{t=1}^n \epsilon_t h(z_t)\right] \\
&= \hat{\mathcal{R}}_S(\mathcal{G}) + \hat{\mathcal{R}}_S(\mathcal{H})
\end{aligned}$$

$\square$

Lemma 3. *For any $\phi_1, \ldots, \phi_n$ where each $\phi_i : \mathbb{R} \mapsto \mathbb{R}$ and is L-Lipschitz, and any $z_1, \ldots, z_n$, we have,*
$$\frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{g\in\mathcal{G}}\sum_{t=1}^n \epsilon_t \phi_t\left(g(z_t)\right)\right] \le \frac{L}{n}\mathbb{E}_\epsilon\left[\sup_{g\in\mathcal{G}}\sum_{t=1}^n \epsilon_t g(z_t)\right]$$

Remark : For any Lipschitz loss we can get rid of loss and only have Rademacher complexity of the class of predictors. That is $\hat{\mathcal{R}}_S(\ell\circ\mathcal{F}) \le L\hat{\mathcal{R}}_S(\mathcal{F})$

Proof.

$$
\begin{aligned}
&\frac{1}{n}\mathbb{E}_{\epsilon_{1:n}}\left[\sup_{g\in\mathcal{G}}\sum_{t=1}^{n}\epsilon_t\phi_t\left(g(z_t)\right)\right] \\
&= \mathbb{E}_{\epsilon_{1:n-1}}\frac{\sup_{g\in\mathcal{G}}\left\{\sum_{t=1}^{n-1}\epsilon_t\phi_t\left(g(z_t)\right)+\phi_n(g(z_n))\right\}+\sup_{g\in\mathcal{G}}\left\{\sum_{t=1}^{n-1}\epsilon_t\phi_t\left(g(z_t)\right)-\phi_n(g(z_n))\right\}}{2} \\
&= \mathbb{E}_{\epsilon_{1:n-1}}\left[\frac{\sup_{g,g'\in\mathcal{G}}\left\{\sum_{t=1}^{n-1}\epsilon_t\left(\phi_t\left(g(z_t)\right)+\phi_t\left(g'(z_t)\right)\right)+\phi_n(g(z_n))-\phi_n(g'(z_n))\right\}}{2}\right] \\
&\leq \mathbb{E}_{\epsilon_{1:n-1}}\left[\frac{\sup_{g,g'\in\mathcal{G}}\left\{\sum_{t=1}^{n-1}\epsilon_t\left(\phi_t\left(g(z_t)\right)+\phi_t\left(g'(z_t)\right)\right)+L|g(z_n))-g'(z_n)|\right\}}{2}\right] \\
&= \mathbb{E}_{\epsilon_{1:n-1}}\left[\frac{\sup_{g,g'\in\mathcal{G}}\left\{\sum_{t=1}^{n-1}\epsilon_t\left(\phi_t\left(g(z_t)\right)+\phi_t\left(g'(z_t)\right)\right)+L(g(z_n))-g'(z_n))\right\}}{2}\right] \\
&= \mathbb{E}_{\epsilon_{1:n-1}}\frac{\sup_{g\in\mathcal{G}}\left\{\sum_{t=1}^{n-1}\epsilon_t\phi_t\left(g(z_t)\right)+Lg(z_n)\right\}+\sup_{g\in\mathcal{G}}\left\{\sum_{t=1}^{n-1}\epsilon_t\phi_t\left(g(z_t)\right)-Lg(z_n)\right\}}{2} \\
&= \frac{1}{n}\mathbb{E}_{\epsilon_{1:n}}\left[\sup_{g\in\mathcal{G}}\sum_{t=1}^{n-1}\epsilon_t\phi_t\left(g(z_t)\right)+L\epsilon_n g(z_n)\right]
\end{aligned}
$$

Repeating the above argument we remove $\phi_1,\ldots,\phi_{n-1}$ and so, we conclude that

$$
\frac{1}{n}\mathbb{E}_{\epsilon}\left[\sup_{g\in\mathcal{G}}\sum_{t=1}^{n}\epsilon_t\phi_t\left(g(z_t)\right)\right] \leq \frac{L}{n}\mathbb{E}_{\epsilon}\left[\sup_{g\in\mathcal{G}}\sum_{t=1}^{n}\epsilon_t g(z_t)\right]
$$

□

4 Example : Rademacher complexity of linear function classes

1. L1 regularizer : Let $\mathcal{F}=\{x\mapsto f^\top x : f\in\mathbb{R}^d, \|f\|_1\leq R\}$. In this case we have

$$
\begin{aligned}
\hat{\mathcal{R}}_S(\mathcal{F}) &= \frac{1}{n}\mathbb{E}_{\epsilon}\left[\sup_{f:\|f\|_1\leq R} f^\top\left(\sum_{t=1}^{n}\epsilon_t x_t\right)\right] \\
&= \frac{R}{n}\mathbb{E}_{\epsilon}\left[\sup_{f:\|f\|_1\leq 1} f^\top\left(\sum_{t=1}^{n}\epsilon_t x_t\right)\right]
\end{aligned}
$$

Now note that the unit ℓ_1 ball on $\mathbb{R}^d$ can be written as a convex hull of $2d$ points, $\{e_1, -e_1, e_2, -e_2, \ldots, e_d, -e_d\}$. Hence by Proposition 2 (4) we have that

$$\begin{aligned}\hat{\mathcal{R}}_S(\mathcal{F}) &= R\ \hat{\mathcal{R}}_S(\{e_1, -e_1, e_2, -e_2, \ldots, e_d, -e_d\}) \\ &\leq \frac{R \log d\ \sqrt{\max_{i\in[d]} \sum_{t=1}^n |x_t[i]|^2}}{n} \\ &\leq \frac{R \sup_{x\in\mathcal{X}} \|x\|_\infty \sqrt{\log d}}{\sqrt{n}}\end{aligned}$$

2. Hilbert norm regularizer : Let $\mathcal{F} = \{x \mapsto \langle f, x\rangle : \|f\|_2 \leq R\}$. This example we already saw in Lecture 4 (we used $R = 1$ there but the analysis carries). For this case we have that,

$$\hat{\mathcal{R}}_S(\mathcal{F}) \leq \frac{R}{n}\sqrt{\sum_{t=1}^n \|x_t\|_2^2} \leq \frac{R \sup_{x\in\mathcal{X}} \|x\|_2}{\sqrt{n}}$$

5 Applications

Example applications : Lasso, SVM, ridge regression, Logistic Regression (including kernel methods), ℓ_1 neural networks, matrix completion (max norm, trace norm), graph prediction

Observation : Hinge loss given by $\ell(y', y) = \max\{1 - y'y, 0\}$ is 1-Lipschitz. Logistic loss given by $\ell(y', y) = \log(1 + e^{-y'y})$ is 1-Lipchitz. Squared loss $\ell(y', y) = (y' - y)^2$ is $4B$ Lipschtiz when $|y|, |y'| \leq B$. Absolute loss $\ell(y', y) = |y - y'|$ is 1-Lipchitz. In all these cases using Lemma 3 we have,

$$\mathbb{E}_S\left[L_D(\hat{y}_{\text{erm}}) - \inf_{f\in\mathcal{F}} L_D(f)\right] \leq 2L\ \mathbb{E}_S\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}} \sum_{t=1}^n \epsilon_t f(x_t)\right]$$

where L is the corresponding Lipchitz constant of the loss.

1. SVM :

$$\begin{aligned}&\text{minimize} \sum_{t=1}^n \max\{1 - \langle f, x_t\rangle \cdot y_t, 0\} \\ &\text{subject to } \|f\|_2 \leq R\end{aligned}$$

This corresponds to class F given by linear predictors with Hilbert norm constrained by R

2. Lasso :

$$\begin{aligned}&\text{minimize} \sum_{t=1}^n (y - \langle f, x_t\rangle)^2 \\ &\text{subject to } \|f\|_1 \leq R\end{aligned}$$

Corresponds to linear predictor with ℓ_1 norm constrained by 1