M-Estimators

Yu Gu

September 16, 2021

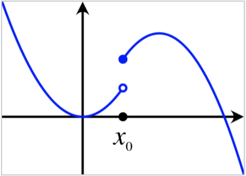
Outline

Upper semicontinuity

A function $f : \mathbb{D} \mapsto \mathbb{R}$ is *upper semicontinuous* if it satisfies either of the following two conditions:

(i) For all $c \in \mathbb{R}$, the set $\{x : f(x) \geq c\}$ is closed.

(ii) For all $x_0 \in \mathbb{D}$, $\limsup_{x \to x_0} f(x) \leq f(x_0)$.



Portmanteau theorem

$X_n \rightsquigarrow X$ if and only if

$$\limsup_{n\to\infty} P^*(X_n \in F) \le P(X \in F)$$

for every closed F.

Maximal inequality with uniform entropy integral

Let $\mathcal{F}$ be a P-measurable class of measurable functions with measurable envelope function F. Then, for every $p \geq 1$,

$$\left\| \|\mathbb{G}_n\|_{\mathcal{F}}^* \right\|_{P,p} \lesssim J(1, \mathcal{F}) \|F\|_{P,2\vee p},$$

where the uniform entropy integral is defined as

$$J(\delta, \mathcal{F}) = \sup_Q \int_0^\delta \sqrt{1 + \log N\big(\epsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q)\big)}\, d\epsilon.$$

Maximal inequality with bracketing integral

Let $\mathcal{F}$ be a class of measurable functions with measurable envelope function F. Then

$$\left\| \|\mathbb{G}_n\|_{\mathcal{F}}^* \right\|_{P,1} \lesssim J_{[]}\big(1, \mathcal{F}, L_2(P)\big) \|F\|_{P,2},$$

where the bracketing integral is defined as

$$J_{[]}(\delta, \mathcal{F}, \|\cdot\|) := \int_0^\delta \sqrt{1 + \log N_{[]}\big(\epsilon\|F\|, \mathcal{F}, \|\cdot\|\big)}\, d\epsilon.$$

Size of Lipschitz class of functions

Suppose the class of functions $\mathcal{F} = \{f_t : t \in T\}$ satisfies

$$|f_s(x) - f_t(x)| \leq d(s,t)F(x)$$

for every $s, t \in T$ and some fixed function F. Then, for any norm $\|\cdot\|$,

$$N_{[]}(2\epsilon\|F\|, \mathcal{F}, \|\cdot\|) \leq N(\epsilon, T, d).$$

M-estimators

- Consider a sequence $\{\mathbb{M}_n(\theta) : \theta \in \Theta\}$ of stochastic processes.
- *M-estimators* are (approximate) maximizers (or minimizers) $\hat{\theta}_n$ of criterion functions $\theta \mapsto \mathbb{M}_n(\theta)$.
- Examples:
  - maximum likelihood estimators
  - least squares estimators

M-estimators (cont.)

- Usually, the criterion function $\mathbb{M}_n(\theta)$ is an empirical process indexed by Θ.

- For i.i.d. observations $X_1, \ldots, X_n$, a common empirical criterion function is of the form

$$\theta \mapsto \mathbb{M}_n(\theta) = \mathbb{P}_n m_\theta.$$

- Let $\{\mathbb{M}(\theta) : \theta \in \Theta\}$ be a limiting process.

Problems of interest

- For M-estimators $\hat{\theta}_n$:
  - consistency for the true parameter θ_0
  - rate of convergence r_n
  - limiting distribution
- For local parameters $\hat{h}_n = r_n(\hat{\theta}_n - \theta_0)$:
  - weak convergence to some random point $\hat{h}$
  - Usually, $\hat{h}$ is the maximizer of $\mathbb{M}(h)$.

Outline

Preliminary arguments

- If the argmax functional were continuous w.r.t. some metric on the space of criterion functions, then weak convergence of the criterion functions would imply weak convergence of the M-estimators by the continuous mapping theorem.

- In keeping with the setup for empirical process, we endow the space of criterion functions with the *uniform metric*.

- The argmax functional is continuous at functions $\mathbb{M}$ that have a unique, *well-separated* maximizer: $\mathbb{M}(\hat{h}) > \sup_{h \notin G} \mathbb{M}(h)$ almost surely for any neighborhood G of $\hat{h}$.

Lemma 1

Let $\mathbb{M}_n$, $\mathbb{M}$ be stochastic processes indexed by a metric space H. Let A and B be arbitrary subsets of H. Suppose that

(i) $\mathbb{M}(\hat{h}) > \sup_{h \notin G, h \in A} \mathbb{M}(h)$ *almost surely, for every open set G that contains $\hat{h}$.*

(ii) $\mathbb{M}_n(\hat{h}_n) \geq \sup_h \mathbb{M}_n(h) - o_p(1)$.

(iii) $\mathbb{M}_n \rightsquigarrow \mathbb{M}$ *in* $\ell^\infty(A \cup B)$.

Then, for every closed set F,

$$\limsup_{n \to \infty} P^*(\hat{h}_n \in F \cap A) \leq P(\hat{h} \in F \cup B^c).$$

Note: $A = B = H \Rightarrow \hat{h}_n \rightsquigarrow \hat{h}$ (portmanteau theorem).

Proof

- By the continuous mapping theorem,

$$\sup_{h\in F\cap A} \mathbb{M}_n(h) - \sup_{h\in B} \mathbb{M}_n(h) \rightsquigarrow \sup_{h\in F\cap A} \mathbb{M}(h) - \sup_{h\in B} \mathbb{M}(h)$$

- Thus, by Slutsky's lemma and the portmanteau theorem,

$$\begin{aligned} &\limsup_{n\to\infty} P^*(\hat{h}_n \in F\cap A) \\ &\overset{(ii)}{\leq} \limsup_{n\to\infty} P^*\left(\sup_{h\in F\cap A} \mathbb{M}_n(h) \geq \sup_{h\in B} \mathbb{M}_n(h) - o_p(1)\right) \\ &\leq P\underbrace{\left(\sup_{h\in F\cap A} \mathbb{M}(h) \geq \sup_{h\in B} \mathbb{M}(h)\right)}_{\Rightarrow \hat{h}\in F\cup B^c}. \end{aligned}$$

- If $\hat{h} \notin F\cup B^c$, then $\hat{h} \in F^c \cap B$, which implies

$$\sup_{h\in B} \mathbb{M}(h) \geq \mathbb{M}(\hat{h}) > \sup_{h\notin F^c, h\in A} \mathbb{M}(h) = \sup_{h\in F\cap A} \mathbb{M}(h).$$

Remarks

- The assumption that $\mathbb{M}_n \rightsquigarrow \mathbb{M}$ uniformly in the whole parameter space is too strong.
- If dropping this assumption, additional properties of $\hat{h}_n$ need to be established in order to obtain $\hat{h}_n \rightsquigarrow \hat{h}$.
- The Argmax theorem requires uniform tightness of $\hat{h}_n$ and uniform convergence of $\mathbb{M}_n$ on compact subspace.

Theorem 2 (Argmax theorem)

Let $\mathbb{M}_n$, $\mathbb{M}$ be stochastic processes indexed by a metric space H. Suppose that

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

(ii) *The sequence $\hat{h}_n$ is uniformly tight and satisfies $\mathbb{M}_n(\hat{h}_n) \geq \sup_h \mathbb{M}_n(h) - o_p(1)$.*

(iii) *$\mathbb{M}_n \rightsquigarrow \mathbb{M}$ in $\ell^\infty(K)$ for every compact $K \subset H$.*

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

Proof

Step 1 Show that on compacta, a unique maximum of an upper semicontinuous function is well-separated. That is,

$$\mathbb{M}(\hat{h}) > \sup_{h \notin G, h \in K} \mathbb{M}(h)$$

almost surely, for every open set G that contains $\hat{h}$.

- If this is not true, then there exist an open set G around $\hat{h}$ and a sequence $h_m \in G^c \cap K$ such that $\mathbb{M}(h_m) \to \mathbb{M}(\hat{h})$.
- Since K is compact, $\{h_m\}$ has a subsequence $\{\tilde{h}_m\}$ that converges to some $h^* \in G^c \cap K$.
- By upper semicontinuity, $\mathbb{M}(h^*) \geq \limsup_m \mathbb{M}(\tilde{h}_m) = \mathbb{M}(\hat{h})$.
- This contradicts the uniqueness of $\hat{h}$, since $\hat{h} \in G$ while $h^* \in G^c$.

Proof (cont.)

Step 2 Apply the previous lemma with $A = B = K$ and obtain that for every closed set F,

$$
\begin{aligned}
&\limsup_{n\to\infty} P^*(\hat{h}_n \in F) \\
\leq\ &\limsup_{n\to\infty} P^*(\hat{h}_n \in F \cap K) + \limsup_{n\to\infty} P^*(\hat{h}_n \in K^c) \\
\leq\ &P(\hat{h} \in F \cup K^c) + \limsup_{n\to\infty} P^*(\hat{h}_n \in K^c) \\
\leq\ &P(\hat{h} \in F) + P(\hat{h} \in K^c) + \limsup_{n\to\infty} P^*(\hat{h}_n \in K^c).
\end{aligned}
$$

Due to uniform tightness of $\hat{h}_n$ and tightness of $\hat{h}$, K can be chosen to make the last two terms arbitrarily small.

Remarks

- The preceding lemma and the Argmax theorem are typically applied to a local parameter h, but they can also be applied to the original parameter θ.
- Since the limiting criterion function $\mathbb{M}(\theta)$ is typically nonrandom, the approach turns into a consistency proof.

Corollary 3 (Consistency)

Let $\mathbb{M}_n$ be stochastic processes indexed by a metric space Θ, and let $\mathbb{M} : \Theta \mapsto \mathbb{R}$ be a deterministic function.

(A) *Suppose that*

(i) $\mathbb{M}(\theta_0) > \sup_{\theta \notin G} \mathbb{M}(\theta)$ *for every open set G that contains θ_0.*

(ii) $\mathbb{M}_n(\hat{\theta}_n) \geq \sup_\theta \mathbb{M}_n(\theta) - o_p(1)$.

(iii) $\|\mathbb{M}_n - \mathbb{M}\|_\Theta \to 0$ *in outer probability.*

Then $\hat{\theta}_n \to \theta_0$ in outer probability.

(B) *Suppose that*

(i) *The map $\theta \mapsto \mathbb{M}(\theta)$ is upper semicontinuous with a unique maximum at θ_0.*

(ii) *The sequence $\hat{\theta}_n$ is uniformly tight and satisfies $\mathbb{M}_n(\hat{\theta}_n) \geq \sup_\theta \mathbb{M}_n(\theta) - o_p(1)$.*

(iii) $\|\mathbb{M}_n - \mathbb{M}\|_K \to 0$ *in outer probability for every compact $K \subset \Theta$.*

Then $\hat{\theta}_n \to \theta_0$ in outer probability.

Equivalent condition for i.i.d. data

In the case of i.i.d. data, $\mathbb{M}_n(\theta) = \mathbb{P}_n m_\theta$ and $\mathbb{M} = \mathbb{P} m_\theta$, the uniform convergence in (iii) is valid if and only if the class of functions $\{m_\theta : \theta \in \Theta\}$ is Glivenko-Cantelli.

Outline

Preliminary arguments

- If $\mathbb{M}(\theta)$ is twice differentiable at a point of maximum θ_0, then $\mathbb{M}'(\theta_0) = 0$ and $\mathbb{M}''(\theta_0)$ is negative definite.
- It is natural to assume that $\mathbb{M}(\theta) - \mathbb{M}(\theta_0) \lesssim -d^2(\theta, \theta_0)$ for every θ in a neighborhood of θ_0.
- The *modulus of continuity* of a stochastic process $\{X(t) : t \in T\}$ is defined by

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

An upper bound for the rate of convergence of $\hat{\theta}_n$ can be obtained from the modulus of continuity of $\mathbb{M}_n - \mathbb{M}$ at θ_0.

Theorem 4 (Rate of convergence)

Let $\mathbb{M}_n$ be stochastic processes indexed by a semimetric space Θ and $\mathbb{M} : \Theta \to \mathbb{R}$ a deterministic function. Suppose that

(i) *For every θ in a neighborhood of θ_0,*

$$\mathbb{M}(\theta) - \mathbb{M}(\theta_0) \lesssim -d^2(\theta, \theta_0).$$

(ii) *For every n and sufficiently small δ, the centered process $\mathbb{M}_n - \mathbb{M}$ satisfies*

$$E^* \sup_{d(\theta,\theta_0)<\delta} \left|(\mathbb{M}_n - \mathbb{M})(\theta) - (\mathbb{M}_n - \mathbb{M})(\theta_0)\right| \lesssim \frac{\phi_n(\delta)}{\sqrt{n}},$$

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

(iii) *The sequence $\hat{\theta}_n$ converges in outer probability to θ_0 and satisfies $\mathbb{M}_n(\hat{\theta}_n) \geq \mathbb{M}_n(\theta_0) - O_p(r_n^{-2})$ for some sequence r_n such that*

$$r_n^2 \phi_n(r_n^{-1}) \leq \sqrt{n} \quad \text{for every } n.$$

Then $r_n d(\hat{\theta}_n, \theta_0) = O_p^*(1)$. If the displayed conditions are valid for every θ and δ, then the condition that $\hat{\theta}_n$ is consistent is unnecessary.

Remarks

- The theorem remains true if replacing the metric function d by an arbitrary function $\tilde{d} : \Theta \times \Theta \mapsto [0, \infty)$ that satisfies $\tilde{d}(\theta_n, \theta_0) \to 0$ whenever $d(\theta_n, \theta_0) \to 0$.
- When $\phi(\delta) = \delta^\alpha$, the rate r_n is at least $n^{1/(4-2\alpha)}$.
- In particular, the "usual" rate $\sqrt{n}$ corresponds to $\phi(\delta) = \delta$.

Proof

Assume for simplicity that $\hat{\theta}_n$ truly maximizes $\mathbb{M}_n(\theta)$. We want to show

$$P^* \left(r_n d(\hat{\theta}_n, \theta_0) > 2^M \right) \to 0 \text{ as } M \to \infty, \text{ for every } n \text{ large enough.}$$

Ideas of proof:

- Partition the parameter space $\Theta \backslash \{\theta_0\}$ into disjoint "shells"

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

  with j ranging over the integers.
- For a given integer M, $r_n d(\hat{\theta}_n, \theta_0) > 2^M$ implies that $\hat{\theta}_n$ is in one of the shells $S_{j,n}$ with $j \geq M$.
- Bound above the probability that $\hat{\theta}_n \in S_{j,n}$.
  - For very large j, use the consistency of $\hat{\theta}_n$.
  - For smaller j, combine the remaining conditions.

Proof (cont.)

Fix $\eta > 0$ small enough such that

$$\sup_{\theta : d(\theta, \theta_0) \leq \eta} \mathbb{M}(\theta) - \mathbb{M}(\theta_0) \lesssim -d^2(\theta, \theta_0)$$

and such that for every $\delta \leq \eta$,

$$E^* \sup_{d(\theta, \theta_0) < \delta} \left| (\mathbb{M}_n - \mathbb{M})(\theta) - (\mathbb{M}_n - \mathbb{M})(\theta_0) \right| \lesssim \frac{\phi_n(\delta)}{\sqrt{n}}.$$

Such η exists by Conditions (i) and (ii).

Proof (cont.)

For n large enough,

$$\begin{aligned}
&P^*\left(r_n d(\hat{\theta}_n, \theta_0) > 2^M\right) \\
&= P^*\left(2^M < r_n d(\hat{\theta}_n, \theta_0) \leq \eta r_n/2\right) + P^*\left(r_n d(\hat{\theta}_n, \theta_0) > \eta r_n/2\right) \\
&\leq \sum_{j \geq M,\, 2^j \leq \eta r_n} P^*(\hat{\theta}_n \in S_{j,n}) + P^*(d(\hat{\theta}_n, \theta_0) > \eta/2).
\end{aligned}$$

The consistency of $\hat{\theta}_n$ for θ_0 guarantees that the second term converges to 0 as $n \to \infty$.

Proof (cont.)

Now we try to bound each term in the summation $\sum_{j \geq M,\, 2^j \leq \eta r_n} P^*(\hat{\theta}_n \in S_{j,n})$.

$$\hat{\theta}_n \in S_{j,n} \Rightarrow \sup_{\theta \in S_{j,n}} \left[\mathbb{M}_n(\theta) - \mathbb{M}_n(\theta_0)\right] \geq 0,$$

$$\text{Condition } (i) \Rightarrow \mathbb{M}(\theta) - \mathbb{M}(\theta_0) \lesssim -d^2(\theta, \theta_0) \lesssim -\frac{2^{2j-2}}{r_n^2},\ \forall \theta \in S_{j,n}.$$

Thus, the summation can be bounded by

$$\sum_{j \geq M,\, 2^j \leq \eta r_n} P^* \left(\sup_{\theta \in S_{j,n}} \left| (\mathbb{M}_n - \mathbb{M})(\theta) - (\mathbb{M}_n - \mathbb{M})(\theta_0) \right| \gtrsim \frac{2^{2j-2}}{r_n^2} \right)$$
$$\lesssim \sum_{j \geq M} \frac{\phi_n(2^j / r_n) r_n^2}{\sqrt{n} 2^{2j}} \lesssim \sum_{j \geq M} 2^{j(\alpha-2)},$$

by Markov's inequality, Condition (ii), the condition that $r_n^2 \phi_n(r_n^{-1}) \leq \sqrt{n}$, and the fact that $\phi_n(c\delta) \leq c^\alpha \phi_n(\delta)$ for every $c > 1$ (since $\phi_n(\delta)/\delta^\alpha$ is decreasing). The term on the right converges to 0 as $M \to \infty$.

Proof (cont.)

If Conditions (i) and (ii) are valid for every θ and δ, then we do not need to split $P^*\left(r_n d(\hat{\theta}_n, \theta_0) > 2^M\right)$ into two parts. We can use the same arguments on the previous slide to complete the proof.

Under i.i.d. setting

- Recall Condition (ii) in the preceding theorem:

$$E^* \sup_{d(\theta,\theta_0)<\delta} \left|(\mathbb{M}_n - \mathbb{M})(\theta) - (\mathbb{M}_n - \mathbb{M})(\theta_0)\right| \lesssim \frac{\phi_n(\delta)}{\sqrt{n}}$$

- For i.i.d. data and empirical criterion functions $\mathbb{M}_n(\theta) = \mathbb{P}_n m_\theta$ and $\mathbb{M}(\theta) = P m_\theta$, Condition (ii) involves the suprema of the empirical process $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - \mathbb{P})$ indexed by classes of functions

$$\mathcal{M}_\delta := \{m_\theta - m_{\theta_0} : d(\theta,\theta_0) < \delta\}.$$

- It is reasonable to assume that these suprema are bounded uniformly in n.

Corollary 5

In the i.i.d. case, assume that

(i) *For every θ in a neighborhood of θ_0,*

$$P\left(m_\theta - m_{\theta_0}\right) \lesssim -d^2(\theta, \theta_0).$$

(ii) *There exists a function ϕ such that $\delta \mapsto \phi(\delta)/\delta^\alpha$ is decreasing for some $\alpha < 2$ and, for every n,*

$$E^* \left\|\mathbb{G}_n\right\|_{\mathcal{M}_\delta} \lesssim \phi(\delta).$$

(iii) *The sequence $\hat{\theta}_n$ converges in outer probability to θ_0 and satisfies $\mathbb{P}_n m_{\hat{\theta}_n} \geq \sup_{\theta\in\Theta} \mathbb{P}_n m_\theta - O_p(r_n^{-2})$ for some sequence r_n such that*

$$r_n^2 \phi_n(r_n^{-1}) \leq \sqrt{n} \quad \textit{for every } n.$$

Then $r_n d(\hat{\theta}_n, \theta_0) = O_p^*(1)$.

Bounds on continuity modulus

- It is important to derive a sharp bound on the modulus of continuity of $\mathbb{G}_n$ before applying the corollary.
- A simple but not necessarily efficient approach is to apply the maximal inequalities to the class $\mathcal{M}_\delta$, which yield

$$E_P^* \|\mathbb{G}_n\|_{\mathcal{M}_\delta} \lesssim J(1, \mathcal{M}_\delta)(P^* M_\delta^2)^{1/2},$$
$$E_P^* \|\mathbb{G}_n\|_{\mathcal{M}_\delta} \lesssim J_{[]}\big(1, \mathcal{M}_\delta, L_2(P)\big)(P^* M_\delta^2)^{1/2}.$$

- These bounds depend mostly on the envelope function M_δ.
- Assuming that the entropy integrals are bounded as $\delta \downarrow 0$, we obtain an upper bound $\phi(\delta) = (P^* M_\delta^2)^{1/2}$ on the modulus.
- By the preceding corollary, r_n is at least the solution of

$$r_n^4 P^* M_{1/r_n}^2 \sim n.$$

Outline

Overview

Key steps to obtain the limiting distribution of M-estimators of Euclidean parameters:

- Establish the consistency of $\hat{\theta}_n$ for the true parameter θ_0.
- Establish the rate of convergence r_n of $\hat{\theta}_n$.
- Define *rescaled criterion functions* as a multiple of the map

$$h \mapsto \mathbb{M}_n(\theta_0 + h/r_n) - \mathbb{M}_n(\theta_0),$$

  which are maximized at local parameters $\hat{h}_n = r_n(\hat{\theta}_n - \theta_0)$.
- Show that suitably rescaled criterion functions converge weakly to a limiting process $\mathbb{M}$ in $\ell^\infty(h : \|h\| \leq K)$ for every K.

If the sample paths $h \mapsto \mathbb{M}(h)$ are upper semicontinuous and possess a unique maximizer $\hat{h}$, then by the Argmax theorem, $\hat{h}_n \rightsquigarrow \hat{h}$.

Overview (cont.)

- For illustration, we derive the limiting distribution of Euclidean M-estimators under the pointwise Lipschitz condition.
- We will combine the Argmax theorem and the rate of convergence theorem.
- More general results on Euclidean M-estimators are given in Theorem 3.2.10 of VW.

Notations

- The parameter space Θ is an open subset of Euclidean space, equipped with the Euclidean distance.
- We assume i.i.d observations and use the empirical process notations: $\mathbb{M}_n(\theta) = \mathbb{P}_n m_\theta$ and $\mathbb{M}(\theta) = P m_\theta$.
- Like before, for any $\delta > 0$, define the class of functions

$$\mathcal{M}_\delta := \{m_\theta - m_{\theta_0} : d(\theta, \theta_0) < \delta\}.$$

Assumptions

1. θ_0 is a point of maximum of $\mathbb{M}(\theta)$ in the interior of Θ.
2. $\hat{\theta}_n$ maximize $\mathbb{M}_n(\theta)$ for every n and is consistent for θ_0.
3. $\mathbb{M}(\theta)$ has a nonsingular second derivative matrix V.
4. There exist some square-integrable functions $\dot{m}$ and $\dot{m}_{\theta_0}$ such that

$$\left|m_{\theta_1}(x) - m_{\theta_2}(x)\right| \leq \dot{m}(x) \left\|\theta_1 - \theta_2\right\|, \quad (1)$$

$$P\left(m_\theta - m_{\theta_0} - (\theta - \theta_0)^T \dot{m}_{\theta_0}\right)^2 = o(\|\theta - \theta_0\|^2), \quad (2)$$

for all $\theta_1, \theta_2, \theta$ in some neighborhood of θ_0.

$\sqrt{n}$ rate of convergence

- Under the Lipschitz assumption, $F_\delta = \delta \dot{m}$ is an envelope function for the class $\mathcal{M}_\delta$. By the theorem on bracketing numbers,

$$N_{[]}\left(2\epsilon \|F_\delta\|_{P,2}, \mathcal{M}_\delta, L_2(P)\right) \le N(\epsilon, B(\theta_0, \delta), \|\cdot\|) \lesssim \epsilon^{-p}.$$

- Applying the maximal inequality yields

$$\mathrm{E}^* \|\mathbb{G}_n\|_{\mathcal{M}_\delta} \lesssim \|F_\delta\|_{P,2} \lesssim \delta.$$

- Thus, the modulus of continuity condition in the rate theorem is satisfied for $\phi(\delta) = \delta$. Hence, we obtain the $\sqrt{n}$ rate for $\hat{\theta}_n$.

Limiting distribution

- Define rescaled criterion functions

$$\mathbb{U}_n(h) := n\left(\mathbb{M}_n\left(\theta_0 + h/\sqrt{n}\right) - \mathbb{M}_n\left(\theta_0\right)\right)$$

  and local parameters $\hat{h}_n = \sqrt{n}(\hat{\theta}_n - \theta_0)$. Obviously, $\hat{h}_n$ maximizes $\mathbb{U}_n(h)$ for every n.
- We rewrite $\mathbb{U}_n$ as

$$\begin{aligned}\mathbb{U}_n &= \mathbb{G}_n\left[\sqrt{n}\left(m_{\theta_0+h/\sqrt{n}} - m_{\theta_0}\right) - h^T\dot{m}_{\theta_0}\right] \\ &\quad + h^T\mathbb{G}_n\dot{m}_{\theta_0} + n\left(\mathbb{M}\left(\theta_0 + h/\sqrt{n}\right) - \mathbb{M}\left(\theta_0\right)\right) \\ &= \mathbb{E}_n(h) + h^T\mathbb{G}_n\dot{m}_{\theta_0} + \frac{1}{2}h^T Vh + o(1).\end{aligned}$$

- Provided that for any compact $K \subset \Theta$, $\|\mathbb{E}_n\|_K = o_p(1)$. Then $\mathbb{U}_n \rightsquigarrow \mathbb{U}$ in $\ell^\infty(K)$, where $\mathbb{U}(h) = h^T Z + \frac{1}{2}h^T Vh$, and Z is the Gaussian limiting distribution of $\mathbb{G}_n\dot{m}_{\theta_0}$.
- By the Argmax theorem, $\hat{h}_n = \sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow \hat{h}$, where $\hat{h}$ is the maximizer of $\mathbb{U}(h)$.