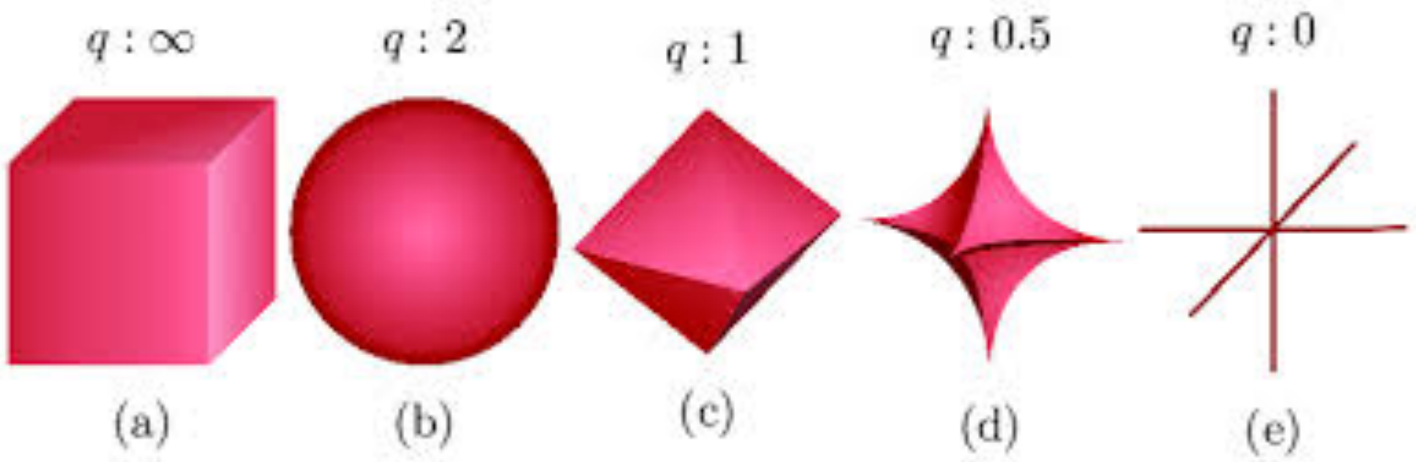
Lecture 11 Localized Complexity

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.

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

$$N(\epsilon, B_1^d, ||\cdot||_1) \leq (1 + \frac{2}{\epsilon})^d \tag{3}$$

Empirical Method of Maurey

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

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

Proof. Let's cover the following set:

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

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

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

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

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

Localized Complexity

Example: Mean Estimation

Idea:Localized Complexity

Localize Leads to Fast Rate

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

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

Also recall that

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

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

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

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

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

We need star shape

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.*

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

Given $h \in \mathcal{F}^*$ with $\|h\|_n \leq t$, define the rescaled function $\tilde{h} = \frac{\delta}{t}h$. We have $\tilde{h} \in \mathcal{F}^*$ by definition with $\|h\|_n \leq \delta$. 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

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

This is equivalent to desired inequality

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

Final Error

$$\delta^* := \min_{\delta>0}\left\{\delta : \frac{G_n(\delta;\mathcal{F}^*)}{\delta} \le \frac{\delta}{2\sigma}\right\}. \qquad \Rightarrow \qquad \sup_{\|g\|_n \le u} \frac{\sigma}{n}\sum \sigma_i g(x_i) \le u\delta^*$$

Method 2: Peeling

Lemma 1 (Peeling Technique) *If there is a function $\phi : [0, \infty) \rightarrow [0, \infty)$ 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

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