Statistics 210B Lecture 1 Notes

Daniel Raban

January 18, 2022

1 Introduction to High-Dimensional Statistics

1.1 Overview of the course

The first half of this course will cover theoretical tools used to establish theorems in high-dimensional statistics:

- Concentration inequalities
- Empirical process theory
- Gaussian process theory and random matrix theory

The second half of this course will cover statistical problems:

- Covariance estimation
- Sparse estimation problem
- Principal component analysis (PCA) in hihg dimension
- Non-parametric regression
- Minimax lower bounds

1.2 A motivating example: sparse estimation

Here is a motivating example:

Example 1.1 (High dimensional sparse estimation)**.** Here is the assumption of our statistical model. We observe

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \in \mathbb{R}^n, \qquad X = \begin{bmatrix} X_1^\top \\ \vdots \\ X_n^\top \end{bmatrix} \in \mathbb{R}^{n \times d}, \qquad X_i \in \mathbb{R}^d.$$

We assume that the relationship $Y = X\theta^* + \omega$ holds, where $\theta^* \in \mathbb{R}^d$ with $\theta^* = \begin{bmatrix} \theta_1^* \\ \vdots \\ \theta_d^* \end{bmatrix}$ and $\omega \in \mathbb{R}^n$ is noise. In the high dimensional case, we have $n \ll d$, so standard linear regression will not be useful.

To deal with the problem in the high-dimensional case, we make the further assumption that θ^* is supported on $S \subseteq \{1, 2, \dots d\}$, with $|S|$ denoted by s; that is, θ_i^* can be nonzero only on the indices in S. This is called an **s-sparse** assumption. Our task is that given (Y, X), we want to estimate θ^*.

We present results without proof, although we will develop these results later in the course.

(a) The naive estimator (assuming $\omega_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$) is

$$\widehat{\theta}_{\text{LS}} := \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{2n} \|Y - X\theta\|_2^2.$$

Classical theory tells us that

$$\begin{aligned} \mathbb{E}[\|\widehat{\theta}_{\text{LS}} - \theta^*\|_2^2] &= \frac{\text{tr}(X^\top X)^{-1}}{n} \sigma^2 \\ &= \Theta\left(\frac{d}{n}\sigma^2\right) \end{aligned}$$

If $n \ll d$, then $\mathbb{E}[\|\widehat{\theta}_{\text{LS}} - \theta^*\|_2^2] \gg 1$. This estimator, however does not use the assumption that $\theta^* \in \mathbb{R}^d$ is s-sparse.

(b) The LASSO estimator[1] is

$$\widehat{\theta}_{\text{LASSO}} := \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{2n} \|Y - X\theta\|_2^2 + \lambda_n \|\theta\|_1,$$

which has an L^1 penalty. Our goal is to show that

$$\|\widehat{\theta}_{\text{LASSO}} - \theta^*\|_2 \lesssim c\sqrt{\frac{s \log d}{n}}.$$

We need the following condition:

Definition 1.1. The matrix X satisfies the **restricted eigenvalue (RE)**[2] condition over S with parameter (κ, α) if

$$\underbrace{\frac{1}{n}\|X\Delta\|_2^2}_{= \frac{1}{n}\langle \Delta, X^\top X \Delta \rangle} \geq \kappa \|\Delta\|_2^2 \qquad \forall \Delta \in \mathbb{C}_\alpha(S) := \{\Delta \in \mathbb{R}^d : \|\Delta_{S^c}\|_1 \leq \alpha \|\Delta_S\|_1\}.$$

[1] LASSO comes from Tbshrani in 1994 and Chen, Donoho, and Saunders in 1994, as well.

[2] This condition was introduced by Bickel, Ritov, and Tsybakov in 2009.

This is a geometric assumption on $\mathcal{L}(\theta) = \frac{1}{2n}\|Y - X(\theta_* + \Delta)\|_2^2$. If X is (κ, α)-RE, then $\mathcal{L}(\Delta)$ is strongly convex in the cone $\mathbb{C}_\alpha(S)$.

Theorem 1.1. *Suppose θ^* is supported on S, with $|S| = s$, and X satisfies the RE condition over S with parameter $(\kappa, 3)$. Further assume that $\lambda_n \geq 2\|\frac{X^\top \omega}{n}\|_\infty$. Then*

$$\|\widehat{\theta}_{\text{LASSO}} - \theta^*\|_2 \leq \frac{3}{\kappa}\sqrt{s}\lambda_n.$$

What does this mean? The sparsity assumption is more natural; for example, if we are dealing with gene data in biology, we may assume that only a few genes will determine a trait. Let's now tackle a few questions about our assumptions:

1. When does RE hold?
2. How large is $2\|X^\top \omega\|_\infty / n$?
3. How can we compare the bound with the least squares estimator?

Make the assumption that $X_I \overset{\text{iid}}{\sim} N(0, \text{Id})$ (which can be generalized) and $\omega_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$. Here are the answers to our questions:

1. **Proposition 1.1.** *Suppose $(X_i)_{i\in[n]} \overset{\text{iid}}{\sim} N(0, \text{Id})$. Fix $S \subseteq [d]$ with $|S| = s$. Then there exist universal constants $0 < c_1 < 1 < c_2$ such that when $n \geq c_2 s \log d$, we have*

$$\mathbb{P}(\tfrac{1}{2n}\|X\Delta\|_2^2 \geq c_1\|\Delta_2\|^2 \quad \forall \Delta \in \mathbb{C}_3(s)) \geq 1 - \frac{e^{-n/32}}{1 - e^{-n/32}}.$$

   This tells us that the $(c_1, 3)$–RE condition is satisfied with high probability (w.h.p.) as long as $n \geq s \log d$. To establish this proposition, we need to use empirical process theory and concentration inequalities.

2. **Lemma 1.1.** *Suppose that $\max_{i\in[n]} \|x_i\|_2/\sqrt{n} \leq B_n$ and $\omega_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$. Then there is a universal constant c such that for all $t > 0$,*

$$\mathbb{P}\left(\frac{\|X^\top \omega\|_\infty}{n} \leq cB_n\sigma\left(\sqrt{\frac{2\log d}{n}} + t\right)\right) \geq 1 - 2e^{-nt^2/2}$$

   Moreover, when $X_i \sim N(0, \text{Id})$, then for all $t \in (0, 1)$,

$$\mathbb{P}\left(\max_{i\in[n]} \frac{\|X_i\|_2^2}{n} \leq 1 + t\right) \geq 1 - ne^{-nt^2/8}.$$

This lemma tells us that

$$\frac{2\|X^\top \omega\|_\infty}{n} \le \underbrace{\widetilde{C}\sigma\sqrt{\frac{\log(d/\delta)}{n}}}_{\lambda_n}$$

with probability at least $1-2\delta$. To establish this lemma, we need concentration inequalities and empirical process theory.

3. Plug in $\lambda = \widetilde{C}\sigma\sqrt{\frac{\log(d/\delta)}{n}}$ to get

$$\|\widehat{\theta}_{\text{LASSO}} - \theta^*\|_2 \le \frac{3}{\kappa}\sqrt{s}\lambda_n = \frac{3}{\kappa}\widetilde{C}\sigma\sqrt{\frac{s\log(d/\delta)}{n}}$$

with probability at least $1-3\delta$. This means that as long as $n \gtrsim s\log(d/\delta)$,

$$\|\widehat{\theta}_{\text{LASSO}} - \theta^*\|_2^2 \ll 1.$$

In comparison, $\mathbb{E}[\|\widehat{\theta}_{\text{LS}} - \theta^*\|_2^2] = \Theta(\frac{d}{n}\sigma^2)$, which needs $n \ge d$ to be small.

1.3 Relationships with other statistical topics

Here are the relationships between this course and other courses:

- Stat 210A Theoretical Statistics: In statistical decision theory, we have a statistical model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ with a statistical procedure $\delta : D \to \Theta$ and a loss function $|ell : \Theta \times \Theta \to \mathbb{R}$. We can then calculate the risk function $R(\theta;\delta) = \mathbb{E}_\theta[\ell(\theta;\delta(Z))]$. We can compare risk functions for different procedures by looking at summarized statistics of the risk function:
  - Bayes risk: We assume $\theta \sim \pi$, so $R_B(\pi;\delta) = \mathbb{E}_{\theta\sim\pi}[R(\theta;\delta)]$.
  - Minimax: We can look at $R_M(\Theta;\delta) = \sup_{\theta\in\Theta} R(\theta;\delta)$.

  In our example, LASSO is approximately minimax optimal
- CS 281A/Stat 241A Statistical learning theory: This focuses on a different (but related) collection of models (empriical risk minimization). We study them using a similar set of tools (concentration inequalities, empirical process theory).
- Stat 260 Mean field asymptotics in statistical learning: Here, we focus on the same collections of statistical models but study them in another regime ($n, d \to \infty$ with $n/d \to$ constant asymptotics). We use different collection of tools (statistical physics, AMP, Gaussian comparison). This needs stronger assumptions but gives more refined results.

Other useful courses are convex optimization and information theory. These courses are important in order to learn deep learning theory and reinforcement learning theory. In the next lecture, we will start learning about concentration inequalities.