IDS.160 – Mathematical Statistics: A Non-Asymptotic Approach

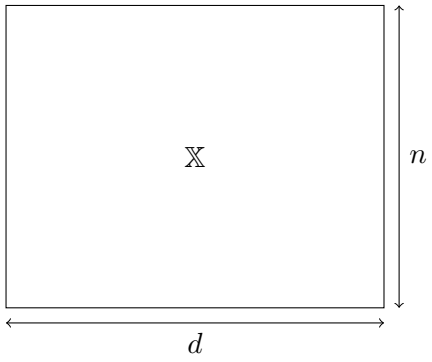
Lecturer: Philippe Rigollet
Scribe: Philippe Rigollet

Lecture 1
Feb. 4, 2020

Goals: This lecture is an introduction to the concepts covered in this class. In particular, we will discuss the difference between the *asymptotic* and *non-asymptotic* approaches to mathematical statistics.

We also give a brief overview of the some of the topics covered in class: Covariance matrix estimation, matrix estimation, empirical risk minimization, neural networks and minimax lower bounds.

A good example to keep in mind is a dataset organized as an n by d matrix $\mathbb{X}$ where, for example, the rows correspond to patients and the columns correspond to measurements on each patient (height, weight, ...). Row i is a random vector $X_i^\top \in \mathbb{R}^d$ of the measurements performed on patient i.



We now compare the *asymptotic* and *non-asymptotic* approaches to mathematical statistics.To better illustrate the difference between the two approaches, let us consider some examples.

1. ASYMPTOTIC VS. NON-ASYMPTOTIC REGIMES

1.1 Mean estimation

Here $d = 1$ and we observe $X_1, \dots, X_n \overset{\text{i.i.d}}{\sim} P$ where P is a distribution over $\mathbb{R}$ with mean μ and variance σ^2. We consider the sample mean:

$$\bar{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i \,.$$

We have the following asymptotic results:

- Law of Large Numbers (LLN): $\bar{X}_n \xrightarrow[n\to\infty]{\mathbb{P}} \mu$

- Central Limit Theorem (CLT): $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, \sigma^2)$

We can also make some non-asymptotic statements:

- Quadratic risk: $\mathbb{E}[(\bar{X}_n - \mu)^2] = \dfrac{\sigma^2}{n}$
- Tail bounds: $\mathbb{P}(|\bar{X}_n - \mu| > t) \leq 2e^{-Cnt^2}$, where $C > 0$ is a constant that depends on further assumptions on P, such as having a bounded support (Hoeffding's inequality).

1.2 Covariance matrix estimation

Assume now that we observe $X_1, \dots, X_n \overset{\text{i.i.d}}{\sim} P$ where P is a distribution over $\mathbb{R}^d$ with mean μ and covariance matrix $\Sigma = \mathbb{E}[X_1 X_1^\top]$. The sample covariance matrix $\hat{\Sigma}$ is defined as

$$\hat{\Sigma} := \frac{1}{n} \sum_{i=1}^{n} X_i X_i^\top .$$

We still have some asymptotic statements:

- LLN: $\hat{\Sigma}_{i,j} \xrightarrow[n\to\infty]{\mathbb{P}} \Sigma_{i,j}$, for all $i, j = 1, \dots, d$.
- CLT: $\sqrt{n}(\hat{\Sigma}_{i,j} - \Sigma_{i,j}) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, \operatorname{var}(\mathbb{X}_{1,i}\mathbb{X}_{1,j}))$.

To compute explicitly the variance $\operatorname{var}(\mathbb{X}_{1,i}\mathbb{X}_{1,j})$, one would need to make assumptions on the fourth moment P but this value does not matter for our considerations here.

In this case, letting $n \to \infty$ implicitly assumes that $n \gg d$. But what if n is of the order of d or even if $n \ll d$?

Can we make similar statements simultaneously for all the entries of $\hat{\Sigma}$? In other words, can we guarantee that the matrix $\hat{\Sigma}$ converges to the matrix Σ?

For example, let's say[1] that we are interested in understanding the random variable

$$|\hat{\Sigma} - \Sigma|_\infty := \max_{i,j} |\hat{\Sigma}_{i,j} - \Sigma_{i.j}|$$

Here is an attempt using a union bound.

$$\mathbb{P}(|\hat{\Sigma} - \Sigma|_\infty > t) = \mathbb{P}(\exists (i,j) \,:\, |\hat{\Sigma}_{i,j} - \Sigma_{i.j}| > t) \leq \sum_{1 \leq i,j \leq d} \mathbb{P}(|\hat{\Sigma}_{i,j} - \Sigma_{i.j}| > t) \leq C \frac{d^2}{nt^2} ,$$

assuming that we use Chebyshev's inequality to control

$$\mathbb{P}(|\hat{\Sigma}_{i,j} - \Sigma_{i.j}| > t) \leq \frac{\operatorname{var}(\hat{\Sigma}_{i,j})}{t^2} \leq \frac{C}{nt^2} .$$

In particular, this attempt fails if d grows faster than $\sqrt{n}$, that is, $d = \omega(\sqrt{n})$. While the above attempt uses loose arguments, we can show that convergence of $\hat{\Sigma}$ to Σ fails if $d/n \to \gamma \in (0, 1]$ (asymptotically fixed aspect ratio). This regime is sometimes refereed to as the *high-dimensional asymptotic* regime. We can see this by showing that the spectrum

[1] Later in the class, we'll be interested in understanding the operator norm of $\hat{\Sigma} - \Sigma$.

of $\hat{\Sigma}$ does not converge to that of Σ even in the simple case where $\Sigma = I_d$. To that end, we can use some tools from random matrix theory (RMT)[2].

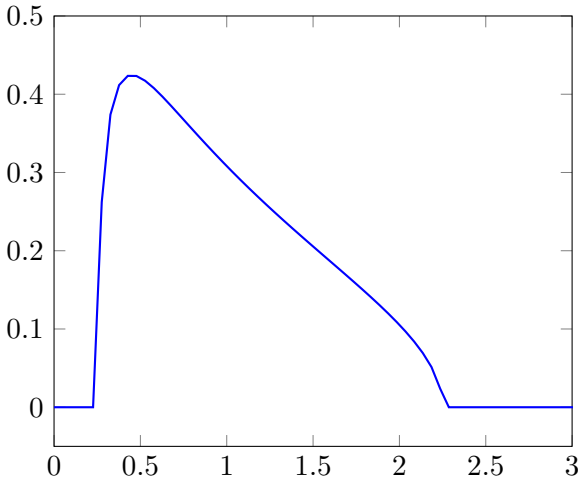
Let $\hat{\lambda}_1, \dots, \hat{\lambda}_d$ denote the eigenvalues of $\hat{\Sigma}$ and let $\lambda_1 = \dots = \lambda_d = 1$ denote those of I_d. We are going to see that the set $\{\hat{\lambda}_1, \dots, \hat{\lambda}_n\}$ does not converge to $\{1\}$. In fact it is a well known fact of RMT that

$$\frac{1}{d}\sum_{j=1}^{d} \delta_{\hat{\lambda}_j} \to T, \qquad \text{as } n \to \infty,\ d \to \infty, \frac{d}{n} \to \gamma \in (0, 1]\,,$$

where T is a random variable distributed according to the Marčenko-Pastur distribution and has density

$$f(t) = \frac{\sqrt{(\gamma_+ - t)(t - \gamma_-)}}{2\pi\gamma t} \mathbb{1}_{[\gamma_-, \gamma_+]}(t)\,, \qquad \gamma_\pm = (1 \pm \sqrt{\gamma})^2\,.$$

This density for $\gamma = 1/2$ looks like that



In particular, we can see that the eigenvalues do not concentrate at 1.

If one is interested only in the largest eigenvalue of $\hat{\Sigma}$, then it is true that $\lambda_{\max}(\hat{\Sigma}) \to \gamma_+$ and the asymptotic distribution of $\lambda_{\max}(\hat{\Sigma})$ is also known and corresponds to the Tracy-Widom distribution. In particular, one can extract the quantiles of this distribution to perform statistical inference on Σ such as hypothesis testing or confidence intervals.

Random matrix theory results are very delicate. In this class, we will see that we can get with much less effort similar qualitative results. For example, we will show that

$$\mathbb{P}(\lambda_{\max}(\hat{\Sigma}) > 1 + C(\sqrt{\frac{d}{n}} + t + \sqrt{t})) \leq e^{-nt}\,.$$

in other words, the order of the fluctuation, which is $\sqrt{d/n}$ holds for all d and n.

[2]Random matrix theory is a deep topic of probability and is not required in this class. However, knowledge of the qualitative results from this field often proves useful to get a grasp of the order of magnitude of important quantities such as the leading eigenvalue of $\hat{\Sigma}$.

1.3 Asymptotic vs non-asymptotic

In light of these examples we can see the difference between asymptotic and non-asymptotic results.

In the context of mean estimation for $d = 1$, a direct consequence of the CLT (classical asymptotic regime) is that we can build an asymptotic confidence interval for μ:

$$\mathbb{P}\big(\mu \in [\bar{X}_n - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X}_n + 1.96\frac{\sigma}{\sqrt{n}}]\big) \xrightarrow[n\to\infty]{} .95$$

It is quite useful to have exact constants (here 1.96) arising from quantiles of the standard Gaussian distribution. Similarly, sharp constants may be obtained in the high-dimensional asymptotic regime by considering the quantiles of the Tracy-Widom distribution for example.

This makes for precise confidence intervals that should be contrasted with the ones obtained in the non-asymptotic regime, using for example Hoeffding's inequality:

$$\mathbb{P}\big(\mu \in [\bar{X}_n - C\frac{\sigma}{\sqrt{n}}, \bar{X}_n + C\frac{\sigma}{\sqrt{n}}]\big) \geq .95\,,$$

where C is a constant typically larger than 1.96. Qualitatively, note that the width of the confidence interval is captured by the non-asymptotic regime: it is of order $\sigma/\sqrt{n}$.

This difference is also salient when considering hypothesis testing where it is often desirable to have sharp thresholds in order to maximize power. From this perspective, the classical asymptotic regime is more desirable.

However, note that these sharp constants are only valid as $n \to \infty$ and in particular, it requires that $n \gg d$. The high-dimensional asymptotic regime gives a partial remedy for this limitation but given n and d, it is unclear whether we are in this regime. Indeed, one may ask which of the following pairs (n, d) are in this regime:

$$(1000, 1),\ (1000, 10),\ (1000, 100),\ (1000, 1000)$$

For each of these one can estimate γ by d/n but it is unclear if asymptotic statements are valid.

Instead the non-asymptotic regime is valid for all n and d. It does not yield sharp constants but captures the performance/accuracy of the estimator $\bar{X}_n$ for all n.

In conclusion, the *classical asymptotic* or *high-dimensional asymptotic* regimes are preferred for statistical inference tasks such as confidence intervals and hypothesis testing whereas the *non-asymptotic* regime is preferred to produce a qualitative description of the performance of a possibly complicated and high-dimensional method such as the ones arising in machine learning.

In the rest of this lecture, we give an overview of the topics covered in this class from the useful perspective of the non-asymptotic regime.

2. A BRIEF OVERVIEW OF TOPICS COVERED

2.1 Empirical risk minimization

Recall our favorite statistical estimation method: maximum likelihood.

We are given a statistical model $(\mathbb{R}, \{P_\theta\}_{\theta\in\Theta})$ where the density of P_θ with respect to the Lebesgue measure is given by p_θ. Assume that we observe $X_1, \dots, X_n \overset{\text{i.i.d}}{\sim} P_{\theta^*}$ for some unknown $\theta^* \in \Theta$. The maximum likelihood estimator $\hat{\theta}$ is given by

$$\hat{\theta} = \operatorname*{argmax}_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \log p_\theta(X_i)\,.$$

Wald's theorem ensures that under some technical conditions,

$$\sqrt{nI(\hat{\theta})}(\hat{\theta} - \theta^*) \to \mathcal{N}(0,1)\,.$$

where $I(\theta)$ denotes the Fisher information: $I(\theta) = -\mathbb{E}_\theta\big[\frac{\partial^2}{\partial\theta^2}\log p_\theta(X_1)\big]$.

The maximum likelihood method belongs to a larger class of methods called *empirical risk minimization*. To see this recall that if the model is identifiable then θ^* is the unique minimizer of the expected negative log-likelihood given by

$$-\mathbb{E}_{\theta^*}[\log p_\theta(X_1)]\,.$$

Therefore, we can view the maximum likelihood method as replacing the expected value with an average and then proceeding to optimizing the resulting function. This is precisely the idea behind empirical risk minimization.

Consider a loss function $\ell(X,\theta)$ such that $\theta^* = \operatorname*{argmin}_{\theta\in\Theta} \mathbb{E}_{\theta^*}\ell(X,\theta)$. The quantity $\mathbb{E}_{\theta^*}\ell(X,\theta)$ is the *risk* of θ and can can estimate it by it *empirical risk*:

$$\frac{1}{n}\sum_{i=1}^{n} \ell(X_i,\theta)\,.$$

Empirical risk minimization consists in minimizing this function over Θ:

$$\hat{\theta} = \operatorname*{argmin}_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \ell(X_i,\theta)\,.$$

We will consider various cases for Θ. For example

- $\Theta = \mathbb{R}^d$
- Θ is a space of smooth functions
- Θ is a combinatorial subset of the boolean hypercube $\{0,1\}^d$

While Wald's theory can apply in the first case, the other two are more tricky and asymptotic normality of $\hat{\theta}$ often does not hold. Nevertheless, we will develop some tools to quantify the accuracy of $\hat{\theta}$ in the non-asymptotic regime.

2.2 Matrix denoising

Assume that we observe a $\sqrt{d}\times\sqrt{d}$ matrix $Y = W + \Xi$ where W is a matrix of interest and Ξ is a noise matrix. Such problems arise in matrix completion and community detection for example.

Without further assumptions, estimating W is hopeless: our best guess is simply Y. In fact, we will see that we can give nontrivial guarantees to estimate W when it has a natural low-dimensional structure, for example if it has low rank.

2.3 Neural networks

The problem of fitting a neural network can be viewed as a regression problem where one observes independent copies of a predictor/response pair $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ that satisfy

$$Y = f(X) + \varepsilon\,,$$

where ε is a noise random variable and $f : \mathbb{R}^d \to \mathbb{R}$ is an unknown regression function.

Neural networks have been applied successfully by fitting regression functions of the form

$$f(x) = W^L \sigma(W_{L-1}\sigma(W_{L-2} \cdots \sigma(W_1 x) \cdots)\,,$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ is a nonlinearity (ReLU, sigmoid, etc.) applied to each coordinate and W_j is a d_j-by-d_{j-1} matrix with $d_0 = d$ and $d_L = 1$. The number of parameters $\sum_j d_j d_{j-1}$ is typically much larger than the sample size n and cannot represent the true dimensionality of the problem if one can estimate f accurately. We will describe some recent developments on how the "true" dimensionality of such functions may be measured.

2.4 Minimax lower bounds

A typical non-asymptotic statistical guarantee is of the form

$$\sup_{\theta \in \Theta} \mathbb{E}_\theta \|\hat{\theta} - \theta\|^2 \leq C\frac{d}{n}\,.$$

In other words, these are uniform guarantees (unlike asymptotic statements which are pointwise). For a given problem, one may ask whether we can do better, either by finding a better proof of performance for $\hat{\theta}$ or by changing the estimator $\hat{\theta}$ altogether. Indeed, since non-asymptotic results are qualitative in nature, it is important to make sure that are painting the correct picture.

A *minimax lower bound* is a result of the form

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta \|\hat{\theta} - \theta\|^2 \geq C\frac{d}{n}\,.$$

where the infimum is over all estimators (measurable functions of the data). This reads as:

> For all estimator $\hat{\theta}$, there exists $\theta \in \Theta$ that cannot be estimated (in squared norm) faster than order d/n.

Minimax lower bounds are very informative companions to non-asymptotic lower bounds. In this class we will develop a general machinery that borrows from information theory to develop minimax lower bounds systematically.

Summary: The *classical asymptotic* or *high-dimensional asymptotic* regimes are preferred for statistical inference tasks such as confidence intervals and hypothesis testing whereas the *non-asymptotic* regime is preferred to produce a qualitative description of the performance of a possibly complicated and high-dimensional method such as the ones arising in machine learning.

The object of interest often has a latent low-dimensional structure which makes seemingly impossible estimation tasks (matrix denoising, neural networks training) possible.

Minimax lower bounds allow us to assess the optimality of non-asymptotic bounds by giving the fundamental limitations of *any* estimator constructed from the data at hand.