Lecture 21

Last time

- Ridge Regression
- Random design
- Beyond linear features

Today

- Radamacher complexity
- Polynomial discrimination

Radamacher Complexity

We have been studying the relationship between

$$\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \ell(\theta, z_i) \quad \text{and} \quad \min_{\theta \in \Theta} \mathbb{E}_z \, \ell(\theta, z).$$

In what follows we will try to understand how to bound the uniform distance between these two.

Goal: Consider $x_i \sim \mathbb{P}$ taking values on some set $\mathcal{X}$ and let $\mathcal{F}$ be a family of functions of the form $f : \mathcal{X} \to \mathbb{R}$. Our aim is to bound

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(x_i) - \mathbb{E} f(x) \right|.$$

Note that unlike before we will not focus on a specific $\mathcal{F}$.

The following measure of complexity will be instrumental.

Def: The **Radamacher Complexity** of a set $A \subseteq \mathbb{R}^n$ is given by

$$\mathcal{R}(A) := \mathbb{E}_{\varepsilon} \sup_{a \in A} |\langle a, \varepsilon \rangle|,$$

Some people define it without $|\cdot|$.

with $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)$ with ε_i iid Radamacher random variables. $\dashv$

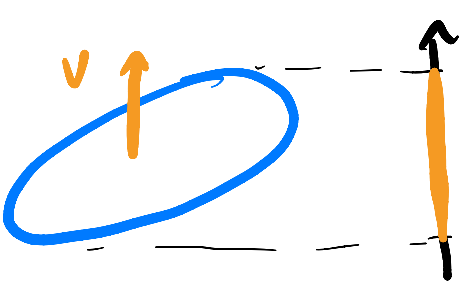
A **geometric interpretation**: the support function of a set A is

$$\sigma_A(v) = \sup_{a \in A} \langle v, a \rangle$$

Notice that if v has norm 1, then

Why?

$$\text{"Width of } A \text{ in the } v \text{ direction"} = \sigma_A(v) + \sigma_C(-v)$$



Moreover

$$\mathbb{E}\, \sigma_A(\varepsilon) \leq \mathcal{R}_A \leq 2\, \mathbb{E}\, \sigma_A(\varepsilon).$$

$\dashv$

Define

$$\mathcal{F}(X) = \{(f(X_1), \dots, f(X_n)) : f \in \mathcal{F}\}.$$

Def: Given a distribution $\mathbb{P}$. The Radamacher complexity of $\mathcal{F}$ is

$$\mathcal{R}_n(\mathcal{F}) := \mathbb{E}_X \, \mathcal{R}\left(\tfrac{1}{n}\mathcal{F}(X)\right)$$

$$= \mathbb{E}_{X,\varepsilon}\left[\sup \left|\tfrac{1}{n}\sum \varepsilon_i f(X_i)\right|\right].$$

Theorem: For any $n \geq 1$,

$$\mathbb{E}\,\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \leq 2\,\mathcal{R}_n(\mathcal{F}).$$

A lower bound also holds (Wainwright Prop 4.11)

Proof: Let $y_1, \dots, y_n \overset{iid}{\sim} \mathbb{P}$ and independent of X. Then,

$$\mathbb{E}\,\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$$

$$= \mathbb{E}_X\left[\sup_{f \in \mathcal{F}} \left|\tfrac{1}{n}\sum (f(X_i) - \mathbb{E}_{y_i} f(y_i))\right|\right]$$

$$= \mathbb{E}_X \left[\sup_{f\in\mathcal{F}} \left| \mathbb{E}_y \frac{1}{n} \sum (f(x_i) - f(y_i)) \right| \right]$$

Jensen's

$$\leq \mathbb{E}_{X,Y} \left[\sup_{f\in\mathcal{F}} \left| \frac{1}{n} \sum (f(x_i) - f(y_i)) \right| \right]$$

symmetric r.v.

$$= \mathbb{E}_{X,Y,\varepsilon} \left[\sup_{f\in\mathcal{F}} \left| \frac{1}{n} \sum \varepsilon_i (f(x_i) - f(y_i)) \right| \right]$$

$$\leq 2 \mathbb{E}_{X,\varepsilon} \left[\sup \left| \frac{1}{n} \sum \varepsilon_i f(x_i) \right| \right]$$

$$\leq 2 \mathcal{R}_n(\mathcal{F}). \qquad \square$$

We also obtain high probability bounds when $\mathcal{F}$ is bounded.

Theorem: Suppose $\mathcal{F}$ satisfies

$$\|f\|_\infty = \sup_{x\in\mathcal{X}} |f(x)| \leq b \qquad \forall f \in \mathcal{F}.$$

Then, for any $n \geq 1$, $t \geq 0$,

$$\mathbb{P}\left(\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \leq 2\mathcal{R}_n(\mathcal{F}) + t \right) \geq 1 - e^{-\frac{nt^2}{2b^2}}.$$

Proof: This is a consequence of McDiarmid's inequality. It suffices to show the bounded difference property with bound $\frac{2b}{n}$.

Set $\bar{f}(x) = f(x) - \mathbb{E}f(X)$. Then

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \bar{f}(x_i) \right|.$$

Let x' differ from x only in its ith component. Fix any $\bar{g} \in \mathcal{F}$, then

$$\left| \frac{1}{n} \sum \bar{g}(x_i) \right| - \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum \bar{f}(x_i') \right|$$

$$\leq \left| \frac{1}{n} \sum \bar{g}(x_i) \right| - \left| \frac{1}{n} \sum \bar{g}(x_i') \right|$$

$$\leq \left| \frac{1}{n} \left(\bar{g}(x_i) - g(x_i') \right) \right|$$

$$\leq 2b/n,$$

take $\sup_{g \in \mathcal{F}}$ and swap the role of x and x' to establish the bound. $\square$

Classes with polynomial discrimination

We transform our problem into bounding Radamacher complexities. Next we develop tools to do so.

Def: $\mathcal{F}$ has polynomial discrimination of order $\nu \geq 1$ if $\forall n \geq 1$ and $\forall x_1, \dots, x_n \in \mathcal{X}$,

$$\# \mathcal{F}(x_1, \dots, x_n) \leq (n+1)^\nu.$$

$\dashv$

Note that even in the simple case where $\forall f \in \mathcal{F}$ $\operatorname{Img} f = \{\pm 1\}$ we can have

$$\# \mathcal{F}(x_1, \dots, x_n) = 2^n.$$

Proposition: Suppose $\mathcal{F}$ has polynomial discrimination of order ν. Then

$$\mathcal{R}_n(\mathcal{F}) \leq \mathbb{E}\left[\sup_{f\in\mathcal{F}} \sqrt{\frac{1}{n}\sum f(x_i)^2}\right]\sqrt{\frac{2\nu\log(n+1)}{n}}$$

Proof: The proof relies on the following lemma

Lemma: Suppose $A \subseteq \mathbb{R}^n$

$$R(A) \leq \max_{a\in A} \|a\|_2 \sqrt{2\log(\# A)}$$

Remember our bound for max of sub-Gaussians.

Recall that

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}\, \mathcal{R}\left(\tfrac{1}{n}\mathcal{F}(X)\right)$$

$$\leq \frac{1}{n}\mathbb{E}_X \max_{f\in\mathcal{F}} \left\| \begin{pmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix} \right\|_2 \sqrt{2\log \mathcal{F}(X)}.$$

The final bound follows since $\mathcal{F}$ has polynomial discrimination. $\square$

In particular if $\mathcal{F}$ is b-bounded then

$$R_n(\mathcal{F}) \leq b\sqrt{\frac{2\nu \log(n+1)}{n}}.$$

Example: Consider $\mathcal{F}$ to be $\{0,1\}$-valued indicator functions of half-intervals

$$f_t(x) = \begin{cases} 1 & \text{if } x \leq t, \\ 0 & \text{otherwise.} \end{cases}$$

Recall the order stats: $x_{(1)} \leq \dots \leq x_{(n)}$,

$$\big(f_t(x_{(1)}), \dots, f_t(x_{(n)})\big) = (1, \dots, 1, 0, \dots, 0).$$

First $x_{(i)} > t$

Thus, $\# \mathcal{F}(X) \leq n+1$. $\dashv$

Corollary (Glivenko-Cantelli): Let

$$F(t) = \mathbb{P}[X \leq t]$$

and $\hat{F}_n$ be the empirical CDF using n iid samples $X_i \sim \mathbb{P}$.
Then, for all $\delta \geq 0$,

$$\mathbb{P}\left[\|\hat{F}_n - F\|_\infty \geq 4\sqrt{\frac{\log(n+1)}{n}} + \delta\right] \leq e^{-n\delta^2/2}.$$

$\dashv$