6.S951 Modern Mathematical Statistics Fall 2024

Lecture 22 — December 3, 2024

Prof. Stephen Bates

Scribe: Oswin So

1 Outline

Agenda:

1. M-Estimators: Overview, Consistency

Last time:

1. Moment estimates
2. Exponential family
3. Asymptotic normality of MLE in exponential family

2 M-Estimators

M-estimators are estimators defined as solutions to optimization problems.

Setting: $X_i \in \mathcal{X}$, $i = 1, \ldots, n$ i.i.d. from P.

Definition 1 (M-estimator)**.** *Given a function $M_\theta(x) : \Theta \times \mathcal{X} \to \bar{\mathbb{R}}$, we define $M_n : \Theta \to \bar{\mathbb{R}}$ as*

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

Then, the M-estimator $\hat{\theta}$ is defined as the minimizer of $M_n(\theta)$, i.e.,

$$\hat{\theta} = \arg\min_{\theta \in \Theta} M_n(\theta).$$

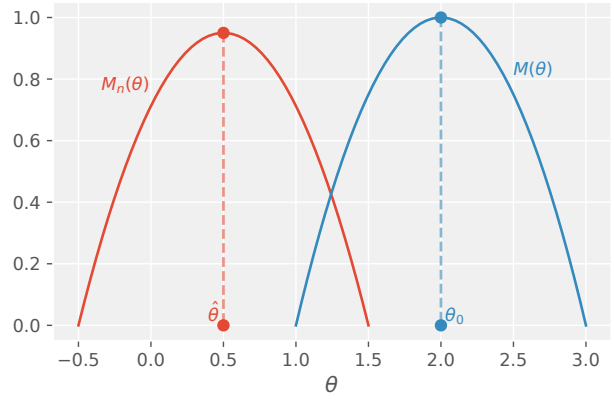
We additionally define θ_0 as

$$\theta_0 = \arg\max_{\theta \in \Theta} \mathbb{E}_{X \sim P}\big[M_\theta(X)\big].$$

We will show the following two properties about M-estimators:

1. Consistency: $\hat{\theta} \xrightarrow{p} \theta_0$.
2. Asymptotic normality: $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$.

Visualization: We visualize the setting below.



Notice that $M_n(\theta) \xrightarrow{p} M(\theta)$ for each θ by the law of large numbers. As n grows, M_n gets closer to M. The question we are interested in is: does the maximizer of M_n get close to the maximizer of M? The answer is yes, under regularity conditions.

Example 1 (Maximum Likelihood)**.** *The maximum likelihood estimator is an M-estimator. Let p_θ denote the density of X under parameter θ. We then define M_θ as*

$$M_\theta(x) = \log p_\theta(x).$$

Example 2 (Quantiles)**.** *Quantile estimation is an M-estimator by using the pinball loss function. For quantile $\tau \in (0, 1)$, we define M_θ as*

$$M_\theta(x) = -(\theta - x)(\tau - \mathbf{1}_{\{(\theta-x)>0\}}).$$

Example 3 (Least Squares)**.** *We can pose the least squares estimator as an M-estimator.*

$$M_\theta((x, y)) = (y - x^\intercal\theta)^2$$

Example 4 (Ridge Regression)**.** *We can also add regularization to the least squares estimator to get the ridge regression estimator.*

$$M_\theta((x, y)) = (y - x^\intercal\theta)^2 - \lambda\|\theta\|^2$$

Example 5 (Median Regression)**.**

$$M_\theta((x, y)) = -|y - x^\intercal\theta|$$

M_n in this case is piecewise linear.

3 Consistency of M-Estimators

We wish to show that $\hat{\theta} \xrightarrow{p} \theta_0$. Recall that $M_n(\theta) \xrightarrow{p} M(\theta)$. This is a good start, but is not sufficient to prove what we want. We will need two "strengthenings" to prove consistency.

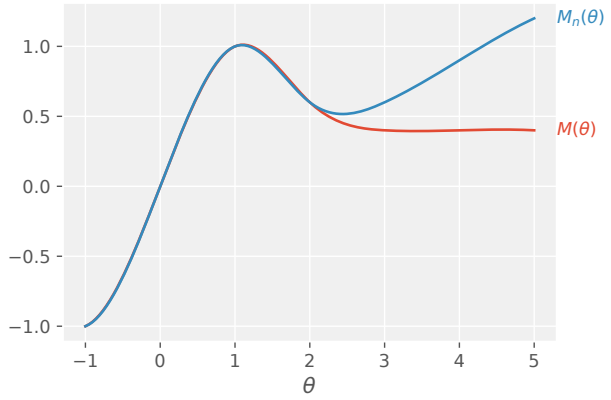
Theorem 2. *Suppose*

1. *Uniform convergence:* $\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \xrightarrow{p} 0$.
2. *Separation: For all* $\epsilon > 0$,

$$\sup_{\theta : \|\theta - \theta_0\| > \epsilon} M(\theta) < M(\theta_0).$$

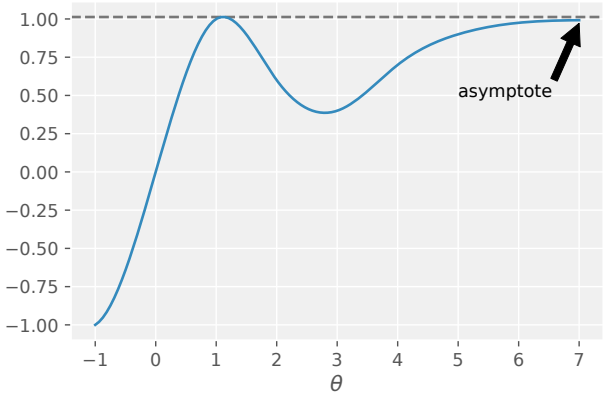
Remark 1. *Consistency still holds for* $\hat{\theta}$ *that only approximately minimizes* M_n.

Role of uniform convergence: It is possible that M_n converges to M pointwise, but the maximizer of M_n does not converge to the maximizer of M, as in the example below.



We enforce uniform convergence to ensure that this does not happen.

Role of separation: Without separation, it is possible that statistical noise to M_n results in a maximizer $\hat{\theta}$ that is very far away from the true maximizer θ_0 even if M_n is close to M as in the example below.



Proof. By definition of $\hat{\theta}$,

$$M_n(\hat{\theta}) \geq M_n(\theta_0) \xrightarrow{p} M(\theta_0).$$

Using the notation $o_p(1)$ to denote a sequence that converges to 0 in probability, we then have

$$M_n(\hat{\theta}) \geq M(\theta_0) - o_p(1).$$

Rearranging, subtracting $M(\hat{\theta})$ on both sides then using uniform convergence, we get

$$\begin{aligned} M(\theta_0) - M(\hat{\theta}) &\leq M_n(\hat{\theta}) - M(\hat{\theta}) + o_p(1) \\ &\leq \sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| + o_p(1) \\ &\xrightarrow{p} 0. \end{aligned}$$

We have thus proven that $M(\hat{\theta}) \xrightarrow{p} M(\theta_0)$.

Next, by separation (ii), we obtain that $\hat{\theta} \xrightarrow{p} \theta_0$. No other maximizer can be close to θ_0 by separation. □

All the work in the proof is done by assuming uniform convergence. When do we get uniform convergence? See IDS 160. Some sufficient conditions for uniform convergence include

- Finite VC dimension
- Finite Rademachar or Gaussian complexity

We also have a more "classical" condition using compactness and continuity in the following theorem.

Theorem 3. *Suppose Θ is compact, M_θ and M are continuous in θ, and*

$$\mathbb{E}[\sup_{\theta \in \Theta} M_\theta(X)] < \infty$$

Then,

$$\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \xrightarrow{p} 0.$$

Proof. See VDV. □

4 Asymptotic Normality in the Exponential Family

Returning to the setting of last lecture (moment estimators and exponential family):

Theorem 4. *Let $X_i \overset{i.i.d.}{\sim} P$ for any data generating distribution P.*

Let $\hat{\theta}$ be the MLE in the exponential family. Then,

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \Gamma),$$

where

$$\theta_0 = \arg\max_{\theta \in \Theta} \mathbb{E}\big[l_\theta(X_i)\big], \qquad \Gamma = {e'_{\theta_0}}^{-1} \Big[\operatorname{Cov}_P \, t(X_i) \Big] \Big({e'_{\theta_0}}^{-1} \Big)^\intercal$$

This is true even if P is not in the model, i.e.,

$$P \neq P_\theta \quad \text{for any } \theta \in \Theta.$$

Moreover, this behavior happens for many M-estimators, not just moment estimators.

Remark 2. *If $P = P_\theta$ for some $\theta \in \Theta$ (model is well specified), then (to be proven in the HW)*

$$e'_{\theta_0} = \operatorname{Cov}_P t(X_i).$$

This results in cancellations in Γ and simplifies the expression to the inverse Fisher information:

$$\begin{aligned}\Gamma &= \Big(\operatorname{Cov}_P t(X_i) \Big)^{-1}, \\ &= \Big(\mathbb{E}\big[\dot{l}(X_i)\dot{l}(X_i)^\intercal\big] \Big)^{-1},\end{aligned}$$

and we have used $\dot{l}$ to denote the derivative.

The same story holds for M-estimators, but with the covariance term modified.