Introduction to Empirical Processes and Semiparametric Inference Lecture 19: M-estimators

Yair Goldberg, Ph.D.,

and Michael R. Kosorok, Ph.D.

University of North Carolina-Chapel Hill

M-Estimators

M-estimators are (approximate) maximizers (or minimizers) $\hat{\theta}_n$ of objective functions $\theta \mapsto M_n(\theta)$.

Examples include:

- maximum likelihood estimators
- least squares estimators
- least absolute deviation estimators

Usually the objective funciton $\theta \mapsto M_n(\theta)$ is an empirical (data generated) process while $\theta \mapsto M(\theta)$ is a limiting process of some kind.

Often,

$$\theta \mapsto M_n(\theta) = \mathbb{P}_n m_\theta(X),$$

where $\{m_\theta(X) : \theta \in \Theta\}$ is a class of measurable functions $X \mapsto m_\theta(X)$ on the sample space $\mathcal{X}$.

The Argmax theorem studies the limiting distribution of M-estimators through the limiting behavior of the associated objective functions.

The Argmax Theorem

Let M_n, M be stochastic processes indexed by a metric space H.

Assume

(A) The sample paths $h \mapsto M(h)$ are upper semicontinuous and possess a unique maximum at a (random) point $\hat{h}$, which as a random map in H is tight.

(B) $M_n \rightsquigarrow M$ in $\ell^\infty(K)$ for every compact $K \subset H$.

(C) The sequence $\hat{h}_n$ is uniformly tight and satisfies $M_n(\hat{h}_n) \geq \sup_{h \in H} M_n(h) - o_P(1)$

then $\hat{h}_n \rightsquigarrow \hat{h}$ in H.

A sequence X_n is asymptotically tight if for every $\epsilon > 0$, there is a compact set K such that $\liminf P_*(X_n \in K^\delta) > 1 - \epsilon$ for every $\delta > 0$, where $K^\delta = \{x : d(x, K) < \delta\}$.

A sequence X_n is uniformly tight if for every $\epsilon > 0$, there is a compact set K such that $P(X_n \in K) > 1 - \epsilon$.

In $\mathbb{R}^p$, X_n is asymptotically tight iff X_n is uniformly tight.

Rate of Convergence

Let $\theta \mapsto M(\theta)$ be twice differentiable at a point of unique maximum θ_0.

Then $\frac{\partial}{\partial\theta}M(\theta_0) \equiv 0$.
while $\frac{\partial^2}{\partial\theta^2}M(\theta_0)$ is negative definite.

Hence we can expect that

$$M(\theta) - M(\theta_0) \leq -cd^2(\theta, \theta_0)$$

for some $c > 0$ in a neighborhood of θ_0.

Sometimes we replace the metric function d by a function

$$\tilde{d} : \Theta \times \Theta \mapsto [0, \infty)$$

that satisfies $\tilde{d}(\theta_n, \theta_0) \to 0$ whenever $d(\theta_n, \theta_0) \to 0$.

This is useful, for example, when different parameters of the model have different rates of convergence.

The modulus of continuity of a stochastic process $\{X(t) : t \in T\}$ is defined by

$$m_x(\delta) \equiv \sup_{s,t \in T: d(s,t) \leq \delta} |X(s) - X(t)| .$$

An upper bound for the rate of convergence of an M-estimator can be obtained from the modulus of continuity of $M_n - M$ at θ_0.

Theorem 14.4: Rate of convergence

Let M_n be a sequence of stochastic processes indexed by a semimetric space (Θ, d) and $M : \Theta \mapsto \mathbb{R}$ a deterministic function.
Assume that

(A) For every θ in a neighborhood of θ_0, there exists a $c_1 > 0$ such that

$$M(\theta) - M(\theta_0) \quad \leq \quad -c_1 \tilde{d}^2(\theta, \theta_0),$$

(B) For all n large enough and sufficiently small δ, the centered process $M_n - M$ satisfies

$$E^* \sup_{\tilde{d}(\theta,\theta_0)<\delta} \sqrt{n}\,|(M_n - M)(\theta) - (M_n - M)(\theta_0)| \quad \leq \quad c_2\phi_n(\delta),$$

for $c_2 < \infty$ and functions ϕ_n such that $\delta \mapsto \phi_n(\delta)/\delta^\alpha$ is decreasing for some $\alpha < 2$ not depending on n.

(C) The sequence $\hat{\theta}_n$ converges in outer probability to θ_0, and satisfies

$$M_n(\hat{\theta}_n) \geq \sup_{\theta \in \Theta} M_n(\theta) - O_P(r_n^{-2})$$

for some sequence r_n that satisfies

$$r_n^2 \phi_n(r_n^{-1}) \leq c_3 \sqrt{n}, \quad \text{for every } n \text{ and some } c_3 < \infty \, .$$

Then

$$r_n \tilde{d}(\hat{\theta}_n, \theta_0) = O_P(1) \, .$$

Remark

The “modulus of continuity” of the empirical process gives an upper bound on the rate.

When $\phi(\delta) = \delta^{\alpha}$ then the rate is at least $n^{1/(4-2\alpha)}$.

For $\phi(\delta) = \delta$ we get the $\sqrt{n}$ rate.

Proof

We assume for simplicity that $\hat{\theta}_n$ maximize $M_n(\theta)$ and that $\tilde{d} = d$.

Our goal is to show that $r_n d(\hat{\theta}_n, \theta_0) = O_P(1)$. This is equivalent to showing that for all n large enough $P^*(r_n d(\hat{\theta}_n, \theta_0) > 2^K) < \epsilon$ for some constant K.

For each n, the parameter space (minus the point θ_0) can be partitioned into "peels"

$$S_{j,n} = \{\theta : 2^{j-1} < r_n d(\theta, \theta_0) \leq 2^j\}$$

with j ranging over the integers.

Fix $\eta > 0$ small enough such that

$$\sup_{\theta : d(\theta,\theta_0)<\eta} M(\theta) - M(\theta_0) \quad \leq \quad -c_1 d^2(\theta,\theta_0)\,.$$

and such that for all $\delta < \eta$

$$E^* \sup_{d(\theta,\theta_0)<\delta} \sqrt{n}\,\left|(M_n - M)(\theta) - (M_n - M)(\theta_0)\right| \quad \leq \quad c_2 \phi_n(\delta),$$

Such η exists by assumptions (A) and (B).

Note that if $r_n d(\hat{\theta}_n, \theta_0) > 2^K$ for a given integer K, then $\hat{\theta}_n$ is in one of the peels $S_{j,n}$, with $j > K$.

Thus

$$P^*\left(r_n d(\hat{\theta}_n, \theta_0) > 2^K\right)$$

$$\leq \sum_{j \geq K, 2^j \leq \eta r_n} P^*\left(\sup_{\theta \in S_{j,n}} [M_n(\theta) - M_n(\theta_0)] \geq 0\right)$$

$$+P^*\left(2d(\hat{\theta}_n, \theta_0) \geq \eta\right)$$

By Assumption (A), for every $\theta \in S_{n,j}$, such that $2^j < \eta r_n$,

$$M(\theta) - M(\theta_0) \leq -c_1 d^2(\theta, \theta_0) \leq -c_1 2^{2j-2} r_n^{-2}$$

By Assumption (B), Markov's inequality, and the fact that $\phi_n(c\delta) \leq c^\alpha \phi_n(\delta)$ for every $c > 1$,

$$P^* \left(\sup_{\theta \in S_{j,n}} |(M_n - M)(\theta) - (M_n - M)(\theta_0)| \geq \frac{c_1 2^{2j-2}}{r_n^2} \right)$$

$$\leq \frac{c_2 \phi_n \left(\frac{2^j}{r_n} \right) r_n^2}{\sqrt{n}(c_1 2^{2j-2})} \leq \frac{2 c_2 c_3 2^{j\alpha - 2j + 2}}{c_1} .$$

Summarizing

$$
\begin{aligned}
& P^*\left(r_n d(\hat{\theta}_n, \theta_0) > 2^K\right) \\
\leq & \sum_{j \geq K, 2^j \leq \eta r_n} P^*\left(\sup_{\theta \in S_{j,n}} [M_n(\theta) - M_n(\theta_0)] \geq 0\right) \\
& + P^*\left(2d(\hat{\theta}_n, \theta_0) \geq \eta\right) \\
\leq & \sum_{j > K} \frac{2c_2 c_3 2^{j\alpha - 2j + 2}}{c_1} + P^*\left(2d(\hat{\theta}_n, \theta_0) \geq \eta\right)
\end{aligned}
$$

The first term is smaller than ϵ for all K large enough. The second term is smaller than ϵ for all n large enough since $\hat{\theta}_n$ is consistent. This proves that $r_n d(\hat{\theta}_n, \theta_0) = O_P(1)$

Regular Euclidean M-Estimators

Let $m_\theta : \mathcal{X} \mapsto \mathbb{R}$ where $\theta \in \Theta \subset \mathbb{R}^p$.

Let $M_n(\theta) = \mathbb{P}_n m_\theta$ and $M(\theta) = P m_\theta$.

Theorem 2.13

Assume

(A) θ_0 maximizes $M(\theta)$ and $M(\theta)$ has a non-singular second derivative matrix V.

(B) There exist measurable functions $F_\delta : \mathcal{X} \mapsto \mathbb{R}$ and $\dot{m}_{\theta_0} : \mathcal{X} \mapsto \mathbb{R}^p$ such that

$$\begin{aligned} |m_{\theta_1}(x) - m_{\theta_2}(x)| &\leq F_\delta(x)\|\theta_1 - \theta_2\|, \\ P(m_{\theta_1} - m_{\theta_0} - \dot{m}_{\theta_0}\|\theta_1 - \theta_0\|)^2 &= o(\|\theta_1 - \theta_0\|^2)\,, \end{aligned}$$

and $PF_\delta^2 < \infty$, $P\|m_\theta\|^2 < \infty$ in some neighborhood $\Theta_0 \subset \Theta$ that contains θ_0.

(C) $\hat{\theta}_n \xrightarrow{\mathrm{P}} \theta_0$ and $M_n(\hat{\theta}_n) \geq \sup_{\theta\in\Theta} M_n(\theta) - O_P(n^{-1})$

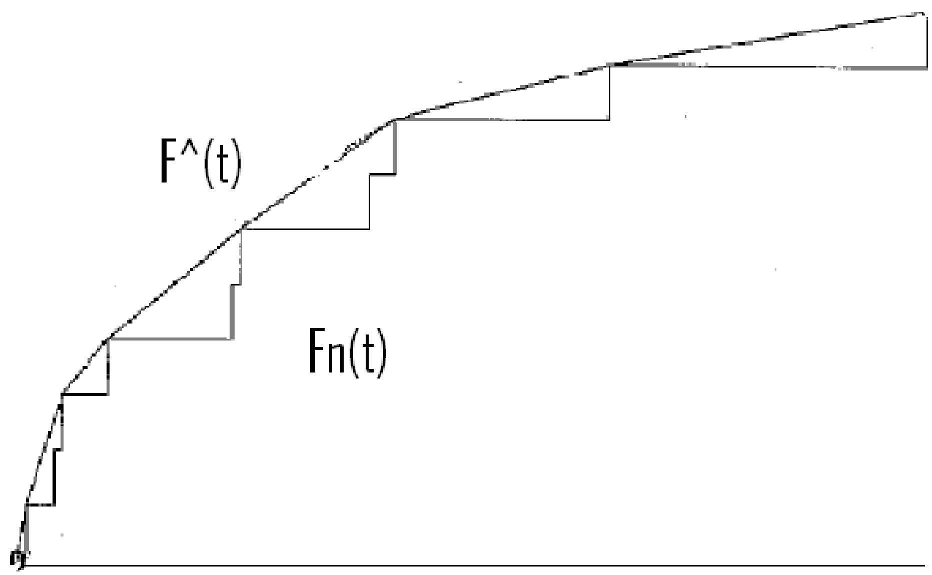
Then $\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow -V^{-1}Z$ where Z is the limiting distribution of $\mathbb{G}_n\dot{m}_{\theta_0}$.

Monotone Density Estimation

Let $X_1, \ldots, X_n$ be a sample of size n from a Lebesgue density f on $[0, \infty)$ that is known to be decreasing. Note that this means that F is concave.

Fix $t > 0$. We assume that f is differentiable at t with derivative $-\infty < f'(t) < 0$.

The maximum likelihood estimator $\hat{f}_n$ of f is the non-increasing step function equal to the left derivative of $\hat{F}_n$, the *least concave majorant* of the empirical distribution function $\mathbb{F}_n$ which is known as the Grenander estimator (Grenander, 1956).


F^(t)
Fn(t)

Consistency

LEMMA 1. ***Marshall's lemma***

$$\sup_{t \geq 0} \left| \hat{F}_n(t) - F(t) \right| \leq \sup_{t \geq 0} \left| \mathbb{F}_n(t) - F(t) \right| .$$

The proof is an exercise.

Fix $0 < \delta < t$. Note that

$$\frac{\hat{F}_n(t+\delta) - \hat{F}_n(t)}{\delta} \leq \hat{f}_n(t) \leq \frac{\hat{F}_n(t) - \hat{F}_n(t-\delta)}{\delta}.$$

By Marshall's lemma,

$$\frac{\hat{F}_n(t+\delta) - \hat{F}_n(t)}{\delta} \quad \overset{\text{as}*}{\to} \quad \frac{F(t+\delta) - F(t)}{\delta}$$

$$\frac{\hat{F}_n(t-\delta) - \hat{F}_n(t)}{\delta} \quad \overset{\text{as}*}{\to} \quad \frac{F(t-\delta) - F(t)}{\delta}$$

By the assumptions on F and the arbitrariness of δ, we obtain $\hat{f}_n(t) \overset{\text{as}*}{\to} f(t)$.

Rate of Convergence

The inverse function representation

Define the stochastic process

$$\hat{s}_n(a) = \arg\max_{s \geq 0}\{\mathbb{F}_n(s) - as\}, \text{ for a} > 0\,.$$

The largest value is selected when multiple maximizers exist.

The function $\hat{s}_n$ is a sort of inverse of the function $\hat{f}_n$ in the sense that $\hat{f}_n(t) \leq a$ if and only if $\hat{s}_n(a) \leq t$ for every $t \geq 0$ and $a > 0$.

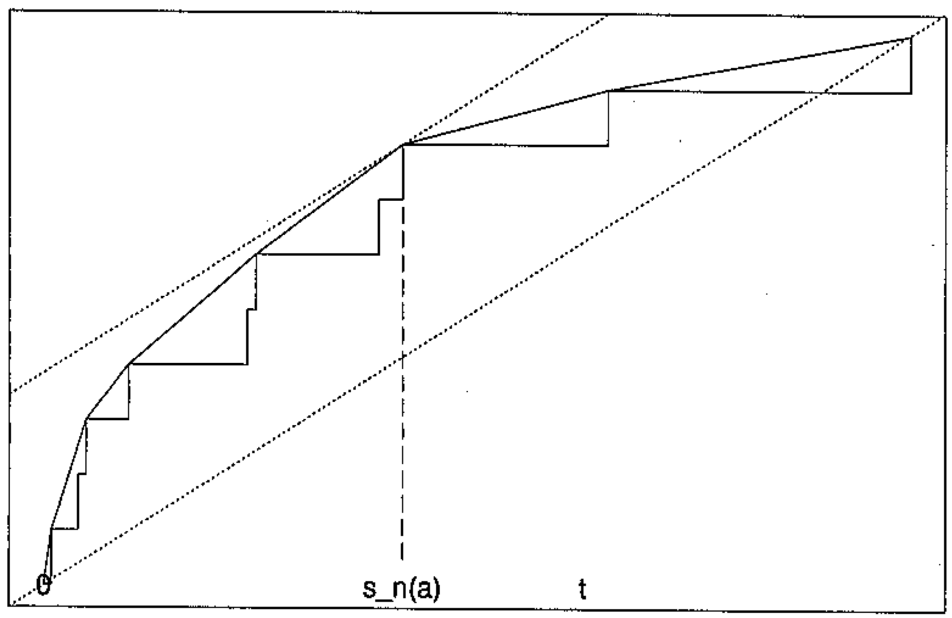


Figure 1: $\hat{s}_n(a) = \arg\max_{s \geq 0}\{\mathbb{F}_n(s) - as\},$ for a > 0.

Define

$$\begin{aligned} M_n(g) &\equiv \mathbb{F}_n(t+g) - \mathbb{F}_n(t) - f(t)g - xgn^{-1/3} \\ M(g) &\equiv F(t+g) - F(t) - f(t)g\,. \end{aligned}$$

By changing variable $s \mapsto t+g$ in the dentition of $\hat{s}_n$ combined with the fact that the location of the maximum of a function does not change when the function is shifted vertically we have

$$\begin{aligned} \hat{s}_n(f(t)+xn^{-1/3}) - t &\equiv \arg\max_{\{g>-t\}} \{\mathbb{F}_n(t+g) \\ &\quad -(f(t)+xn^{-1/3})(t+g)\} \\ &= \arg\max_{\{g>-t\}} M_n(g) \end{aligned}$$

Define $\hat{g}_n = \arg\max_{\{g>-t\}} M_n(g)$.

Our goal is to show that the conditions of Theorem 14.4 hold for $\hat{g}_n$ with rate of $n^{1/3}$ where

$$\theta = g \ , \theta_0 = 0 \ , d(\theta, \theta_0) = |\theta - \theta_0| \ .$$

Note that by the existence of the derivative for f at t we have

$$M(g) = F(t+g) - F(t) - f(t)g = \frac{1}{2} f'(t) g^2 + o(g^2)$$

Since by assumption $f'(t) < 0$, Assumption (A), namely, $M(\theta) - M(\theta_0) \leq -c_1 d^2(\theta, \theta_0)$, holds.

Recall that Assumption (B) of Theorem 14.4 states:

For all n large enough and sufficiently small δ, the centered process $M_n - M$ satisfies

$$E^* \sup_{d(\theta,\theta_0)<\delta} \sqrt{n}\,|(M_n - M)(\theta) - (M_n - M)(\theta_0)| \quad \leq \quad c_2\phi_n(\delta),$$

for $c_2 < \infty$ and functions ϕ_n such that $\delta \mapsto \phi_n(\delta)/\delta^\alpha$ is decreasing for some $\alpha < 2$ not depending on n.

Recall

$$\begin{aligned} M_n(g) &\equiv \mathbb{F}_n(t+g) - \mathbb{F}_n(t) - f(t)g - xgn^{-1/3} \\ M(g) &\equiv F(t+g) - F(t) - f(t)g\,. \end{aligned}$$

and thus $M_n(0) = M(0) = 0$.

Hence

$$
\begin{aligned}
& E^* \sup_{|g|<\delta} \sqrt{n}\,|M_n(g) - M(g)| \\
\leq \quad & E^* \sup_{|g|<\delta} |\mathbb{G}_n(\mathbf{1}\{X \leq t+g\} - \mathbf{1}\{X \leq t\})| \\
& + O(\sqrt{n}\delta n^{-1/3}) \\
\lesssim \quad & \phi_n(\delta) \equiv \delta^{1/2} + \sqrt{n}\delta n^{-1/3}.
\end{aligned}
$$

Clearly

$$
\frac{\phi_n(\delta)}{\delta^\alpha} = \frac{\delta^{1/2} + \sqrt{n}\delta n^{-1/3}}{\delta^\alpha}
$$

is decreasing for $\alpha = 3/2 < 2$.

Assumption (C) of Theorem 14.4:

The sequence $\hat{\theta}_n$ converges in outer probability to θ_0,

and satisfies

$$M_n(\hat{\theta}_n) \geq \sup_{\theta \in \Theta} M_n(\theta) - O_P(r_n^{-2})$$

for some sequence r_n that satisfies

$$r_n^2 \phi_n(r_n^{-1}) \leq c_3 \sqrt{n}, \quad \text{for every } n \text{ and some } c_3 < \infty \, .$$

- $M(g) = F(t+g) - F(t) - f(t)g$ is continuous and has a unique maximum at $g = 0$.
- $M_n(g) \rightsquigarrow M(g)$ uniformly on compacts.
- $M_n(\hat{g}_n) = \sup_g M_n(g)$.

Thus by the argmax theorem $\hat{g}_n \rightsquigarrow 0$.

Choose $r_n = n^{1/3}$. Then

$$r_n^2 \phi_n(r_n^{-1}) = n^{2/3}\phi_n(n^{-1/3}) = n^{1/2} + n^{1/6}n^{-1/3} = O(n^{1/2})$$

Thus Assumption (C) holds.

Hence $n^{1/3}\hat{g}_n = O_P(1)$.

Weak Convergence

Denote $\hat{h}_n = n^{1/3}\hat{g}_n = n^{1/3} \arg\max_{\{g>-t\}} M_n(g)$.

Rewriting, and multiplying by $n^{2/3}$, we have $n^{2/3}M_n(n^{-1/3}h)$

$$
\begin{aligned}
= \quad & n^{2/3}(\mathbb{P}_n - P)\left(\mathbf{1}\{X \le t + hn^{-1/3}\} - \mathbf{1}\{X \le t\}\right) \\
& + n^{2/3}\left[F(t + hn^{-1/3}) - F(t) - f(t)hn^{-1/3}\right] - xh\,.
\end{aligned}
$$

It can be shown that

$$
n^{2/3}M_n(n^{-1/3}h) \rightsquigarrow \mathbb{H}(h) \equiv \sqrt{f(t)}\mathbb{Z}(h) + \frac{1}{2}f'(t)h^2 - xh,
$$

where $\mathbb{Z}$ is a two-sided Brownian motion.

We use the argmax theorem to prove that

$$\arg\max\{n^{2/3}M_n(n^{-1/3}h) = \hat{h}_n \rightsquigarrow \hat{h} = \arg\max\ \mathbb{H}.$$

We need to show

- $\mathbb{H}$ is continuous and has a unique maximum.
- $n^{2/3}M_n(n^{-1/3}h) \stackrel{\mathsf{P}}{\to} \mathbb{H}(h)$ uniformly on compacts.
- $M_n(n^{-1/3}\hat{h}_n) = \sup_h M_n(n^{-1/3}h)$.

Using the rescaling attributes of Brownian motion, we have

$$\arg\max \mathbb{H} = \left|\frac{4f(t)}{[f'(t)]^2}\right|^{1/3} \arg\max_h\{\mathbb{Z}(h) - h^2\} + \frac{x}{f'(t)}\,.$$

Simple algebra yields

$$\begin{aligned} P&\left(\left|\frac{4f(t)}{[f'(t)]^2}\right|^{1/3} \arg\max_h\{\mathbb{Z}(h) - h^2\} + \frac{x}{f'(t)} \le 0\right) \\ &= \quad P\left(\left|4f'(t)f(t)\right|^{1/3} \arg\max_h \left\{\mathbb{Z}(h - h^2\right\} \le x\right), \end{aligned}$$

By the inverse function representation we have

$$\begin{aligned}
&P(n^{1/3}(\hat{f}_n(t) - f(t)) \leq x) \\
&\quad = \quad P(\hat{f}_n(t) \leq f(t) + xn^{-1/3}) \\
&\quad = \quad P(\hat{s}_n(f(t) + xn^{-1/3}) < t) \\
&\quad = \quad P(\arg\max_h \{M_n(n^{-1/3}h)\} \leq 0) \\
&\quad = \quad P(\hat{h}_n \leq 0) \\
&\quad \to \quad P(\hat{h} \leq 0) \\
&\quad = \quad P\left(\left|4f'(t)f(t)\right|^{1/3} \arg\max_h \left\{\mathbb{Z}(h) - h^2\right\} \leq x\right)
\end{aligned}$$

Summarizing:

$$n^{1/3}(\hat{f}_n(t) - f(t)) \rightsquigarrow |4f'(t)f(t)|^{1/3}\mathbb{C},$$

where the random variable $\mathbb{C} \equiv \arg\max_h\{\mathbb{Z}(h) - h^2\}$ has Chernoff's distribution.