MATH 281C: Mathematical Statistics

Lecture 9

Let us first recall our main bound on the expected suprema of empirical processes. If $\mathcal{F}$ is a class of real-valued functions on $\mathcal{X}$ with envelope F, then (assuming that $PF^2 < \infty$), we have

$$\mathbb{E} \sup_{f \in \mathcal{F}} |P_n f - Pf| \le \frac{C}{\sqrt{n}} \|F\|_{L^2(P)} J(F, \mathcal{F}), \tag{9.1}$$

where

$$J(F, \mathcal{F}) = \int_0^1 \sqrt{1 + \log \sup_Q M(\epsilon \|F\|_{L^2(Q)}, \mathcal{F}, L^2(Q))}\, \mathrm{d}\epsilon.$$

An important implication of this bound is that when $\mathcal{F}$ is a Boolean function class with finite VC dimension D,

$$\mathbb{E} \sup_{f \in \mathcal{F}} |P_n f - Pf| \le C \sqrt{\frac{D}{n}}. \tag{9.2}$$

In this lecture, we will discuss an application of the bounds (9.1) and (9.2) to a problem in M-estimation. Our treatment here will be a mix of rigor and heuristics. We will come back to M-estimation in the next lecture where all the heuristic arguments will be rigorized.

1 Application to an M-estimation problem

Consider a mode estimation problem. Suppose $X_1, \dots, X_n$ are iid observations from a univariate density p. Assume that p has a single mode θ_0, and that it is symmetric around $\theta_0 \in \mathbb{R}$. In addition, assume that p is smooth and bounded, and that $p'(x) < 0$ for $x > \theta_0$ and $p'(x) > 0$ for $x < \theta_0$. You can think of p as the normal density with mean θ_0, or the Cauchy density centered at θ_0.

Consider now the problem of estimating θ_0. For this, define

$$M(\theta) = \int_{\theta-1}^{\theta+1} p(x)\mathrm{d}x = \mathbb{P}(|X_1 - \theta| \le 1) = PI_{\{\theta - 1 \le X_1 \le \theta + 1\}}. \tag{9.3}$$

Note that

$$M'(\theta) = p(\theta + 1) - p(\theta - 1) \quad \text{and} \quad M''(\theta) = p'(\theta + 1) - p'(\theta - 1).$$

Because of the assumptions on p, it is clear that $M'(\theta_0) = 0$, and for $\theta \neq \theta_0$, we have $M'(\theta) < 0$ for $\theta > \theta_0$ and $M'(\theta) > 0$ for $\theta < \theta_0$. This implies that $\theta \mapsto M(\theta)$ has a unique maximum at θ_0. Also, $M''(\theta_0) = p'(\theta_0 + 1) - p'(\theta_0 - 1) < 0$.

Because θ_0 uniquely maximizes $M(\theta)$ over $\theta \in \mathbb{R}$, a reasonable method of estimating θ_0 is to estimate it by $\widehat{\theta}_n$, where $\widehat{\theta}_n$ is any maximizer of $M_n(\theta)$ over $\theta \in \mathbb{R}$ with

$$M_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} I\{X_i \in [\theta - 1, \theta + 1]\}.$$

It is now natural to ask the following questions:

1. Is $\widehat{\theta}_n$ consistent as an estimator for θ_0, i.e., is it true that $|\widehat{\theta}_n - \theta_0|$ converges in probability to zero?

2. Assuming consistency, what is the rate of convergence r_n of $|\widehat{\theta}_n - \theta_0|$ to zero?

3. What is the limiting distribution of $r_n(\widehat{\theta}_n - \theta_0)$?

Below we will prove consistency of $\widehat{\theta}_n$ rigorously. We will also provide a heuristic argument for finding the rate r_n, which we will make rigorous in the next lecture. The limiting distribution will be addressed in a few weeks after the discussion on uniform central limit theorems.

The fundamental first step for analyzing M-estimators is the following inequality:

$$0 \le M(\theta_0) - M(\widehat{\theta}_n) \le M(\theta_0) - M_n(\theta_0) - M(\widehat{\theta}_n) + M_n(\widehat{\theta}_n)$$

We can rewrite the above inequality in Empirical Process notation. For $\theta \in \mathbb{R}$, define the function $m_\theta : \mathbb{R} \to \mathbb{R}$ by

$$m_\theta(x) = I\{\theta - 1 \le x \le \theta + 1\}.$$

With this notation, the inequality becomes

$$0 \le P(m_{\theta_0} - m_{\widehat{\theta}_n}) \le (P_n - P)(m_{\widehat{\theta}_n} - m_{\theta_0}). \tag{9.4}$$

We refer to this inequality as the *basic inequality*.

To derive the consistency of $\widehat{\theta}_n$ from (9.4), we can crudely bound the RHS of (9.4) as

$$(P_n - P)(m_{\widehat{\theta}_n} - m_{\theta_0}) \le 2\sup_{\theta\in\mathbb{R}} |P_n m_\theta - P m_\theta|.$$

It is easy to check that $\{m_\theta : \theta \in \mathbb{R}\}$ is a Boolean class of functions with VC dimension 2. By inequality (9.2),

$$\mathbb{E}\sup_{\theta\in\mathbb{R}} |P_n m_\theta - P m_\theta| \lesssim n^{-1/2} \quad \text{and hence} \quad \sup_{\theta\in\mathbb{R}} |P_n m_\theta - P m_\theta| \xrightarrow{\mathbb{P}} 0.$$

Combining this with (9.4), we obtain

$$M(\theta_0) - M(\widehat{\theta}_n) = P(m_{\theta_0} - m_{\widehat{\theta}_n}) \xrightarrow{\mathbb{P}} 0. \tag{9.5}$$

By the assumptions on p, the following is true: for every $\epsilon > 0$,

$$M(\theta_0 \pm \epsilon) = \sup_{\theta\in\mathbb{R}:|\theta-\theta_0|\ge\epsilon} M(\theta) < M(\theta_0). \tag{9.6}$$

Together, (9.5) and (9.6) imply that $|\widehat{\theta}_n - \theta_0| \xrightarrow{\mathbb{P}} 0$. To see this, first fix $\epsilon > 0$ and use (9.6) to obtain an $\eta > 0$ such that

$$\sup_{\theta \in \mathbb{R}: |\theta - \theta_0| \geq \epsilon} M(\theta) < M(\theta_0) - \eta.$$

It follows then that

$$\mathbb{P}(|\widehat{\theta}_n - \theta_0| \geq \epsilon) \leq \mathbb{P}\{M(\widehat{\theta}_n) < M(\theta_0) - \eta\} = \mathbb{P}\{M(\theta_0) - M(\widehat{\theta}_n) > \eta\} \to 0,$$

where the last (converging to zero) assertion follows from (9.5). This proves $|\widehat{\theta}_n - \theta_0| \xrightarrow{\mathbb{P}} 0$.

This argument for proving consistency of an M-estimator is quite general, and can be isolated in the following theorem; see Theorem 5.7 in van der Vaart (1998).

Theorem 1.1 (Consistency)**.** Let $\{M_n\}$ be a sequence of random functions of $\theta \in \Theta$, and let M be a fixed deterministic function of $\theta \in \Theta$. Let $\widehat{\theta}_n$ be any maximizer of $\{M_n(\theta), \theta \in \Theta\}$, and let θ_0 be the unique maximizer $\theta \mapsto M(\theta)$. Suppose the following two conditions hold

1. $\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \xrightarrow{\mathbb{P}} 0$.
2. For every $\epsilon > 0$, $\sup_{\theta \in \Theta: d(\theta, \theta_0) \geq \epsilon} M(\theta) < M(\theta_0)$. Here d is a metric on Θ.

Then $d(\widehat{\theta}_n, \theta_0) \xrightarrow{\mathbb{P}} 0$ as $n \to \infty$.

The assumption $\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \xrightarrow{\mathbb{P}} 0$ is often too strong for consistency (and also not always easy to check), but there exist results under weaker conditions.

Now that the consistency of $\widehat{\theta}_n$ is established, the next natural question is about the rate of convergence. We can first try to go over the consistency argument again to see it gives an explicit rate of convergence for $|\widehat{\theta}_n - \theta_0|$. We will first argue heuristically.

The consistency argument given above was based on the inequality:

$$0 \leq M(\theta_0) - M(\widehat{\theta}_n) \leq (P_n - P)(m_{\widehat{\theta}_n} - m_{\theta_0}) \leq 2 \sup_{\theta \in \mathbb{R}} |P_n m_\theta - P m_\theta|. \tag{9.7}$$

For consistency, we used that the RHS above converges in probability to zero. But inequality (9.2) actually implies that

$$\mathbb{E} \sup_{\theta \in \mathbb{R}} |P_n m_\theta - P m_\theta| \leq C n^{-1/2},$$

which gives

$$\sup_{\theta \in \mathbb{R}} |P_n m_\theta - P m_\theta| = O_{\mathbb{P}}(n^{-1/2}).$$

Inequality (9.7) then implies

$$0 \leq M(\theta_0) - M(\widehat{\theta}_n) = O_{\mathbb{P}}(n^{-1/2}). \tag{9.8}$$

From here, to obtain an explicit rate for $|\widehat{\theta}_n - \theta_0|$, we need to use some structure of the function $M(\cdot)$. Note that the the second derivative of M at θ_0 equals

$$M''(\theta_0) = p'(\theta_0 + 1) - p'(\theta_0 - 1),$$

which is strictly negative. As a result, there exists a constant C and a neighborhood of θ_0 such that for all θ in that neighborhood, we can write

$$M(\theta_0) - M(\theta) \geq C(\theta - \theta_0)^2. \tag{9.9}$$

The constant C is related to $M''(\theta_0)$. Using this, we can heuristically write

$$M(\theta_0) - M(\widehat{\theta}_n) \geq C(\widehat{\theta}_n - \theta_0)^2. \tag{9.10}$$

In other words, we are assuming that $\widehat{\theta}_n$ belongs to the neighborhood of θ_0 where (9.9) holds. Because of the consistency of $\widehat{\theta}_n$ (which we have rigorously proved), inequality (9.10) can be made rigorous. Combining (9.10) with (9.8), we deduce that

$$(\widehat{\theta}_n - \theta_0)^2 = O_{\mathbb{P}}(n^{-1/2}),$$

and hence

$$|\widehat{\theta}_n - \theta_0| = O_{\mathbb{P}}(n^{-1/4}).$$

We have therefore obtained $n^{-1/4}$ as a rate of convergence for $|\widehat{\theta}_n - \theta_0|$. It turns out however that

$$|\widehat{\theta}_n - \theta_0| = O_{\mathbb{P}}(n^{-1/3}) \quad \text{(optimal rate of convergence)}$$

In other words, $n^{-1/4}$ is slower than the actual rate of convergence and is a refection of some looseness in our proof technique. The main source of looseness is in the inequality:

$$(P_n - P)(m_{\widehat{\theta}_n} - m_{\theta_0}) \leq 2 \sup_{\theta \in \mathbb{R}} |P_n m_\theta - P m_\theta| = O_{\mathbb{P}}(n^{-1/2}). \tag{9.11}$$

It turns out that the LHS above is much smaller than the RHS. To get a heuristic understanding of the size of the LHS above, let us first compute bounds for

$$\mathbb{E}|(P_n - P)(m_\theta - m_{\theta_0})|$$

for a fixed θ that is close to θ_0. Clearly,

$$\begin{aligned}\mathbb{E}|(P_n - P)(m_\theta - m_{\theta_0})| &\leq \sqrt{\mathrm{Var}(P_n(m_\theta - m_{\theta_0})} \\ &= \frac{1}{\sqrt{n}}\sqrt{\mathrm{var}(m_\theta(X_1) - m_{\theta_0}(X_1))} \leq \frac{1}{\sqrt{n}}\sqrt{\mathbb{E}\{m_\theta(X_1) - m_{\theta_0}(X_1)\}^2}.\end{aligned} \tag{9.12}$$

Now for θ close to θ_0 (and $\theta < \theta_0$), we have

$$m_\theta(x) - m_{\theta_0}(x) = I(\theta - 1 \leq x \leq \theta_0 - 1) - I(\theta + 1 \leq x \leq \theta_0 + 1).$$

Thus

$$\mathbb{E}\{m_\theta(X_1) - m_{\theta_0}(X_1)\}^2 \leq P(\theta - 1 \leq X_1 \leq \theta_0 - 1) + \mathbb{P}(\theta + 1 \leq X_1 \leq \theta_0 + 1) \leq 2p(\theta_0)|\theta - \theta_0|.$$

Combining this with (9.12), we obtain

$$\mathbb{E}|(P_n - P)(m_\theta - m_{\theta_0})| \leq \sqrt{\frac{2p(\theta_0)}{n}}|\theta - \theta_0|^{1/2},$$

so that

$$(P_n - P)(m_\theta - m_{\theta_0}) = O_{\mathbb{P}}(|\theta - \theta_0|^{1/2} n^{-1/2}).$$

This is true for a fixed θ that is close to θ_0. Heuristically, this suggests that

$$(P_n - P)(m_{\widehat{\theta}_n} - m_{\theta_0}) = O_{\mathbb{P}}(|\widehat{\theta}_n - \theta_0|^{1/2} n^{-1/2})$$

We shall formally justify this later. Note that this bound is an stronger compared to our earlier bound (9.11). Plugging this in the RHS of the basic inequality (9.4), and using the quadratic bound (9.10) on the LHS of (9.4), we deduce that

$$C(\widehat{\theta}_n - \theta_0)^2 \le O_{\mathbb{P}}(|\widehat{\theta}_n - \theta_0|^{1/2} n^{-1/2}).$$

Canceling $|\widehat{\theta}_n - \theta_0|^{1/2}$ from both sides, we deduce that

$$|\widehat{\theta}_n - \theta_0|^{1/2} = O_{\mathbb{P}}(n^{-1/3}). \tag{9.13}$$

As mentioned earlier, this is the correct rate for $\widehat{\theta}_n - \theta_0$. Indeed, it turns out that

$$n^{1/2}(\widehat{\theta}_n - \theta_0) \xrightarrow{\mathrm{d}} \operatorname*{argmax}_{h \in \mathbb{R}}(aZ_h - bh^2)$$

as $n \to \infty$, where $Z_h, h \in \mathbb{R}$ is a two-sided Brownian motion starting from 0, and a, b are two constants depending on p and θ_0. We will prove this limiting result later but it tells us now that $n^{-1/3}$ is the correct rate of convergence.

We will make the rate result (9.13) rigorous in the next lecture. A key ingredient in the rigorous argument will involve establishing the inequality

$$\mathbb{E} \sup_{\theta:|\theta-\theta_0|\le\delta} |(P_n - P)(m_\theta - m_{\theta_0})| \le C\sqrt{\frac{\delta}{n}} \tag{9.14}$$

for δ sufficiently small. The above inequality can be derived as a consequence of (9.1). Indeed, to apply (9.1), we first need to obtain an envelope for the class $\{m_\theta - m_{\theta_0} : \theta \in [\theta_0 - \delta, \theta_0 + \delta]\}$. It is not hard to see that

$$F(x) := I_{[\theta_0-1-\delta,\, \theta_0-1+\delta]}(x) + I_{[\theta_0+1-\delta,\, \theta_0+1+\delta]}(x)$$

is an envelope function. Furthermore,

$$PF^2 \le P(\theta_0 - 1 - \delta \le X_i \le \theta_0 - 1 + \delta) + \mathbb{P}(\theta_0 + 1 - \delta \le X_1 \le \theta_0 + 1 + \delta) \le 4p(\theta_0)\delta.$$

Thus, (9.1) implies

$$\mathbb{E} \sup_{\theta:|\theta-\theta_0|\le\delta} |(P_n - P)(m_\theta - m_{\theta_0})| \le C\sqrt{\frac{\delta}{n}} J(F, \mathcal{F}),$$

where $\mathcal{F} = \{m_\theta - m_{\theta_0} : |\theta - \theta_0| \le \delta\}$. This will prove (9.14) provided $J(F, \mathcal{F}) < \infty$. This will follow from the fact that the class $\mathcal{F}$ has finite VC *subgraph dimension* (to be defined shortly). Note that this is not a Boolean class of functions so we need VC subgraph dimension as opposed to VC dimension.

Let us now summarize this discussion of a heuristic argument for the rate of M-estimators. Although we did for a special M-estimator which corresponded to $m_\theta := I(\theta - 1 \leq x \leq \theta + 1)$, the ideas are fairly general. The most important ingredient is the basic inequality (9.4). The LHS $P(m_{\theta_0} - m_{\widehat{\theta}_n})$ is bounded from below by an assumption on the second derivative of $\theta \mapsto Pm_\theta$ at $\theta = \theta_0$. The RHS can be understood by calculating

$$\mathbb{E}\{m_\theta(X_1) - m_{\theta_0}(X_1)\}^2.$$

For the specific choice of $m_\theta(x) = I(\theta - 1 \leq x \leq \theta + 1)$, it turned out

$$\mathbb{E}\{m_\theta(X_1) - m_{\theta_0}(X_1)\}^2 \leq C|\theta - \theta_0|.$$

For other m_θ, the RHS might be different (for example, it is common to have $C(\theta - \theta_0)^2$ on the RHS). Plugging these bounds in the basic inequality will yield an inequality involving $|\widehat{\theta}_n - \theta_0|$ and n, which can be solved to get explicit rates (such as $n^{-1/3}$) in this problem. This heuristic will be justified next time. Before concluding this section, let us state a result from van der Vaart and Wellner (1996) on the rate of convergence of M-estimators. We will formally prove this result later but, based on the above heuristics, its conclusion should be quite obvious.

Theorem 1.2 (Page 294 of van der Vaart and Wellner (1996))**.** Let $X_1, \dots, X_n$ be iid observations from a distribution P. Suppose $\Theta \subseteq \mathbb{R}$ is an open set and let $m_\theta, \theta \in \Theta$ be a collection of real-valued functions on $\mathcal{X}$ that are indexed by Θ. Suppose there exist $\alpha > 0$ and a function M on $\mathcal{X}$ with $PM^2 < \infty$ for which

$$|m_{\theta_1}(x) - m_{\theta_2}(x)| \leq M(x)|\theta_1 - \theta_2|^\alpha \quad \text{for all } \theta_1, \theta_2 \in \Theta. \tag{9.15}$$

Let $\widehat{\theta}_n$ and θ_0 denote maximizers of $P_n m_\theta$ and Pm_θ over $\theta \in \Theta$. If $\theta \mapsto Pm_\theta$ has two derivatives at θ_0 with the second derivative strictly negative, then

$$|\widehat{\theta}_n - \theta_0| = O_{\mathbb{P}}(n^{-1/(4-2\alpha)}). \tag{9.16}$$

Heuristic Argument. We use the heuristics summarized above to justify (9.16) based on the assumptions made in the theorem. Note that assumption (9.15) implies that

$$\mathbb{E}\{m_\theta(X_1) - m_{\theta_0}(X_1)\}^2 \leq \mathbb{E}\{M^2(X_1)|\theta - \theta_0|^{2\alpha}\} = |\theta - \theta_0|^{2\alpha} PM^2.$$

This suggests the heuristic

$$(P_n - P)(m_{\widehat{\theta}_n} - m_{\theta_0}) \lesssim_{\mathbb{P}} n^{-1/2}|\widehat{\theta}_n - \theta_0|^\alpha.$$

Combining with the basic inequality and the lower bound $C(\widehat{\theta}_n - \theta_0)^2$ on $P(m_{\theta_0} - m_{\widehat{\theta}_n})$, we obtain

$$(\widehat{\theta}_n - \theta_0)^2 \lesssim_{\mathbb{P}} n^{-1/2}|\widehat{\theta}_n - \theta_0|^\alpha,$$

which yields

$$|\widehat{\theta}_n - \theta_0| \lesssim_{\mathbb{P}} n^{-1/(4-2\alpha)}$$

When $\alpha = 1$, Theorem 1.2 gives the usual $n^{-1/2}$ rate for $|\widehat{\theta}_n - \theta_0|$.

References

van der Vaart, A. (1998). *Asymptotic Statistics.* Cambridge University Press.

van der Vaart, A. and Wellner, J. A. (1996). *Weak Convergence and Empirical Process: With Applications to Statistics.* Springer-Verlag.