Lecture 1: Introduction

Isabella Zhu

4 February 2025

§1 Logistics

Class is divided into two parts with spring break as partition.

Part 1. Taught by Rigollet, concentration inequalities, i.e. how close $\frac{1}{n}\sum_{i=1}^{n} X_i$ is to $\mathbb{E}[X]$. Linear regression, sparse linear regression, matrix estimation.

Part 2. Taught by Wainwright, confidence intervals and empirical process theory. Nonparametric regression/least squares. Lower bounds (?)

4 to 6 psets, no exams.

§2 Asymptotic vs. Non-Asymptotic

Consider set of points in $\mathbb{R}^d$, arrange in matrix X where

$$X = \begin{bmatrix} - & p_1 & - \\ - & p_2 & - \\ \cdots & \cdots & \cdots \\ - & p_n & - \end{bmatrix}$$

with dimensions $n \times d$.

Some statistical questions.

1. **Mean estimation**. If $X_1, X_2, \ldots X_n \overset{\text{i.i.d.}}{\sim} P$ over $\mathbb{R}$. Use CLT to inform how good of estimate $\bar{X}_n$ is of μ. This is *classical asymptotics*.
2. **Quadratic risk**. $\mathbb{E}[\bar{X}_n - \mu]^2 = \frac{\sigma^2}{n}$.
3. **Tail bounds**. $\mathbb{P}(|\bar{X}_n - \mu| > t) \leq ke^{-cnt^2/\sigma^2}$ for every n and every t.

Notice that 2 and 3 are *non-asymptotic*, so valid for all n, not just n large.

In non-asymptotic statistics, we care about the dimension d of each point.

Some statistical questions (related to dimension).

1. **Covariance estimation**. $X_1, X_2, \dots X_n \overset{\text{i.i.d.}}{\sim} P$ on $\mathbb{R}^d$. Define X as above. We have
$$Cov(X) = \mathbb{E}[XX^T] = \Sigma \geq 0.$$
Can estimate with $\frac{1}{n}\sum_{i=1}^n X_i X_i^T$.

   Classical asymptotics would tell us
$$\sqrt{n}u^T(\hat{\Sigma} - \Sigma)v \to N(0, var((X^Tu)(X^Tv)))$$
   For non-asymptotic, we get
$$\mathbb{E}||\hat{\Sigma} - \Sigma||_F^2 = \frac{d^2}{n}$$
   but this is annoying to analyze when n is not much greater than d.

2. **Tail bounds.** Specifically,
$$\mathbb{P}(\max_{ij} |\hat{\Sigma}_{ij} - \Sigma_{ij}| > t)$$
   which is something we will investigate almost every class.

   This can be bounded with union bound and Chebyshev
$$\begin{aligned}\mathbb{P}(\max_{ij} |\hat{\Sigma}_{ij} - \Sigma_{ij}| > t) &= P(\forall i, j : \ |\hat{\Sigma}_{ij} - \Sigma_{ij}| > t)\\ &\leq \sum_{ij} \mathbb{P}(|\hat{\Sigma}_{ij} - \Sigma_{ij}| > t)\\ &\leq \sum_{ij} \frac{var(\hat{\Sigma}_{ij})}{t^2} \propto \frac{d^2}{nt^2}\end{aligned}$$
   Note that this is a terrible bound. We will improve on this in later classes.

Remark 2.1. To drive the point home, *classical asymptotics* refers to when $n \to \infty$, or in other words, when $n >> d$. A *non-asympotic* approach covers the case when n is smaller than or comparable to d.

§2.1 Random Matrix Theory

Remark 2.2. Don't need to understand where these results come from, but we will be using these results.

Take a matrix of random Gaussian noise and take the spectral decomposition to get eigenvalues $\lambda_1, \dots \lambda_n$.

Assumption 2.3

Let $\Sigma = I_d$ for simplification. $X_1, \ldots X_n \overset{\text{i.i.d.}}{\sim} N(0, I_d)$.

Let $n \to \infty$ and $d \to \infty$, constrained to constant **aspect ratio** which means

$$\frac{d}{n} \to \gamma < 1$$

Consider eigenvalues $\hat{\lambda_1}, \ldots \hat{\lambda_d}$ of $\hat{\Sigma}$. Note that $\Sigma = I_d$ has eigenvalues $1, 1, \ldots 1$.

Question 2.4. Does $\{\hat{\lambda_1}, \ldots \hat{\lambda_d}\}$ converge to $\{1\}$?

Turns out, the answer is no.

Theorem 2.5

The empirical probability distribution of eigenvalues converges as follows:

$$\frac{1}{d}\sum_{j=1}^{d} \delta_{\hat{\sigma}_j} \to P_\gamma$$

where P_γ is a probability distribution with density

$$f_\gamma = \frac{1}{2\pi}\frac{\sqrt{(\gamma_+ - x)(x - \gamma_-))}}{\gamma x}$$

where $x \in [\gamma_-, \gamma_+]$ and $\gamma_\pm = (1 \pm \sqrt{\gamma})^2$.

Corollary 2.6

(Bai-Yin) $\lambda_{max}(\hat{\Sigma}) \to (1 + \sqrt{\gamma})^2$. This implies that the eigenvalues do not in fact converge to 1.

Corollary 2.7

(Tracy-Widan) $\lambda_{max}(\hat{\Sigma})$ fluctuates with variance on the order of $\frac{1}{n^{2/3}}$ instead of the typical $\frac{1}{\sqrt{n}}$.

For non-asymptotic version, we have

$$\mathbb{P}(\lambda_{max}(\hat{\Sigma}) \geq (1 + \sqrt{d/n} + t)^2) \leq ce^{-cnt^2/2}.$$

but this bound is suboptimal as Tracy-Widen implies we should be able to do better.

§3 Summary

We discussed three types of asymptotics today.

1. **Classical asymptotics** CLT, gives exact constants, but requires $n \to \infty$.

2. **High-dimensional asymptotics** $n \to \infty$, $d \to \infty$, maintaining constant aspect ratio. Random matrix theory results. Gives exact constants while allowing for large d. Delicate method, limited scope.

3. **Non-asymptotic** Replace the $z_{\alpha/2}$ from CLT with an unspecified constant c, which is either large or unspecified unfortunately. This is what we will be investigating in this class.