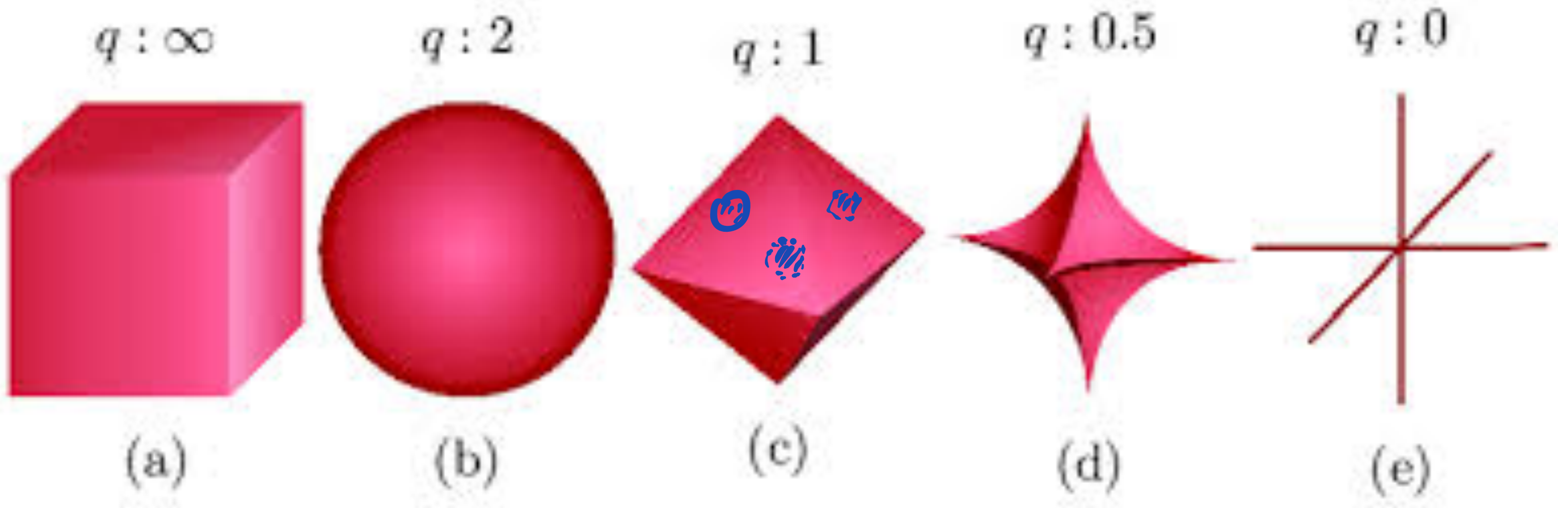
Lecture 11 Localized Complexity

IEMS 402 Statistical Learning

Northwestern

Empirical Method of Maurey

L1 Ball



Volume Based Bound

Last lecture, we discussed the problem of getting a covering number N for L_1 balls using L_2 balls.

$$N(\epsilon, B_1^d, ||\cdot||_2) \tag{1}$$

Using a volume argument, we were able to establish the following result.

$$N(\epsilon, B_1^d, ||\cdot||_2) \leq N(\epsilon, B_1^d, ||\cdot||_1) \tag{2}$$

$$N(\epsilon, B_1^d, ||\cdot||_2) \leq (1 + \frac{2}{\epsilon})^d \tag{3}$$

$$\Rightarrow \quad \log N \leq d \log \frac{1}{\epsilon}$$

Empirical Method of Maurey

Theorem 1. *When* $\epsilon > \frac{1}{\sqrt{d}}$, $N \leq (2d+1)^{O(1/\epsilon^2)}$

As a result, $\log N \lesssim \frac{1}{\epsilon^2}\log(d)$.

number of (random) basis that you need

$\leq d^{1/\epsilon^2}$

Proof. Let's cover the following set:

$$B_1^{d,+} = \{x \in \mathcal{R}^d \mid \|x\|_1 \leq 1 \text{ and } x_i \geq 0 \ \forall i\}$$

L_1 norm

number of possible sum of sample + data is $\binom{t}{d}$

randomly sample t basis. $\Rightarrow$ error: $\epsilon = O\left(\frac{1}{\sqrt{t}}\right)$

The above set means that $\sum x_i \leq 1 \ \forall x_i \geq 0$.

We can think about a probability distribution over $\{e_1, \ldots, e_d, 0\}$:

$$z = \sum_{i=1}^{d} x_i e_i + (1 - \|x\|_1) \cdot 0$$

$\Rightarrow$ as sum of $\begin{pmatrix}1\\0\\\vdots\\0\end{pmatrix} \begin{pmatrix}0\\1\\0\\\vdots\\0\end{pmatrix} \cdots \begin{pmatrix}0\\0\\\vdots\\1\end{pmatrix}, \vec{0}$

Empirical Method of Maurey

This implies the following probabilities.

$$\mathbb{P}[z = e_j] = x_j \, \forall j \in [d]$$
$$\mathbb{P}[z = 0] = 1 - \|x\|_1$$

With these, we can get a mean of the probability distribution.

$$\mathbb{E}[z] = \sum \mathbb{P}[z = e_j] \cdot e_j + \mathbb{P}[z = 0] \cdot 0 = \sum x_j \cdot e_j = x$$

We will draw t samples $z_1, \ldots, z_t$ from the distribution where each z is some e_i. After drawing the samples, we can take the average of the samples:

$$\bar{z} = \frac{1}{t} \sum_{i=1}^{t} z_i$$

We want to show that $\mathbb{E}[\|\bar{z} - x\|_2^2] \leq \epsilon^2$. If we can do this, then if we take all possible $\bar{z}$, we get an ϵ-cover of the space using those $\bar{z}$ since then all x we can choose will be within ϵ of some point in the cover by what we argue above.

Empirical Method of Maurey vs Volumn

- Volumn : $N = \left(\frac{1}{\varepsilon}\right)^d \Rightarrow \log N = d \cdot \log(1/\varepsilon)$

This is better when ε is small

↳ the optimal dependency over ε

- Empirical method of Maurey : $N = d^{\left(\frac{1}{\varepsilon^2}\right)}$

$\Rightarrow \log N = \frac{1}{\varepsilon^2} \log(d)$

The dimension dependency $\log(d)$ is better.

Lasso

The dependency on ε is bad

Localized Complexity

Example: Mean Estimation

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i = \arg\min_{\theta} \underbrace{\frac{1}{n}\sum_{i=1}^{n} (\theta - x_i)^2}_{\hat{L}(\theta)}$$

$x_i \sim \mathrm{Unif}([-1,1])$

$$L(\theta) = \mathbb{E}(\theta - x)^2 = \|\theta - \theta^*\|^2 + \mathrm{Var}(x)$$

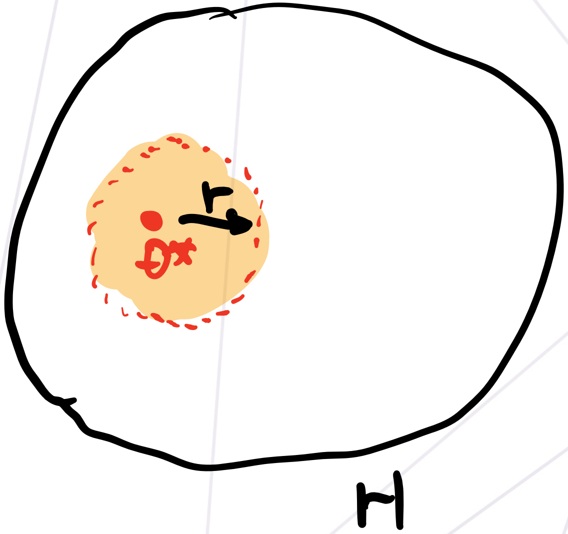
① $L(\hat{\theta}) - L(\theta) \leq \mathrm{Rad} = O\left(\frac{\text{Complexity}}{\sqrt{n}}\right)$

② Right Bound

$$\|\hat{\theta} - \theta^*\| = O\left(\frac{1}{\sqrt{n}}\right)$$

$$\Rightarrow L(\hat{\theta}) - L(\theta^*) = \|\hat{\theta} - \theta^*\|^2 = O\left(\frac{1}{n}\right) \Leftarrow \text{Fast Rate}$$

Idea:Localized Complexity



$$\phi(r) = \mathrm{Rad}\left(\{\theta \in H,\ \|\theta - \hat\theta\| \le r\}\right)$$

$$= \mathbb{E}_\sigma \sup_{\{\theta \in H,\ \|\theta-\hat\theta\| \le r\}} \sum \sigma_i\, l_\theta(x_i)$$

expect r is the right result:

$\|\hat\theta - \theta^*\| \lesssim r$

$\Downarrow$ ss

$\phi(r)$ $\Rightarrow$ using Rademacher

$\phi(r) = r \Rightarrow r$ is the final error ??

previous: $\sup_{\theta \in H} |\mathbb{E} - \hat{\mathbb{E}} f|$

Now : $\sup_{\theta \text{ lies in the neighbor of } \theta^*} |\mathbb{E} - \hat{\mathbb{E}} f|$

Localize Leads to Fast Rate

localitie $\|\theta - \theta^*\| \lesssim \delta$

Assume Loss $\lesssim \|\theta - \theta^*\|^\alpha$

— localized Rademacher complexity linear respect to δ

$$\frac{\delta}{\sqrt{n}} = \delta^\alpha \Rightarrow \delta = (\sqrt{n})^{\frac{1}{1-\alpha}}$$

$$\Rightarrow \text{final loss} = \delta^\alpha = \left(\frac{1}{\sqrt{n}}\right)^{\frac{\alpha}{\alpha-1}}$$

Non-parametric Least Square

To estimate the unknown regression function f^*, we consider the empirical risk minimizer (ERM), which is given by

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2. \tag{2}$$

Method 1

Proof of Theorem 1: Since $\hat{f}$ is optimal to the ERM problem (2) and $f^* \in \mathcal{F}$ is feasible, we have

$$\frac{1}{n}\sum_{i=1}^{n}(\overset{f^*(x_i)+w_i}{y_i} - \hat{f}(x_i))^2 \leq \frac{1}{n}\sum_{i=1}^{n}(y_i - \cancel{f^*(x_i)}^{\,w_i})^2. \tag{3}$$

Also recall that

$$= \Sigma_{i=1}^{n} (f^* - \hat{f})^2 + 2(f^* - \hat{f})w_i + \cancel{w_i^2}$$

$$y_i = f^*(x_i) + \sigma w_i, \quad 1 \leq i \leq n.$$

We plug this expression into y_i's in equation (3), open the squares and rearrange terms. Doing so gives the "basic inequality"

loss ≤ complexity → is a linear growth.

loss is quadratic growth

$$\frac{1}{2}\|\hat{f} - f^*\|_n^2 \leq \frac{\sigma}{n}\sum_{i=1}^{n} w_i(\hat{f}(x_i) - f^*(x_i)) \tag{4}$$

≤ Rad Complexity.

Introducing the shorthand $\Delta := \hat{f} - f^* \in \mathcal{F}^*$, we rewrite the above basic inequality compactly as

$\|\Delta\|^2 \leq C \cdot \|\Delta\|$

$\Rightarrow \|\Delta\| \leq C$

$$\frac{1}{2}\boxed{\|\Delta\|_n^2} \leq \frac{\sigma}{n}\boxed{\sum_{i=1}^{n} w_i \Delta(x_i)}. \tag{5}$$

(☆) ⇒ $\|\Delta\|$ can be bounded by the linear growth speed of the localized complexity

square term

linear term.

We need star shape

localized complexity

$G_n(\delta; F^*) := \mathrm{Rad}(\{\|\theta - \theta^*\| \le \delta, \theta \in H\})$

Lemma 1. *If $\mathcal{F}^*$ is star-shaped, then the function $\delta \mapsto \frac{G_n(\delta;\mathcal{F}^*)}{\delta}$ is non-increasing on $(0, \infty)$. Hence δ^* exists and is finite.*

$\delta_1 \le \delta_2 \Rightarrow \frac{G_n(\delta_1)}{\delta_1} \ge \frac{G_n(\delta_2)}{\delta_2}$

Proof For any $0 < \delta < t$, we want to show that $\frac{G_n(t,\mathcal{F}^*)}{t} \le \frac{G_n(\delta;\mathcal{F}^*)}{\delta}$.

Given $h \in \mathcal{F}^*$ with $\|h\|_n \le t$, define the rescaled function $\tilde{h} = \frac{\delta}{t}h$. We have $\tilde{h} \in \mathcal{F}^*$ by definition with $\|h\|_n \le \delta$. It is easy to see that

$$\frac{1}{n}\left(\frac{\delta}{t}\sum_{i=1}^{n} w_i h(x_i)\right) = \frac{1}{n}\sum_{i=1}^{n} w_i \tilde{h}(x_i).$$

Taking the supreme and expectation on both side over h, we obtain that

$$\frac{\delta}{t}\mathbb{E}\left[\sup_{h\in\mathcal{F}^*:\|h\|_n\le t}\frac{1}{n}\sum_{i=1}^{n} w_i h(x_i)\right] \le \mathbb{E}\left[\sup_{\tilde{h}\in\mathcal{F}^*:\|\tilde{h}\|_n\le \delta}\frac{1}{n}\sum_{i=1}^{n} w_i \tilde{h}(x_i)\right].$$

This is equivalent to desired inequality

$$\frac{G_n(t,\mathcal{F}^*)}{t} \le \frac{G_n(\delta,\mathcal{F}^*)}{\delta}$$

the true solution

orange: localized complexity of δ_1

gray: localized complexity of δ_2

the line connects θ and θ^* should lies in the hypothesis space !

Final Error

localized complexity

$G_n(\delta; F^*) = \frac{\delta^2}{2\sigma}$ → the bou .

$$\delta^* := \min_{\delta>0}\left\{\delta : \frac{G_n(\delta;\mathcal{F}^*)}{\delta} \le \frac{\delta}{2\sigma}\right\}.$$

linear growth speed at δ^*

$$G_n(u, F^*) = \Rightarrow \mathbb{E}_\sigma \sup_{\|g\|_n \le u} \frac{\sigma}{n}\sum \sigma_i g(x_i) \le u\delta^*$$

Thm. If $u \geq \delta^*$, then $\mathbb{E}_\sigma \sup_{\|g\|_n \le u} \frac{\sigma}{n}\sum \sigma_i g(x_i) \le u\delta^*$. matches

$$G_n(u, F^*) = u \cdot \frac{G_n(u, F^*)}{u} \le u \cdot \frac{G_n(\delta^*; F)}{\delta^*} = u \cdot \boxed{\delta^*}$$

because $u \geq \delta^*$

Thm. proof of $\|\hat{f} - f^*\| \le O(\delta^*)$

① if $\|\hat{f} - f^*\| \le \delta^*$ ✓

② if $\|\hat{f} - f^*\| \geq \delta^*$. using ✱ : $\|\Delta\|^2 \le G_n(\|\Delta\|) \le \|\Delta\| \cdot \delta^*$

$\Rightarrow \|\Delta\| \le \delta^*$.

Not Required

Method 2: Peeling

empirical process

$\sup_{g \in F} |\mathbb{E} - \hat{\mathbb{E}} \; g|$

Lemma 1 (Peeling Technique) *If there is a function* $\phi : [0, \infty) \to [0, \infty)$ *and* $r^* > 0$ *s.t.* $\forall r > \hat{r}^*$, *we have*

- $\phi(4r) \leq 2\phi(r)$
- $R_n(G_r)) \leq \phi(r)$

Then we have for all $r > \hat{r}^*$ *we have*

$$\mathbb{E}_{\sigma_i, z_i}\left[\frac{\frac{1}{n}\sum_{i=1}^{n} \sigma_i g(z_i)}{\mathbb{E} g + r}\right] \leq \frac{4\phi(r)}{r}$$

$\Rightarrow$ find r^* such that $\phi(r^*) = r^*$

— normalized Empirical Process