High Dimensional Probability and Statistics **2nd Semester, 2021-2022**

Lecture 0: Short Introduction

Instructor: Ke Wei *Scribe: Ke Wei (Updated: 2022/02/19)*

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.

Example 0.1 *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 an estimator. Then a natural question arises: how well is the estimator? This question can be answer in different ways. For example,*

- *By the law of large numbers, 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. The central limit theorem implies that*

$$\frac{\sum_{k=1}^n X_k}{\sigma\sqrt{n}} \to \text{ standard normal distribution},$$

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

- *Assuming the variance of the variable is σ^2, the mean square error (MSE) is*

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{k=1}^n X_k - \mu\right)^2\right] \le \frac{\sigma^2}{n}.$$

This lecture considers another way to evaluate the performance of the estimator, which in more quantitative for the finite n case. More precisely, we consider the probability of $\frac{1}{n}\sum_{k=1}^n X_k$ deviating from μ by a small quantity,

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

If δ is small, then it means with high probability $\frac{1}{n}\sum_{k=1}^n$ is close to μ.

Inequalities of the type (0.1) are known as **concentration inequalities**. Evidently, it is special case of the following more general form

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

This form of inequality covers many other important applications, including the generalization error analysis in statistical learning.

Example 0.2 *Given a pair of random variable (X, Y), a central task in statical learning is to find the relationship between X and Y. This is typically formed as the problem of finding a function (hypothesis) h in a function class $\mathcal{H}$ such that the population risk*

$$R(h) = \mathbb{E}\left[\mathcal{L}(h(X), Y)\right]$$

is minimized. Here $\mathcal{L}(\cdot, \cdot)$ represents certain loss function. However, since we do not know the distribution of by only have access to a set of i.i.d samples $X_1, \cdots, X_n$, a computationally tractable alternative is to minimize the empirical risk,

$$\widehat{R}_n(h) = \frac{1}{n}\sum_{k=1}^{n} \mathcal{L}(h(X_k), Y_k).$$

Letting h^* be the minimizer of $R(h)$ and $\hat{h}_n^*$ be the minimizer of $\widehat{R}_n(h)$, in order for $\hat{h}_n^*$ to generalize well for the entire distribution, we wish $R(\hat{h}_n^*)$ should be close to $R(h^*)$. This can be achieved if $\widehat{R}_n(h)$ is close to $R(h)$ for all $h \in \mathcal{H}$ since then they will have their minimizers close to each other. More precisely, we have

$$\begin{aligned} R(\hat{h}_n^*) - R(h^*) &= \left(R(\hat{h}_n^*) - \widehat{R}_n(\hat{h}_n^*)\right) + \left(\widehat{R}_n(\hat{h}_n^*) - \widehat{R}_n(h^*)\right) + \left(\widehat{R}_n(h^*) - R(h^*)\right) \\ &\leq \left|R(\hat{h}_n^*) - \widehat{R}_n(\hat{h}_n^*)\right| + \left|\widehat{R}_n(h^*) - R(h^*)\right| \\ &\leq 2\sup_{h\in\mathcal{H}} \left|\widehat{R}_n(h) - R(h)\right|. \end{aligned}$$

Thus, in order to bound the generalization error $R(\hat{h}_n^*) - R(h^*)$, it suffices to bound

$$\sup_{h\in\mathcal{H}} \left|\widehat{R}_n(h) - R(h)\right| = \sup_{h\in\mathcal{H}} \left|\frac{1}{n}\sum_{k=1}^{n} \mathcal{L}(h(X_k), Y_k) - \mathbb{E}\left[\mathcal{L}(h(X, Y))\right]\right| \tag{0.2}$$

Furthermore, in order to provide a high probability (upper) bound for (0.2)*, we can proceed in two steps: first show that*

$$f(Z_1, \cdots, Z_n) := \sup_{h\in\mathcal{H}} \left|\frac{1}{n}\sum_{k=1}^{n} \mathcal{L}(h(X_k), Y_k) - \mathbb{E}\left[\mathcal{L}(h(X, Y))\right]\right|, \quad \text{where } Z_k = (X_k, Y_k) \tag{0.3}$$

concentrates around its mean $\mathbb{E}\left[f(Z_1, \cdots, Z_n)\right]$and then provide a bound for $\mathbb{E}\left[f(Z_1, \cdots, Z_n)\right]$. Clearly, the form of f in this example is much more complicated than that in Example 0.1.

As noted in Example 0.2, we also need to bound the expectation of the supremum of a set of random variables for generalization error analysis. A general form is the following **expectation of suprema**:

$$\mathbb{E}\left[\sup_{t\in T} X_t\right],$$

where T is an index set. For the particular case as in (0.3), it is usually referred to as **uniform law of large numbers**. In addition to generalization error analysis, computing the expectation of

suprema also appears in other important applications, such as the estimation of the spectral norm of random matrices.

The concentration inequalities for random variables can be extended to **random matrices**. With a light abuse of notation, capital letters are also used to denote matrices. We will focus on the following type of inequality:

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

where $\|\cdot\|_2$ denotes the spectral norm of a matrix. 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 minmax framework, and study the **minimax lower bounds** over a family of estimation problems.