

Machine Learning Theory (CS 6783)

Lecture 8: Rademacher Complexity Properties

1 Recap

1. For any statistical learning problem we have,

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

2. Effective size of function class on data

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

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

3. Massart's finite Lemma, for any set $V \subset \mathbb{R}^n$:

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

4. Rademacher Complexity:

- (a) If $\mathcal{H} \subset \mathcal{G}$, then $\hat{\mathcal{R}}_S(\mathcal{H}) \leq \hat{\mathcal{R}}_S(\mathcal{G})$
- (b) For any fixed function $h : \mathcal{Z} \mapsto \mathbb{R}$, $\hat{\mathcal{R}}_S(\mathcal{G} + h) = \hat{\mathcal{R}}_S(\mathcal{G})$
- (c) $\hat{\mathcal{R}}_S(\text{cvx}(\mathcal{G})) = \hat{\mathcal{R}}_S(\mathcal{G})$
- (d) $\hat{\mathcal{R}}_S(\mathcal{G} + \mathcal{H}) = \hat{\mathcal{R}}_S(\mathcal{G}) + \hat{\mathcal{R}}_S(\mathcal{H})$

2 Contraction Lemma

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

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

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

Lemma 1. For any ϕ_1, \dots, ϕ_n where each $\phi_i : \mathbb{R} \mapsto \mathbb{R}$ and is L -Lipschitz, and any z_1, \dots, z_n , we have,

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

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

Proof.

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

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

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

□

3 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, $\{e_1, -e_1, e_2, -e_2, \dots, e_d, -e_d\}$. Hence by Proposition ?? (4) we have that

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

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,

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

4 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 $4B$ Lipschitz when $|y|, |y'| \leq B$. Absolute loss $\ell(y', y) = |y - y'|$ is 1-Lipchitz. In all these cases using Lemma 1 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 Lipschitz 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

3. ℓ_1 neural network with K layers. Loss could be squared loss or logistic loss. Let \mathcal{F}_1 be some arbitrary base class of predictors. Recursively define the subsequent i layer neural network predictor as follows

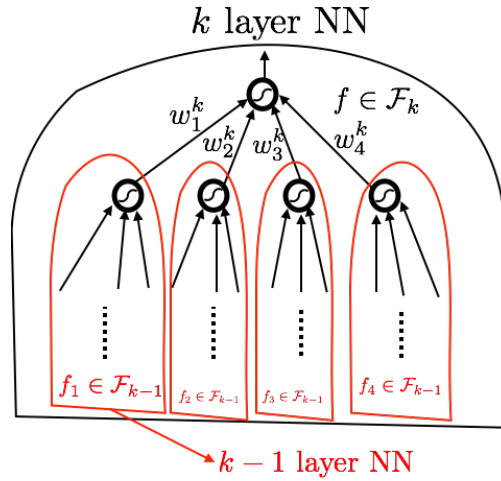
$$\mathcal{F}_i = \left\{ x \mapsto \sum_j w_j^i \sigma(f_j(x)) : \forall j, f_j \in \mathcal{F}_{i-1}, \|w^i\|_1 \leq B_i \right\}$$

where σ is a 1-Lipchitz loss function. Then

$$\begin{aligned} \hat{\mathcal{R}}_S(\mathcal{F}_i) &= \frac{1}{n} \mathbb{E}_\epsilon \left[\sup_{\substack{\|w^i\|_1 \leq B_i \\ \forall j, f_j \in \mathcal{F}_{i-1}}} \sum_{t=1}^n \sum_j \epsilon_t w_j^i \sigma(f_j(x_t)) \right] \\ &\leq \frac{1}{n} \mathbb{E}_\epsilon \left[\sup_{\substack{\|w^i\|_1 \leq B_i \\ \forall j, f_j \in \mathcal{F}_{i-1}}} \|w^i\|_1 \max_j \left| \sum_{t=1}^n \epsilon_t \sigma(f_j(x_t)) \right| \right] \\ &\leq \frac{B_i}{n} \mathbb{E}_\epsilon \left[\sup_{\forall j, f_j \in \mathcal{F}_{i-1}} \max_j \left| \sum_{t=1}^n \epsilon_t \sigma(f_j(x_t)) \right| \right] \\ &= \frac{B_i}{n} \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}_{i-1}} \left| \sum_{t=1}^n \epsilon_t \sigma(f(x_t)) \right| \right] \\ &\leq \frac{2B_i}{n} \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}_{i-1}} \sum_{t=1}^n \epsilon_t \sigma(f(x_t)) \right] \\ &= 2B_i \hat{\mathcal{R}}_S(\sigma \circ \mathcal{F}_{i-1}) \\ &\leq 2B_i \hat{\mathcal{R}}_S(\mathcal{F}_{i-1}) \end{aligned}$$

Hence we can conclude that

$$\hat{\mathcal{R}}_S(\mathcal{F}_i) \leq \left(\prod_{i=1}^k 2B_i \right) \hat{\mathcal{R}}_S(\mathcal{F}_1)$$



5 Covering Number

Conditioned on x_1, \dots, x_n , we are interested in bounding :

$$\frac{1}{n} \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^n \epsilon_t f(x_t) \right] = \frac{1}{n} \mathbb{E}_\epsilon \left[\sup_{\mathbf{v} \in \mathcal{F}_{|x_1, \dots, x_n}} \sum_{t=1}^n \epsilon_t \mathbf{v}[t] \right]$$

Recall the projection of \mathcal{F} on sample :

$$\mathcal{F}_{|x_1, \dots, x_n} = \{(f(x_1), \dots, f(x_n)) \in \mathbb{R}^d : f \in \mathcal{F}\}$$

For real valued functions of course $|\mathcal{F}_{|x_1, \dots, x_n}|$ could very well be infinite. But now given the n data points, we can ask how large a set do we need to discretize $\mathcal{F}_{|x_1, \dots, x_n}$ to accuracy β .

Definition 1. $V \subset \mathbb{R}^n$ is an ℓ_p cover of \mathcal{F} on x_1, \dots, x_n at scale $\beta > 0$ if for all $f \in \mathcal{F}$, there exists $\mathbf{v}_f \in V$ such that

$$\left(\frac{1}{n} \sum_{t=1}^n |f(x_t) - \mathbf{v}_f[t]|^p \right)^{1/p} \leq \beta$$

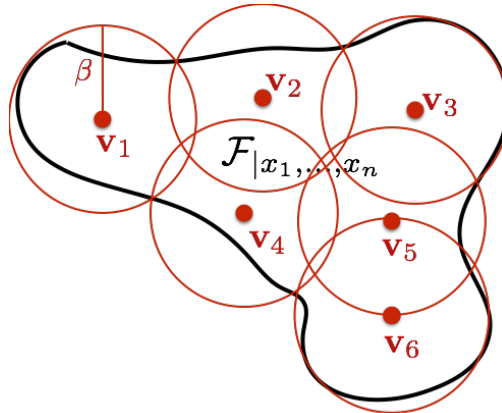
Empirical covering number

$$\mathcal{N}_p(\mathcal{F}, \beta; x_1, \dots, x_n) = \min\{|V| : V \text{ is an } \ell_p \text{ cover of } \mathcal{F} \text{ on } x_1, \dots, x_n \text{ at scale } \beta\}$$

Covering number

$$\mathcal{N}_p(\mathcal{F}, \beta, n) = \sup_{x_1, \dots, x_n} \mathcal{N}_p(\mathcal{F}, \beta; x_1, \dots, x_n)$$

You can think of $V \subset \mathbb{R}^n$ as a finite discretization of $\mathcal{F}_{|x_1, \dots, x_n} \subset \mathbb{R}^n$ to scale β in the normalized ℓ_p distance as shown in Figure below. It can easily be verified that for any $p, p' \in [1, \infty)$ such that $p' \leq p$, $\mathcal{N}_{p'}(\mathcal{F}, \beta; x_1, \dots, x_n) \leq \mathcal{N}_p(\mathcal{F}, \beta; x_1, \dots, x_n)$.



$$V = \{\mathbf{v}_1, \dots, \mathbf{v}_6\}$$