36-789 - High Dimensional Statiststics II **Fall 2015**

Lecture 1: Oct 27

Lecturer: Alessandro Rinaldo *Scribes: Kirthevasan Kandasamy*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

1.1 Minimax Theory

Minimax theory is used to show that a statistical procedure has good performance. It characterises the intrinsic difficulty of a statistical problem.

Example 1.1 ***Sparse Regression:*** *Let $Y = X\beta + \epsilon$ where $\beta \in \mathbb{R}^d$, $\epsilon \sim \mathcal{N}_n(\mathbf{0}, I_n)$ and $d > n$. (We are interested in $d \gg n$.) Let $\beta \in B_0(k)$ where $B_0(k) = \{\beta \in \mathbb{R}^d; \|\beta\|_0 \leq k\}$. Let $\hat{\beta} = \mathrm{argmin}_{\beta \in B_0(k)} \|Y - X\beta\|^2$ be the estimator. It can be shown that for some $c > 0$,*

$$\max_{\beta \in B_0(k)} \mathbb{E}_\beta[\|\hat{\beta} - \beta\|^2] \leq \frac{c\sigma^2 k \log(d/k)}{n}$$

Under some assumptions on X, we can show that there exists $c' > 0$,

$$\inf_{\hat{\beta}} \max_{\beta \in B_0(k)} \mathbb{E}_\beta[\|\hat{\beta} - \beta_0\|^2] \geq \frac{c'\sigma^2 k \log(d/k)}{n}$$

Here $\sigma^2 k \log(d/k)/n$ is the minimax rate and the estimator $\hat{\beta}$ is said to be minimax optimal.

1.1.1 General Set Up

1.1.1.1 Step 1

Let $\mathcal{P}$ be a class of prob distributions on $(\mathcal{X}, \mathcal{B})$, $\mathcal{P} = \{p_\theta; \theta \in \Theta\}$. Below are some examples.

1. $p_\theta = \mathcal{N}(\theta, I_d)$ where $\theta \in \Theta \subset \mathbb{R}^d$. For instance, $\Theta = \{\theta; \|\theta\|_0 \leq k\}$ is the set of k-sparse vectors, or $\Theta = B_q(r)$ is the ℓ_q-norm ball of radius r.
2. Approximately sparse regression: $Y = X\beta + \epsilon$ where $\beta \in \{\beta \in \mathbb{R}^d; \left(\sum_{i=1}^d |\beta_i|^q\right)^{1/q} \leq r\}$ and $q \in [0, 1]$.
3. Hypothesis testing: $Y = X\beta + \epsilon$ and β is k-sparse. The null is $H_0 : \beta = 0$.
4. Nonparametric regression: Observe Y_1^n, where $Y_i = f(X_i) + \epsilon_i$, $X_i \sim \mathrm{Unif}(0, 1)$ and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ for all i and $f \in \mathcal{F}$ where $\mathcal{F}$ is a class of smooth functions. We would like a lower bound of the form,

$$\inf_{\hat{f}} \sup_{f \in \mathcal{F}} \mathbb{E}[\|\hat{f} - f\|_2^2] \;\geq\; c\psi_n$$

   for some constant c.

5. Density Estimation: We observe $X_1^n \sim p \in \mathcal{P}$ where $\mathcal{P}$ is a class of smooth densities. We would like a result of the form,
$$\inf_{\hat{p}} \sup_{p \in \mathcal{P}} \mathbb{E}_p[\|\hat{p} - p\|_q] \geq c\psi_n$$

N.B. 1.2 *When $\mathcal{P} = \{p_\theta; \theta \in \Theta\}$ it is instructive to think of θ as the map $\theta : \mathcal{P} \to \Theta$.*

1.1.1.2 Step 2

Let $d : \Theta \times \Theta \to \mathbb{R}$ be a semi-metric; i.e. it satisfies the following:

- $d(\theta, \theta') = d(\theta', \theta)$ for all $\theta, \theta' \in \Theta$.
- $d(\theta, \theta') \leq d(\theta, \theta'') + d(\theta'', \theta')$ for all $\theta, \theta', \theta'' \in \Theta$.
- $d(\theta, \theta) = 0$.

1.1.1.3 Step 3

Define the "loss function" $w : [0, \infty) \to [0, \infty)$, where $w \neq \mathbf{0}$ and $w(0) = 0$. Some examples are $w(x) = x^2$ and $w(x) = \mathbb{1}(x > \tau)$.

1.1.2 Minimax Risk

Given $X_1^n \sim p \in \mathcal{P}$ we would like to estimate $\theta(p)$ for an estimator $\hat{\theta} : \mathcal{X}^n \to \Theta$. The point-wise risk at p is $\mathbb{E}_p[w(d(\hat{\theta}, \theta))]$. Some examples:

1. In estimation, $w(d(\hat{\theta}, \theta)) = \|\hat{\theta} - \theta\|_2^2$.
2. In hypothesis testing, $\mathcal{P} = \{p_0, p_1\}$ and $\phi : \mathcal{X}^n \to \{0, 1\}$ ($\phi(X_1^n) = 1$ means we reject H_0). The risk is $c_0\mathbb{E}[\phi(X)] + c_1\mathbb{E}[1 - \phi(X)]$. (This doesn't fall into the above framework.)

Definition 1.3 (Maximal Risk) *The maximal risk for an estimator $\hat{\theta}$ is*
$$\gamma_n(\hat{\theta}) = \sup_{p \in \mathcal{P}} \mathbb{E}_p[w(d(\hat{\theta}, \theta))]$$

Typically, we have an upper bound of the form $\gamma_n(\hat{\theta}) \leq C\psi_n$ where $\psi_n \to 0$ as $n \to 0$.

Definition 1.4 (Minimax Risk) *The minimax risk is the infimum of γ_n over all estimators $\hat{\theta}$,*
$$\mathcal{R}(n, \mathcal{P}) = \inf_{\hat{\theta}} \sup_{p \in \mathcal{P}} \mathbb{E}_p[w(d(\hat{\theta}, \theta))]$$

It is a problem dependent quantity and doesn't depend on the statistical procedure.

Suppose we have an estimator satisfying $\gamma_n(\hat{\theta}) \in \mathcal{O}(\psi_n^*)$. Then ψ_n^* is the optimal rate of convergence if $\mathcal{R}(n, \mathcal{P}) \in \Omega(\psi_n^*)$. Then we have $\mathcal{R}(n, \mathcal{P}) \in \Theta(\psi_n^*)$.

1.2 A General Strategy to Obtain Minimax Rates

We follow a three step procedure:

1. Reduction to a probability bound:

$$\mathbb{E}_p[w(d(\hat{\theta},\theta))] \geq w(\delta)\mathbb{P}(d(\hat{\theta},\theta) \geq \delta)$$

2. Reduction to a finite number of hypotheses,

$$\inf_{\hat{\theta}} \sup_{p\in\mathcal{P}} \mathbb{P}_p(d(\hat{\theta},\theta(p)) > \delta) \geq \inf_{\hat{\theta}} \underbrace{\max_{p\in\mathcal{P}_M} \mathbb{P}_p(d(\hat{\theta},\theta(p)) > \delta)}_{(\star)}$$

   where $\mathcal{P}_M = \{p_{\theta_0}, p_{\theta_1}, \ldots, p_{\theta_M}\} \subset \mathcal{P}$.

3. Recast $(\star)$ as a hypothesis testing problem where $p_{\theta_0}, p_{\theta_1}, \ldots, p_{\theta_M}$ are possible distributions and we need to select (based on our observations X_1^n) which of them has produced X_1^n.

For step 3, we usually use a 2δ packing argument. Suppose that $d(\theta(p_i),\theta(p_j)) \geq 2\delta$ for all $i \neq j$. Denote the minimum distance test by $\phi^*(X) = \text{argmin}_{k\in\{0,1,\ldots,M\}} d(\hat{\theta},\theta(p_k))$. Then by the triangle inequality, for any $\hat{\theta}$

$$\mathbb{P}_{p_j}(d(\hat{\theta},\theta(p_i)) > \delta) \geq \mathbb{P}_{p_j}(\phi^*(X) \neq j)$$

Then,

$$\inf_{\hat{\theta}} \sup_{p\in\mathcal{P}_M} \mathbb{P}_{p_j}(d(\hat{\theta},\theta(p)) > \delta) \;\geq\; \underbrace{\inf_{\psi} \max_{j\in\{0,1,\ldots,M\}} \mathbb{P}_{p_j}(\phi^*(X) \neq j)}_{p_e(m,\delta)}.$$

Any lower bound for $p_e(m,\delta)$ will give a lower bound for estimation. In particular, if $p_e(m,\delta) \geq c > 0$ then $\mathcal{R}(n,\mathcal{P}) \geq c \cdot w(\delta)$. Usually $\delta = \delta_n$ depends on n and $\delta_n \to 0$. For this to be a mimimax rate we should find $\hat{\theta}$ such that $\gamma_n(\hat{\theta}) \in \mathcal{O}(w(\delta_n))$.

We have two competing objectives in choosing $\mathcal{P}_M = \{p_0, p_1, \ldots, p_M\}$:

- Choose $p_0, p_1, \ldots, p_M$ so that they are maximally separated in the d-metric.
- However, probabilistically they should be indistinguishable. That is, a testing problem involving these distributions should be difficult.