Stats 300b: Theory of Statistics **Winter 2019**

Lecture 14 – February 21

Lecturer: John Duchi *Scribe: Kevin Guo*

Warning: *these notes may contain factual errors*

Reading: VdV Section 5.9

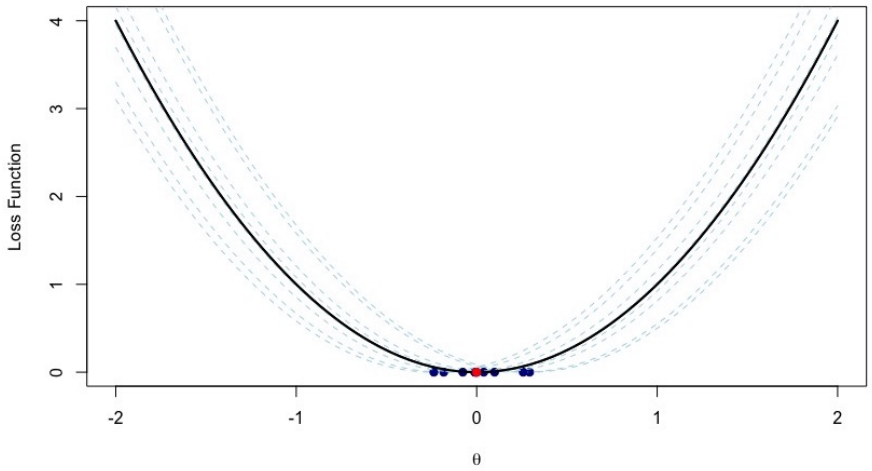
Outline:

- Rates of Convergence
  - Moduli of Continuity
  - Peeling
- High Dimensional Statistics
  - Gaussian Sequence Model

1 Rates of Convergence

Consider an M-estimation problem with loss function $\ell : \Theta \times \mathcal{X} \to \mathbb{R}$. Define the population risk $L(\theta) := P\ell(\theta, X)$ and empirical risk $L_n(\theta) := P_n\ell(\theta, X)$. We would like to estimate $\theta^* = \operatorname{argmin}_{\theta\in\Theta} L(\theta)$. A natural estimator is the minimizer of the empirical risk, $\hat{\theta}_n = \operatorname{argmin}_{\theta\in\Theta} L_n(\theta)$. When can we prove that $\hat{\theta}_n$ will be close to θ^*?

One way to prove that $\hat{\theta}_n$ is close to θ^* is to show that for all θ "far" from θ^*, we have $L_n(\theta) > L_n(\theta^*)$. This means that θ "far" from θ^* do not minimize the empirical risk, so $\hat{\theta}_n$ cannot be "far" from θ^*. Intuitively, it should be enough that the "growth" of the function L overwhelms the local "fluctuations" $|L_n - L|$, which can be seen in the following picture:



In black is the function $L(\theta)$, and in dashed lines are a few samples of $L_n(\theta)$. It is pretty clear that $|\hat{\theta}_n - \theta^*| \geq 2$ will almost never happen, because the growth of the function L is fast enough to ensure that $L(\theta) - L(\theta^*)$ is much larger than the fluctuations $|L_n(\theta) - L(\theta)|$ for $\theta \geq 2$. In simpler terms, we would have to be extraordinarily unlucky for $L_n(\theta)$ to deviate so far from $L(\theta)$ that it drops below $L_n(\theta^*)$ for some θ with $|\theta| \geq 2$. Therefore, we can expect $|\hat{\theta}_n - \theta| \leq 2$ almost all the time.

To make this heuristic argument more rigorous, we will decompose the gap $L_n(\theta) - L_n(\theta^*)$ into a "fluctuation" component and a "growth" component:

$$L_n(\theta) - L_n(\theta^*) = \underbrace{(L_n(\theta) - L(\theta)) - (L_n(\theta^*) - L(\theta^*))}_{\text{"Fluctuation"}} + \underbrace{(L(\theta) - L(\theta^*))}_{\text{"Growth"}} \tag{1}$$

We give the collection of fluctuations a name:

Definition 1.1. *The **localized process** $\Delta_n(\theta)$ is a stochastic process indexed by $\theta \in \Theta$ defined by:*

$$\Delta_n(\theta) := (L_n(\theta) - L(\theta)) - (L_n(\theta^*) - L(\theta^*))$$

In words, $\Delta_n(\theta)$ is the size of the fluctuation of L_n at θ relative to the size of the fluctuation of L_n at θ^*.

We will now give a definition and make an assumption that allows us to control the localized process.

Definition 1.2. *Define the **modulus of continuity** $\omega_n(\delta)$ of the localized process by:*

$$\omega_n(\delta) := \sup_{\theta : d(\theta, \theta^*) \leq \delta} |\Delta_n(\theta)|$$

Fluctuation assumption. Assume that the modulus of continuity satisfies:

$$\mathbb{E}[\omega_n(\delta)] \leq \frac{M}{\sqrt{n}} \delta^\alpha \tag{2}$$

for some $M < \infty$ and $\alpha > 0$. In other words, we assume that the fluctuations $L_n(\theta) - L(\theta)$ are not too large (on average) compared to the fluctuations $L_n(\theta^*) - L(\theta^*)$, where "too large" is measured relative to the distance between θ and θ^*.

We will also make an assumption on the "Growth" term.

Growth assumption. For a metric d on Θ and constants $\beta \geq 1$, $\lambda > 0$, assume that the population risk $L(\cdot)$ satisfies:

$$L(\theta) \geq L(\theta^*) + \lambda d(\theta, \theta^*)^\beta \tag{3}$$

for all θ satisfying $d(\theta, \theta^*) \leq \epsilon$.

Using the above assumptions, we can actually compute a rate of convergence for the M-estimator $\hat{\theta}_n$, if we additionally assume that the estimator is consistent.

Theorem 1. *Suppose that $\hat{\theta}_n \xrightarrow{p} \theta^*$ and the Fluctuation Assumption and Growth Assumption are satisfied with $\alpha < \beta$. Then we have the convergence rate*

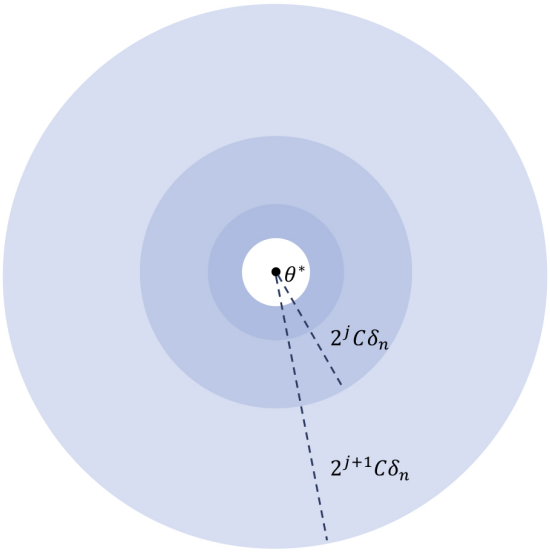
$$d(\hat{\theta}_n, \theta^*) = \mathcal{O}_P\left(n^{-\frac{1}{2(\beta-\alpha)}}\right) \tag{4}$$

Remark. Where does the rate in (4) come from? We expect $L_n(\theta) > L_n(\theta^*)$ whenever the "growth" exceeds the "fluctuation." By comparing the upper bounds from the Growth and Fluctuation assumptions, this happens when $\lambda\delta^\beta \geq M\delta^\alpha/\sqrt{n}$. Solving for δ gives $\delta \asymp n^{-\frac{1}{2(\beta-\alpha)}}$. Of course, we justify these intuitions in the proof of the theorem.

Proof . Set $\delta_n = (\frac{M}{\lambda\sqrt{n}})^{\frac{1}{\beta-\alpha}}$. Fix $C > 1$. If $d(\hat{\theta}_n, \theta^*) \geq C\delta_n$, then (at least) one of the following must occur:

- The distance between $\hat{\theta}_n$ and θ^* exceeds ϵ, so the inequality (3) does not apply (recall that the Growth Assumption only assumes that (3) holds in an ϵ-neighborhood of θ^*).
- There exists an integer $j \in [0, \lfloor \log_2(\epsilon/C\delta_n) \rfloor]$ such that $d(\hat{\theta}_n, \theta^*) \in [2^j C\delta_n, 2^{j+1}C\delta_n]$.

In the second case, we have subdivided the set $\{\theta \in \Theta : C\delta_n \leq d(\theta, \theta^*) \leq C\delta_n 2^{\lfloor \log_2(\epsilon/C\delta_n) \rfloor}\}$ into concentric "shells" of doubling width[1], as illustrated in the following picture:



We will show that the localized process is well-controlled on each shell. This argument is known as

[1]The "outermost" shell may contain points farther from θ^* than ϵ unless ϵ is exactly a multiple of 2 times $C\delta_n$.

peeling. Taking a union bound gives:

$$\begin{aligned}\mathbb{P}\left(d(\hat{\theta}_n, \theta^*) \geq C\delta_n\right) &\leq \mathbb{P}\left(d(\hat{\theta}_n, \theta^*) \in [C\delta_n, \epsilon)\right) + \mathbb{P}\left(d(\hat{\theta}_n, \theta^*) \geq \epsilon\right) \\ &\leq \sum_{j=0}^{\lfloor \log_2(\epsilon/C\delta_n)\rfloor} \mathbb{P}\left(d(\hat{\theta}_n, \theta^*) \in [C2^j\delta_n, C2^{j+1}\delta_n]\right) + \mathbb{P}\left(d(\hat{\theta}_n, \theta^*) \geq \epsilon\right)\end{aligned}$$

Note that the event $d(\hat{\theta}_n, \theta^*) \in [C2^j\delta_n, C2^{j+1}\delta_n]$ implies $\omega_n(C2^{j+1}\delta_n)$ is at least $\lambda(C2^j\delta_n)^\beta$. This is because $d(\hat{\theta}_n, \theta^*) \neq 0$ means $L_n(\hat{\theta}_n) - L_n(\theta^*) < 0$. By (1), this implies that $0 \leq \Delta_n(\theta) + (L(\theta) - L(\theta^*))$, which guarantees $|\Delta_n(\theta)| \geq L(\theta) - L(\theta^*)$. By the Growth Assumption and the fact that ω_n is an increasing function of its argument, we have:

$$\omega_n(C2^{j+1}\delta_n) \geq \omega_n(d(\hat{\theta}_n, \theta^*)) \geq |\Delta_n(\hat{\theta}_n)| \geq L(\hat{\theta}_n) - L(\theta^*) \geq \lambda d(\hat{\theta}_n, \theta^*)^\beta \geq \lambda(C2^j\delta_n)^\beta$$

Therefore, we can further upper bound $\mathbb{P}(d(\hat{\theta}_n, \theta^*) \geq C\delta_n)$ by enlarging the events $\{d(\hat{\theta}_n, \theta^*) \in [C2^j\delta_n, C2^{j+1}\delta_n]\}$:

$$\mathbb{P}\left(d(\hat{\theta}_n, \theta^*) \geq C\delta_n\right) \leq \sum_{j=0}^{\lfloor \log_2(\epsilon/C\delta_n)\rfloor} \mathbb{P}\left(\omega_n(C2^{j+1}\delta_n) \geq \lambda(C2^j\delta_n)^\beta\right) + \mathbb{P}\left(d(\hat{\theta}_n, \theta^*) \geq \epsilon\right)$$

Applying Markov's inequality and the Fluctuation Assumption lets us write:

$$\begin{aligned}\mathbb{P}\left(d(\hat{\theta}_n, \theta^*) \geq C\delta_n\right) &\leq \sum_{j\geq 0} \frac{\mathbb{E}[\omega_n(C2^{j+1}\delta_n)]}{\lambda(C2^j\delta_n)^\beta} + \mathbb{P}\left(d(\hat{\theta}_n, \theta^*) \geq \epsilon\right) \\ &\leq \sum_{j\geq 0} \frac{M}{\lambda\sqrt{n}} \frac{(C2^{j+1}\delta_n)^\alpha}{(C2^j\delta_n)^\beta} + \mathbb{P}\left(d(\hat{\theta}_n, \theta^*) \geq \epsilon\right) \\ &\leq \frac{2^\alpha}{C^{\beta-\alpha}} \sum_{j=0}^{\infty} 2^{j(\alpha-\beta)} \frac{M}{\lambda\sqrt{n}} \delta^{\alpha-\beta} + \mathbb{P}\left(d(\hat{\theta}_n, \theta^*) \geq \epsilon\right)\end{aligned}$$

Recall that $\delta_n = \left(\frac{M}{\lambda\sqrt{n}}\right)^{\frac{1}{\beta-\alpha}}$ so $\delta_n^{\alpha-\beta} = \frac{\lambda\sqrt{n}}{M}$. Therefore, $\frac{M}{\lambda\sqrt{n}}\delta^{\alpha-\beta} = 1$ and we can forget about that term. Summing the geometric series gives the bound:

$$\mathbb{P}\left(d(\hat{\theta}_n, \theta^*) \geq C\delta_n\right) \leq \frac{2^\alpha}{C^{\beta-\alpha}} \frac{1}{1 - 2^{\alpha-\beta}} + \mathbb{P}\left(d(\hat{\theta}_n, \theta^*) \geq \epsilon\right)$$

Since $\hat{\theta}_n$ is consistent for θ^*, $\mathbb{P}(d(\hat{\theta}_n, \theta^*) \geq \epsilon) = o(1)$. Since C is arbitrary, we can choose C large enough to make the first term in this final sum as small as we'd like. Hence, $d(\hat{\theta}_n, \theta^*)/\delta_n$ is uniformly tight and we can conclude $d(\hat{\theta}_n, \theta^*) = \mathcal{O}_P(\delta_n) = \mathcal{O}_P(n^{-\frac{1}{2(\beta-\alpha)}})$. □

Example 1: Very Smooth Losses. Suppose $\Theta \subset \mathbb{R}^d$, and the map $\theta \mapsto \ell(\theta, x)$ is $\mathcal{C}^3$ for every x. Moreover, suppose that in a neighborhood of θ^*, $\theta \mapsto \ell(\theta, x)$ is $M(x)$-Lipschitz with $\mathbb{E}[M(X)^2] < \infty$. We will show that the Growth and Fluctuation Assumptions are satisfied in this setting, so that a consistent empirical-risk minimizer is also $\sqrt{n}$-consistent.

First, we will show that it satisfies the Growth Assumption. Taking a Taylor expansion of $L(\cdot)$ at θ^* gives:

$$L(\theta) = L(\theta^*) + \nabla L(\theta^*)(\theta - \theta^*) + \frac{1}{2}(\theta - \theta^*)^\top \nabla^2 L(\theta^*)(\theta - \theta^*) + \mathcal{O}(||\theta - \theta^*||^3)$$

Since θ^* is a global minimizer of L, $\nabla L(\theta^*) = 0$ and $\nabla^2 L(\theta^*)$ is positive-definite. Therefore, for θ close enough to θ^* we have:

$$L(\theta) \geq L(\theta^*) + \frac{1}{4}\lambda_{\min}(\nabla^2 L(\theta^*)) \cdot ||\theta - \theta^*||^2$$

Therefore, the growth condition is satisfied with $\lambda = \frac{1}{4}\lambda_{\min}(\nabla^2 L(\theta^*))$ and $\beta = 2$.

Next, we will show that the Fluctuation Assumption is satisfied as well. This uses the familiar tools of symmetrization and chaining.

$$\mathbb{E}[\omega_n(\delta)] = \mathbb{E}\left[\sup_{\theta: d(\theta,\theta^*)\leq\delta} \Delta_n(\theta)\right] \leq 2\mathbb{E}\left[\mathbb{E}\left[\sup_{d(\theta,\theta^*)\leq\delta}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i(\ell(\theta, X_i) - \ell(\theta^*, X_i))\right|\middle| X_{1:n}\right]\right]$$

Conditioned on $X_{1:n}$, the quantity $\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i(\ell(\theta, X_i) - \ell(\theta^*, X_i))$ is a $\sqrt{\frac{1}{n}\sum_{i=1}^n M(X_i)^2}$-sub-Gaussian process in the Euclidean norm on the set $\{\theta \in \Theta : d(\theta, \theta^*) \leq \delta\}$ by the Lipschitz assumption. Therefore, by Dudley's Entropy Integral inequality we get:

$$\begin{aligned}
\mathbb{E}[\omega_n(\delta)] &\leq \frac{C}{\sqrt{n}}\mathbb{E}\left[\int_0^{2\delta}\left(\frac{1}{n}\sum_{i=1}^n M(X_i)^2\right)^{1/2}\log N(\delta\mathbb{B}, ||\cdot||_2, \epsilon)d\epsilon\right] \\
&\leq \frac{C}{\sqrt{n}}\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n M(X_i)^2\right)^{1/2}\int_0^{2\delta}\sqrt{d\log\left(1 + \frac{2\delta}{\epsilon}\right)}d\epsilon\right] \\
&\leq C'\delta\sqrt{\frac{d}{n}}\mathbb{E}\left[M(X_i)^2\right]^{1/2}
\end{aligned}$$

Therefore, the Fluctuation Assumption is satisfied with $\alpha = 1$ and $M = C'\sqrt{d}\mathbb{E}[M(X_i)^2]^{1/2}$. ♣

2 High-Dimensional Statistics

Changing gears away from rates of convergence, we are now going to discuss high-dimensional statistics. In many modern problems, the problem dimension d scales with the sample size n, and we may even have $d \gg n$. As a result, many of the results of classical asymptotic theory do not apply. What can we do in such settings?

Here are a few motivating examples of high-dimensional problems:

- **Genome-Wise Association Study (GWAS).** In GWAS, biologists try to discover which genes are associated with certain phenotypes. In this problem, we often have $d =$ millions (one parameter to estimate for each gene) but $n =$ hundreds, since obtaining a sample requires sequencing an individual's genome.
- **Signal Reconstruction.** Suppose there is a discrete signal $x \in \mathbb{R}^d$, and engineers collect n Fourier measurements about x, i.e. engineers observe $y = \mathbf{A}x$ for an $n \times p$ matrix $\mathbf{A}$. If n is much smaller than d, is it possible to reconstruct the complete signal x? It turns out that if x is a sparse linear combination of only a few pure frequencies, it is often possible to perfectly recover x from only a few measurements.

In order to make any progress on high-dimensional problems, we lean heavily on two tools:

- **Concentration Inequalities.** If the problem dimension d is allowed to grow with n, it often does not make sense to talk about the "asymptotic regime" unless the scaling relation of d and n is well-specified. Instead, high-dimensional statistics often aims to obtain non-asymptotic/finite-sample error guarantees, and concentration inequalities are a useful tool to prove such results.
- **Growth conditions/Identifiability.** To obtain meaningful results in high-dimensional problems, it is often necessary to impose additional "structural" constraints on the parameters being estimated in order to reduce the "effective dimension" of the problem. One canonical example is to assume that among the d parameters to be estimated, only $k \ll d$ are nonzero.

Next, we will study a prototypical high-dimensional problem to illustrate how concentration inequalities and growth/identifiability assumptions can tame a seemingly intractable problem.

2.1 Gaussian Sequence Model

The **Gaussian Sequence Model** studies the following problem: given an observation $Y = \theta + \sigma\epsilon$ for $\theta \in \mathbb{R}^n$ and $\epsilon \sim \mathcal{N}(0, \mathbf{I}_{n\times n})$, how can we estimate θ? This is certainly a high-dimensional problem, since the number of parameters to estimate d is the same as the sample size n.

A natural first idea is to use $\hat{\theta} = Y$; after all, this estimator is the MLE, it is the minimum-variance unbiased estimator, and it is minimax optimal over the parameter space $\Theta = \mathbb{R}^n$. Unfortunately, it has pretty terrible risk: $\mathbb{E}[||\hat{\theta} - \theta||^2] = n\sigma^2$. Can we do better if we make the structural assumption that only k of the coordinates of θ are nonzero? Certainly it is possible for an "oracle" with knowledge of the underlying support $S \subset [d]$ to achieve better risk, using the estimator $\hat{\theta}^o$ defined by:

$$\hat{\theta}^0_j = \begin{cases} Y_j & \text{if } j \in S \\ 0 & \text{if } j \notin S \end{cases}$$

This estimator achieves risk $\mathbb{E}[||\hat{\theta}^0 - \theta||^2] = k\sigma^2$, a substantial improvement if $k \ll d$. Obviously, this estimator cannot actually be implemented since, unlike the oracle, we do not have access to the true support S.

However, it turns out we can do *almost* as well by guessing the support of θ and mimicking the oracle on the estimated support $\hat{S}$. A natural way to construct $\hat{S}$ is to guess that if $|Y_j| < \tau$ for some threshold τ, then $j \notin S$ but if $|Y_j| \geq \tau$, then $j \in S$. This gives rise to the **hard-thresholding estimator**:

$$\hat{\theta}_j^\tau = \begin{cases} Y_j & \text{if } |Y_j| \geq \tau \\ \\ 0 & \text{if } |Y_j| < \tau \end{cases}$$

The following theorem, which will be proved next time, asserts that the risk of the hard-thresholding estimator (for an optimal choice of τ) comes within a logarithmic factor of the risk of the oracle estimator.

Theorem 2. . *Let* $\tau^2 = 4\sigma^2 \log(n/k)$. *Then* $\mathbb{E}[||\hat{\theta}^\tau - \theta||^2] \leq Ck\sigma^2 \log(n/k)$ *for* C *a universal constant.*