Lecture 0: Short Introduction

Instructor: Ke Wei *Scribe: Ke Wei (Updated: 2023/02/20)*

Keywords: concentration inequalities, expectation of suprema, uniform law of large numbers, random matrices, minimax lower bounds.

This lecture provides a short introduction on what we are interested in this course and why we are interested in them. We begin with the arguably simplest example.

Given n i.i.d random variables $X_1, \cdots, X_n$ with mean $\mu = \mathbb{E}[X_k]$, maybe the most common approach to infer μ is to use the sample mean $\frac{1}{n}\sum_{k=1}^n X_k$ as (stochastic) estimator. Then what can we say about the convergence of $\frac{1}{n}\sum_{k=1}^n X_k$?

- By law of large numbers (LLN), it is known that $\frac{1}{n}\sum_{k=1}^n X_k$ converges to μ almost surely.
- Suppose the variance of the random variable is σ^2. Central limit theorem (CLT) implies that

$$\frac{\sum_{k=1}^n (X_k - \mu)}{\sigma\sqrt{n}} \to g \sim \mathcal{N}(0,1),$$

  from which a confidence interval can be constructed (in the asymptotic sense). It follows that

$$\begin{aligned}\mathbb{P}\left[\left|\frac{\sum_{k=1}^n (X_k - \mu)}{n}\right| \geq t\right] &= \mathbb{P}\left[\left|\frac{\sum_{k=1}^n (X_k - \mu)}{\sigma\sqrt{n}}\right| \geq \frac{t\sqrt{n}}{\sigma}\right] \\ &\approx \mathbb{P}\left[|g| \geq \frac{t\sqrt{n}}{\sigma}\right] \leq 2e^{-\frac{nt^2}{2\sigma^2}}.\end{aligned}$$

Note that both LLN and CLT are asymptotic results which are not useful if we want to measure how $\frac{1}{n}\sum_{k=1}^n X_k$ deviates from μ for finite n. This lecture considers the finite n or nonasymptotic case. That is, we would like to provide an explicit bound for

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{k=1}^n X_k - \mu\right| \geq t\right], \tag{0.1}$$

in contrast to $\approx$ obtained from CLT. Though an explicit bound can be obtained from CLT through

$$\mathbb{P}\left[\left|\frac{\sum_{k=1}^n (X_k - \mu)}{\sigma\sqrt{n}}\right| \geq \frac{t\sqrt{n}}{\sigma}\right] \leq \mathbb{P}\left[|g| \geq \frac{t\sqrt{n}}{\sigma}\right] + \left|\mathbb{P}\left[\left|\frac{\sum_{k=1}^n (X_k - \mu)}{\sigma\sqrt{n}}\right| \geq \frac{t\sqrt{n}}{\sigma}\right] - \mathbb{P}\left[|g| \geq \frac{t\sqrt{n}}{\sigma}\right]\right|,$$

the resulting bound is not desirable due to the second term (see Theorem 2.1.3 of [1]). Thus, we need to seek alternative approaches that bypass CLT. It turns out the asymptotic CLT result admit more quantitative nonasymptotic variants, though not as concise as CLT.

Deviation bound for (0.1) is known as **concentration inequalities** for the sum of independent random variables. However, in many applications the quantity of interest is not a sum but a nonlinear function of independent random variables. For example, let $X \in \mathbb{R}^{m\times n}$ be a random matrix[1] whose entries are independent random variables. Clearly, the spectral norm of X, denoted $\|X\|_2$ can not be expressed as the sum of its entries. Indeed, we can express $\|X\|_2$ as

$$\|X\|_2 = \max_{\|u\|_2=1,\|v\|_2=1} u^\top X v. \tag{0.2}$$

Thus, it is meaningful to extend concentration inequalities to

$$\mathbb{P}\left[\left|f(X_1,\cdots,X_n) - \mathbb{E}\left[f(X_1,\cdots,X_n)\right]\right| \geq t\right]$$

for more general f in addition to sum.

Note that in order to provide a high probability bound for the size of $f(X_1,\cdots,X_n)$, we still need to understand the size of $\mathbb{E}\left[f(X_1,\cdots,X_n)\right]$ whose information is not contained in concentration inequality. For general f, bounding $\mathbb{E}\left[f(X_1,\cdots,X_n)\right]$ is by no means an easy task. Since we cannot hope to address this problem for every possible f, we specify our attention to **expectation of suprema**:

$$\mathbb{E}\left[\sup_{t\in T} X_t\right], \quad T \text{ is an index set},$$

which arises from a wide of applications, i.e. (0.2) for spectral norm of random matrices. As a special case, we will consider the following form arising from generalization analysis in statistical learning:

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{k=1}^{n} f(X_k) - \mathbb{E}\left[f(X)\right]\right|, \quad \mathcal{F} \text{ is a class of functions.}$$

The result for this case is usually referred to as **uniform law of large numbers (ULLN)**.

The concentration inequalities for random variables can be extended to **random matrices**. We will focus on the bound of

$$\mathbb{P}\left[\left\|\sum_{k=1}^{n}(X_k - \mathbb{E}\left[X_k\right])\right\|_2 \geq t\right].$$

It has applications in for example covariance matrix estimation, sparse linear regression.

For an estimation problem, there can be many different estimators. Thus, a natural question is which one is better or whether an estimator achieves the optimal performance. The answer to this question relies on the criterion that is used. For example, a minimum-variance unbiased estimator (MVUE) is an unbiased estimator that has lower variance than any other unbiased estimators. In this lecture, we consider the minimax framework, and study the **minimax lower bounds** over a family of estimation problems.

Reading Materials

[1] Roman Vershynin, *High-dimensional probability: An introduction with applications in data science.*

[1] With a light abuse of notation, capital letters are also used to denote matrices.