M-Estimation, I: Introduction and Asymptotic Properties

Jesper Riis–Vestergaard Sørensen

University of Copenhagen, Department of Economics

Outline

Introduction

Nonlinear Estimation Chapters

- W. Chapters 12–13: Abstract and technical.
- But generality can be useful!
- Unified framework for estimation.
  - **Ex:** OLS, Nonlinear LS, Maximum likelihood...
- There will be no exam questions in Ch. 12–13 *specifically*.
- But important—and required—background knowledge.

Steps in Econometric Analysis

1. Identification: Given distribution of observables, (how) can we uniquely recover parameters?

2. Estimation: Given sample, how to construct parameter estimates?

   Consistency, rates of convergence

3. Inference: Confidence intervals, prediction intervals, hypothesis testing, etc.

Steps in Econometrics Analysis

- Identification: Has nothing to do with sample.
- Estimation: What formula(e)/algorithm to follow?
- Inference: Requires (asymptotic) distribution theory.

Steps highly interdependent.

- Identification method may suggest estimator.
- Inference method hinges on estimation method.

Nonlinear Regression

Nonlinear Regression Model

- y: scalar outcome.
- $\mathbf{x}$: K-vector of explanatory variables.
- candidate estimators
  $m(\mathbf{x}, \boldsymbol{\theta})$ parametric model for $E(y|\mathbf{x})$.
- $\Theta \subseteq \mathbb{R}^P$ parameter space. Fixed dim P.
- Mean model correctly specified if

  expected value of y given x is exactly given by our model thereof at \theta_0 -> for all past realisations of x

  $$E(y|\mathbf{x}) = m(\mathbf{x}, \boldsymbol{\theta}_o) \quad \forall x \tag{1}$$

  candidate paramteter

  holds for some $\boldsymbol{\theta}_o \in \Theta$.
- $\boldsymbol{\theta}_o$ often called “true value of theta.”

Examples of Functional Form

Ex. $m(x, \theta) = x\theta$
(linear)

like income

- **Ex.** If y nonnegative, may take

$$m(\mathbf{x}, \boldsymbol{\theta}) = \exp(\mathbf{x}\boldsymbol{\theta}). \qquad \text{(exponential regression)}$$

- **Ex.** If $y \in \{0, 1\}$, may take

$$m(\mathbf{x}, \boldsymbol{\theta}) = \frac{1}{1 + \exp(-\mathbf{x}\boldsymbol{\theta})}. \qquad \text{(logistic regression)}$$

sigmoid-function?

- Here: $K = P$. But $K \gtrless P$ allowed.

no of explanatory variables (K) = no of candidate parameters (P)

Error Formulation

- Assume correct specification (NLS.1).
- Defining $u := y - m(\mathbf{x}, \boldsymbol{\theta}_o)$, may write

$$y = m(\mathbf{x}, \boldsymbol{\theta}_o) + u, \quad E(u|\,\mathbf{x}) = 0.$$

- $E(u|\,\mathbf{x}) = 0$ a consequence of model.
    - Not an additional assumption.
- Error formulation useful for abbreviations.

Discussion

- $E(u|\mathbf{x}) = 0$ does *not* imply u and $\mathbf{x}$ independent.
- ... only cond'l *mean* independence.
  - heteroskedasticity
    May have $\text{var}(u|\mathbf{x})$ nonconstant (in $\mathbf{x}$).
  - If $y \geqslant 0$, must have $u \geqslant -m(\mathbf{x}, \boldsymbol{\theta}_o) \ldots$
- Error formulation yields *semi*parametric model for $y|\mathbf{x}$.
  - Parametric model for $E(y|\mathbf{x})$.
  - But haven't specified parametric distribution for $u|\mathbf{x}$.

Identification

Identification

We'll show: $\boldsymbol{\theta}_o$ solves population problem (PP)

$$\min_{\boldsymbol{\theta} \in \Theta} E\left\{ \left[y - m\left(\mathbf{x}, \boldsymbol{\theta}\right)\right]^2 \right\}.$$

- Model m + parameter space Θ known quantities.

- Hence, **IF** given distribution of $(y, \mathbf{x})$, PP problem known.

- $\boldsymbol{\theta}_o$ identified if PP solution *unique*.

    there is one and only one solution to the population
    -> there cannot be another set of parameters that solves the population problem

Identification

$\theta \in \Theta$

$\pm m(\mathbf{x}, \boldsymbol{\theta}_o)$ and expanding square,

$$[y - m(\mathbf{x}, \boldsymbol{\theta})]^2 = \{[y - \overset{= u}{m(\mathbf{x}, \boldsymbol{\theta}_o)]} - [m(\mathbf{x}, \boldsymbol{\theta}) - m(\mathbf{x}, \boldsymbol{\theta}_o)]\}^2$$
$$= [y - m(\mathbf{x}, \boldsymbol{\theta}_o)]^2 + [m(\mathbf{x}, \boldsymbol{\theta}) - m(\mathbf{x}, \boldsymbol{\theta}_o)]^2$$
$$- 2u[m(\mathbf{x}, \boldsymbol{\theta}) - m(\mathbf{x}, \boldsymbol{\theta}_o)].$$

u^2

$E[E[\quad |x]]$

$E[u|x] = 0$

Taking expectations,

$$\underbrace{E\{[y - m(\mathbf{x}, \boldsymbol{\theta})]^2\}}_{\text{Criterion @ } \theta} = \overset{= u^2}{E\{[y - m(\mathbf{x}, \boldsymbol{\theta}_o)]^2\}}$$
$$+ E\{[m(\mathbf{x}, \boldsymbol{\theta}) - m(\mathbf{x}, \boldsymbol{\theta}_o)]^2\}.$$

Criterion @ θ_o

Identification

Have shown

"Population criterion function"

$$E\{[y - m(\mathbf{x}, \boldsymbol{\theta})]^2\} = E\{[y - m(\mathbf{x}, \boldsymbol{\theta}_o)]^2\} + \overbrace{E\{[m(\mathbf{x}, \boldsymbol{\theta}) - m(\mathbf{x}, \boldsymbol{\theta}_o)]^2\}}^{\geq 0}.$$

It follows that

$$E\{[y - m(\mathbf{x}, \boldsymbol{\theta})]^2\} \geqslant E\{[y - m(\mathbf{x}, \boldsymbol{\theta}_o)]^2\} \text{ for all } \boldsymbol{\theta} \in \Theta.$$

$\Rightarrow \boldsymbol{\theta}_o$ solves PP.

$\theta_o \in \operatorname{argmin} E\{[y - m(x,\theta)]^2\}$

Q: Uniqueness? does not yield uniqueness =

Identification Condition

Have shown

the true value of theta is A minimizer of the population problem

$$E\{\,[y - m(\mathbf{x},\boldsymbol{\theta})]^2\,\} = E\{\,[y - m(\mathbf{x},\boldsymbol{\theta}_o)]^2\,\} + E\{\,[m(\mathbf{x},\boldsymbol{\theta}) - m(\mathbf{x},\boldsymbol{\theta}_o)]^2\,\}.$$

expected square distance between our two models for the conditional mean (no 1 is cond.mean for candidate regressors, no 2 is the 'true' cond.mean)

$\boldsymbol{\theta}_o$ uniquely solves PP if and only if

$$E\{\,[m(\mathbf{x},\boldsymbol{\theta}) - m(\mathbf{x},\boldsymbol{\theta}_o)]^2\,\} > 0 \text{ for all } \boldsymbol{\theta} \in \Theta \setminus \{\boldsymbol{\theta}_o\}.$$

Q: When will identification fail?

Whenever we have multiple solutions to the population problem

Identification Failures / Successes

Example: *Linear* regression, $m(\mathbf{x}, \boldsymbol{\theta}) = \mathbf{x}\boldsymbol{\theta}$ with $\Theta = \mathbb{R}^K$.

Here

$$E\{\,[m(\mathbf{x}, \boldsymbol{\theta}) - m(\mathbf{x}, \boldsymbol{\theta}_o)]^2\,\} = E\{\,[\mathbf{x}(\boldsymbol{\theta} - \boldsymbol{\theta}_o)]^2\,\}$$
$$= (\boldsymbol{\theta} - \boldsymbol{\theta}_o)' \, E(\mathbf{x}'\mathbf{x}) \, (\boldsymbol{\theta} - \boldsymbol{\theta}_o).$$

for all \theta diff than \theta_0

$K \times K$

- > 0 if $E(\mathbf{x}'\mathbf{x})$ positive definite.

  columns must be linearly independent
  "FULL RANK"
- Just usual (population) rank condition. OLS.2 (?)

If $E(x'x)$ singular

$\exists\, \upsilon \neq \underline{0}$ s.t. $E(x'x)\upsilon = \underline{0}$. $\theta = \theta_0 + \upsilon$

$= \upsilon' E(x'x)\upsilon = 0.$

Identification Failures

```
= multiple solution to the population problem!
```

Example: *Non*linear regression with

$$m(\mathbf{x}, \boldsymbol{\theta}) = \theta_1 + \theta_2 x_2 + \theta_3 x_3^{\theta_4}.$$

- Suppose $\theta_{o3} = 0$. (Truth linear.)
- At $\boldsymbol{\theta}$ with $\theta_3 = 0\,(= \theta_{o3})$...
- ... criterion function *independent of* θ_4.

```
"\theta_4 disappears"
```

- For this $\boldsymbol{\theta}_o$, identification fails.
- Example of poorly identified model.

Estimation

Estimation

$\boldsymbol{\theta}_o$ solves PP,

$$\boldsymbol{\theta}_o \in \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmin}} \, E\left\{ \left[y - m\left(\mathbf{x}, \boldsymbol{\theta}\right)\right]^2 \right\}.$$

Analogy principle suggests,

$$\hat{\theta} = \widehat{\boldsymbol{\theta}}_N \in \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmin}} \, \frac{1}{N} \sum_{i=1}^{N} \left[y_i - m\left(\mathbf{x}_i, \boldsymbol{\theta}\right)\right]^2 .$$

Nonlinear least squares (NLS) estimator.

For now, assume existence (but not uniqueness) of solution.

Consistency? $\hat{\theta}_N \xrightarrow{p} \theta_o$

Q: Does NLS consistently estimate $\boldsymbol{\theta}_o$?

It turns out answer is "yes," provided (roughly)

1. $\boldsymbol{\theta}_o$ is identified,
2. Criterion function convergence

$$\frac{1}{N}\sum_{i=1}^{N}[y_i - m(\mathbf{x}_i, \boldsymbol{\theta})]^2 \text{ “}\rightarrow\text{” } E\{\,[y - m(\mathbf{x}, \boldsymbol{\theta})]^2\,\}$$

... in suitable sense.

'a heuristic convergence' = in a calculated-guess kind of sense

Next: More detail in general setting.

M-Estimation

M-Estimand "Target estimation"

We now consider more abstract setting.

Let $q(\mathbf{w}, \boldsymbol{\theta})$ denote function of

1. random vector $\mathbf{w}$ [observables, e.g. $\mathbf{w} = (\mathbf{y}, \mathbf{x})$],
2. parameters $\boldsymbol{\theta}$.

True parameter $\boldsymbol{\theta}_o$ assumed unique solution to PP

$$\boldsymbol{\theta}_o = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmin}} E[q(\mathbf{w}, \boldsymbol{\theta})].$$

"M" short for "minimization."

- Or "maximization" (sign change).

q sometimes called loss function.

M-Estimator

Given random (as in i.i.d.) sample $\{\mathbf{w}_i\}_1^N$.

Analogy principle suggests sample problem (SP)

$$\min_{\boldsymbol{\theta}\in\Theta} \frac{1}{N}\sum_{i=1}^{N} q\left(\mathbf{w}_i, \boldsymbol{\theta}\right).$$

Definition: Any SP solution is an M-estimator of $\boldsymbol{\theta}_o$.

Example M-Estimators

- OLS: $q(\mathbf{w}, \boldsymbol{\theta}) = (y - \mathbf{x}\boldsymbol{\theta})^2$.

- NLS: $q(\mathbf{w}, \boldsymbol{\theta}) = [y - m(\mathbf{x}, \boldsymbol{\theta})]^2$.

- Maximum likelihood: $q(\mathbf{w}, \boldsymbol{\theta}) = -\ln f(y|\, \mathbf{x}; \boldsymbol{\theta})$.

- Least absolute deviations (LAD): $q(\mathbf{w}, \boldsymbol{\theta}) = |y - \mathbf{x}\boldsymbol{\theta}|$.

- ...and many, many more.

Scope of Framework

Observables $\mathbf{w}_i$ allow scalar/vector outcome.

- One equation, one cross section $\Rightarrow$ scalar y_i.
- Multiple equations, one cross section $\Rightarrow$ vector $\mathbf{y}_i$.
  - **Ex:** Joint labor supply decision (wife/husband),

$$\begin{aligned} y_i^{\mathrm{w}} &= \text{labor supply, wife, family } i, \\ y_i^{\mathrm{h}} &= \text{labor supply, husband, family } i. \end{aligned}$$

- One equation, panel data $\Rightarrow$ vector $\mathbf{y}_i = (y_{i1}, \ldots, y_{iT})'$.
  - FE: $q(\mathbf{w}_i, \boldsymbol{\theta}) = \sum_{t=1}^{T} (\ddot{y}_{it} - \ddot{\mathbf{x}}_{it}\boldsymbol{\theta})^2$

Formulation very general!

Asymptotic Properties of M-Estimators

Recap: Setting

M-estimand solves population problem (PP),

$$\boldsymbol{\theta}_o \in \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmin}}\, E\left[q\left(\mathbf{w}, \boldsymbol{\theta}\right)\right].$$

M-estimator solves sample problem (SP),

$$\widehat{\boldsymbol{\theta}} \in \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmin}}\, \frac{1}{N} \sum_{i=1}^{N} q\left(\mathbf{w}_i, \boldsymbol{\theta}\right).$$

Q: Properties?

Consistency

Informal Look at Consistency

Criterion functions (minimands) and minimizers:

$N^{-1} \sum_{i=1}^{N} q(\mathbf{w}_i, \boldsymbol{\theta})$ $\qquad$ $E\left[q(\mathbf{w}, \boldsymbol{\theta})\right]$

$\widehat{\boldsymbol{\theta}}$ $\qquad$ $\boldsymbol{\theta}_o$

Q: Relationships?

Informal Look at Consistency

By definition of M-estimand and M-estimator:

$N^{-1}\sum_{i=1}^{N} q(\mathbf{w}_i, \boldsymbol{\theta})$	$E\left[q(\mathbf{w}, \boldsymbol{\theta})\right]$
argmin	argmin
$\widehat{\boldsymbol{\theta}}$	$\boldsymbol{\theta}_o$

Informal Look at Consistency

By (weak) law of large numbers,

$$
\begin{array}{ccc}
N^{-1}\sum_{i=1}^{N} q(\mathbf{w}_i,\boldsymbol{\theta}) & \xrightarrow{p} & E\left[q(\mathbf{w},\boldsymbol{\theta})\right] \\
\big\downarrow \text{argmin} & & \big\downarrow \text{argmin} \\
\widehat{\boldsymbol{\theta}} & & \boldsymbol{\theta}_o
\end{array}
$$

Informal Look at Consistency

$$\begin{array}{ccc} N^{-1}\sum_{i=1}^{N} q(\mathbf{w}_i,\boldsymbol{\theta}) & \xrightarrow{p} & E\left[q(\mathbf{w},\boldsymbol{\theta})\right] \\ \text{argmin} & & \text{argmin} \\ \widehat{\boldsymbol{\theta}} & \xrightarrow[?]{p} & \boldsymbol{\theta}_o \end{array}$$

Seems reasonable...

Q: When does mini*mand* convergence imply mini*mizer* convergence (in prob).

Formal Look at Consistency

Q: When is $\widehat{\boldsymbol{\theta}}$ consistent for $\boldsymbol{\theta}_o$?

Suffices (essentially) following two conditions hold:

1. Identification: $\boldsymbol{\theta}_o$ is identified.
2. Uniform Law of Large Numbers: S minimand converges to P equivalent *uniformly in probability*,

$$\max_{\boldsymbol{\theta} \in \Theta} \left| \frac{1}{N} \sum_{i=1}^{N} q\left(\mathbf{w}_i, \boldsymbol{\theta}\right) - E\left[q\left(\mathbf{w}, \boldsymbol{\theta}\right)\right] \right| \xrightarrow{p} 0.$$

Identification Assumption

At this level of abstractness, *assume* identification, i.e.

any theta different than the true one

$$E\left[q\left(\mathbf{w},\boldsymbol{\theta}\right)\right] > E\left[q\left(\mathbf{w},\boldsymbol{\theta}_o\right)\right] \text{ for all } \boldsymbol{\theta}\in\Theta\backslash\left\{\boldsymbol{\theta}_o\right\}.$$

In words: $\boldsymbol{\theta}_o$ unique solution to PP.

- May make less abstract in applications (later).

Uniform Law of Large Numbers

May *deduce* minimand convergence using:

Theorem (W. Theorem 12.1)

If

1. $\Theta \subseteq \mathbb{R}^P$ *compact (i.e. closed + bounded),*
2. $q(\mathbf{w}, \cdot)$ *continuous (in* $\boldsymbol{\theta}$*),* no matter the value of w (random vector of observables) $\forall w$

and additional technical conditions hold, then

$$\max_{\boldsymbol{\theta}\in\Theta} \left| \frac{1}{N} \sum_{i=1}^{N} q(\mathbf{w}_i, \boldsymbol{\theta}) - E[q(\mathbf{w}, \boldsymbol{\theta})] \right| \xrightarrow{p} 0.$$

sample criterion function converges in probability to its population equivalent

Uniform law of large numbers (ULLN).

Consistency Theorem

Theorem (W. Theorem 12.2)

Under the assumptions of W. Theorem 12.1 (ULLN) and assuming identification of $\boldsymbol{\theta}_o$,

1. $\widehat{\boldsymbol{\theta}}$ *solves SP, and*
2. $\widehat{\boldsymbol{\theta}}$ *is consistent for* $\boldsymbol{\theta}_o$, $\widehat{\boldsymbol{\theta}} \to_p \boldsymbol{\theta}_o$.

More Formal Consistency Argument

Proof Sketch:

1. Compact Θ + $q(\mathbf{w}, \cdot)$ continuous $\Rightarrow$ SP solution exists.

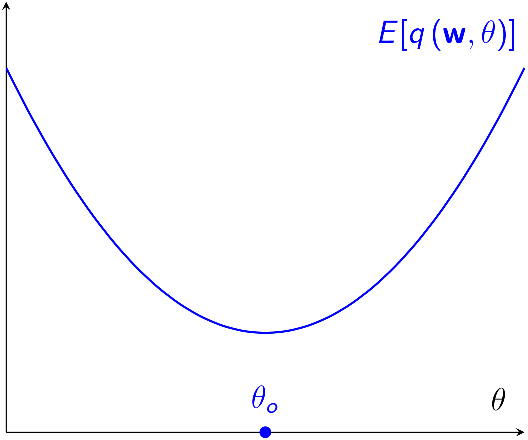
   Weierstrass Ext Value Thm.

   - Why? cont's fctn defined on compact space attains its extrema.

2. ULLN $\Rightarrow$ in limit, S/P minimands coincide (in prob). (maxima / minima)

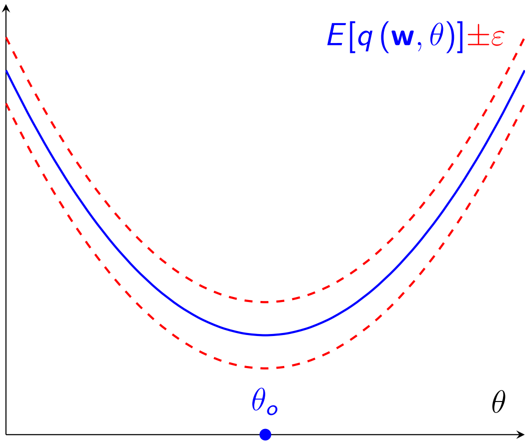
3. Identification implies unique PP solution, so must have $\widehat{\boldsymbol{\theta}} \rightarrow_p \boldsymbol{\theta}_o$.

Graphical Illustration of Consistency

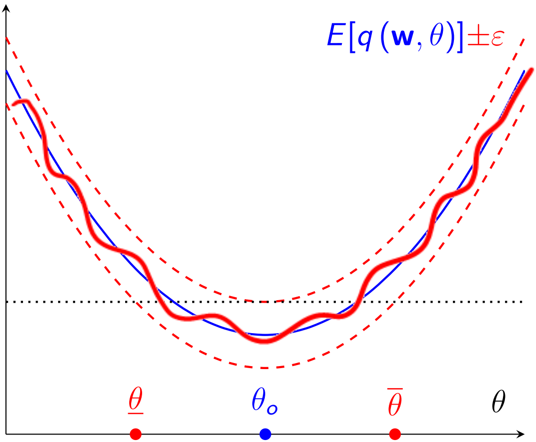


Graphical Illustration of Consistency

When minimand difference $\leqslant \varepsilon$, S minimand in "sleeve"



Graphical Illustration of Consistency



$\widehat{\theta}$ “squeezed in”

Role of Uniform Convergence

Consider (deterministic) functions

$$f_n(\theta) := \begin{cases} \frac{1}{2}, & \theta = 0, \\ 0, & \theta = n, \\ 1, & \text{otherwise.} \end{cases} \implies \operatorname{argmin} f_n = \underline{\quad n \quad} \quad (\approx \hat{\theta})$$

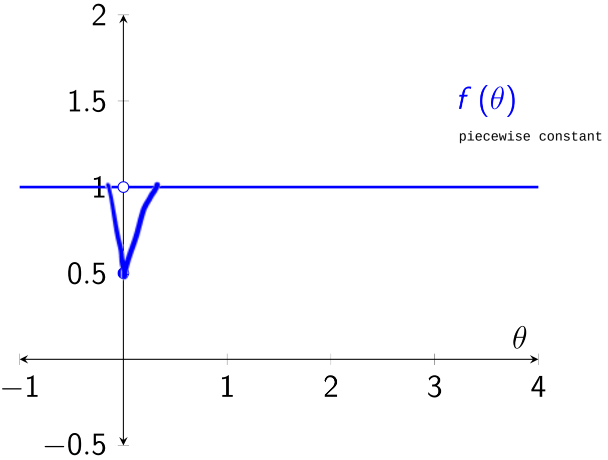
$n \to \infty$

For each θ, $f_n(\theta) \to f(\theta)$ where

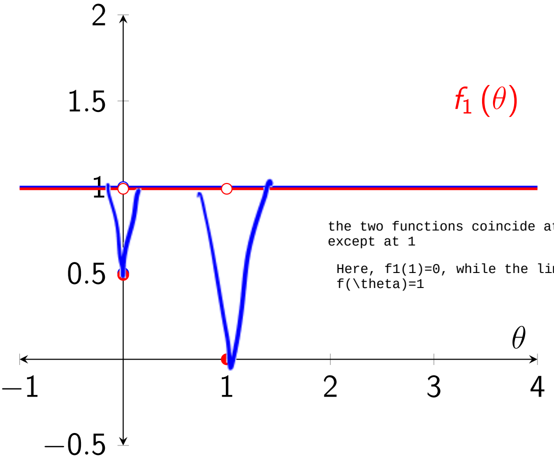
$$f(\theta) := \begin{cases} \frac{1}{2}, & \theta = 0, \\ 1, & \theta \neq 0. \end{cases} \implies \operatorname{argmin} f = \underline{\quad 0 \quad}$$

- Minimizer? escaping to the horizon. The sequence of minimizers grows without bound

Escape to Horizon



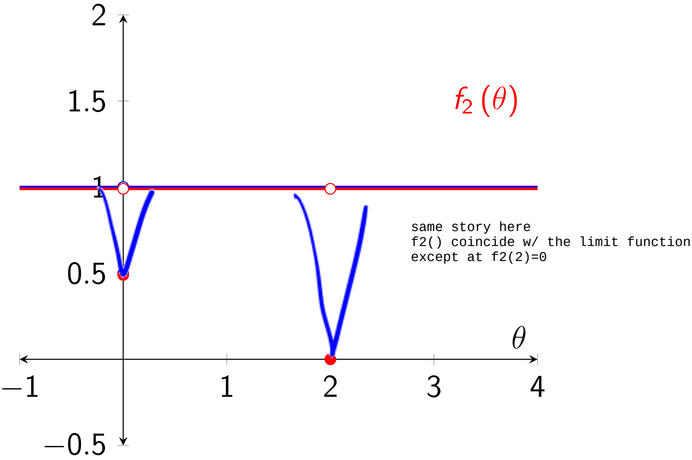
Escape to Horizon



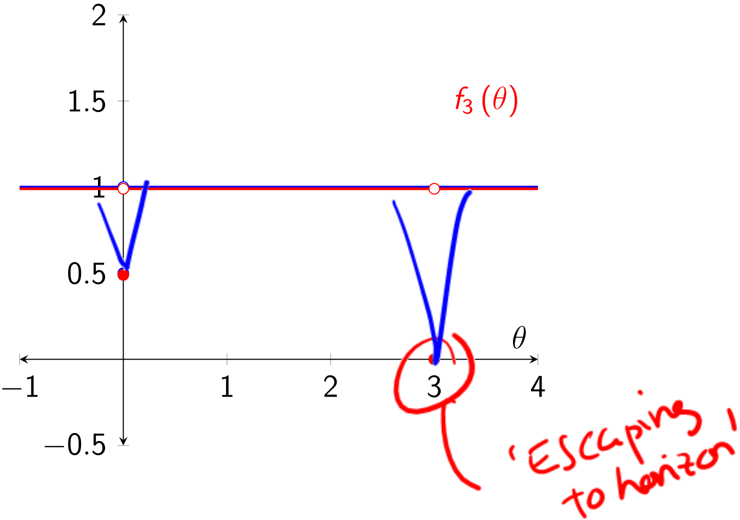
the two functions coincide at all points except at 1

Here, f1(1)=0, while the limit function f(\theta)=1

Escape to Horizon



Escape to Horizon



Role of Uniform Convergence

- **Problem?** why don't we see minimzer convergence? Convergence is not sufficiently uniform.

$$\max_{\theta \in \mathbb{R}} |f_n(\theta) - f(\theta)| = |f_n(n) - f(n)| = |0-1|$$

$$= 1 \not\to 0 \quad n \to \infty.$$

they coincide at every point,
except at a single point

the difference being 1 for all n as n grows without bound

- Similar problem with f_n stochastic.
- Example ruled out by compactness.

compact = closed and bounded (?)

  - $\Theta = \mathbb{R}$ *un*bounded.

since it is unbounded, it is not compact
in other words, it is not continous?

Necessity of Uniform Convergence

- Uniform convergence sufficient but not necessary.
- Think: Linear model + squared loss
$$q\left(\mathbf{w}, \boldsymbol{\theta}\right) = \left(y - \mathbf{x}\boldsymbol{\theta}\right)^2 .$$
  - Natural parameter space entire $\mathbb{R}^P$.
  - Estimator in closed form.
  - Uniform convergence/compactness not needed.
- Here: We use it to *deduce* minimizer convergence.

Normality

Additional Assumptions

Have for consistency invoked:

- $\boldsymbol{\theta}_o$ identified `unique solution/minimizer of the population objective function`
- Θ compact
  `A compact set is for example [0,1], but not (0,1) -> we have closed endpoints`
  `need to have finite amount of parameters in the parameter space`
- $q(\mathbf{w}, \cdot)$ continuous
- (+ technical...)

Asymptotic normality requires *stronger* assumptions.

Additional Assumptions

For asymptotic normality, add:

- $\boldsymbol{\theta}_o$ interior to Θ. [Draw]
- $q(\mathbf{w}, \cdot)$ twice continuously differentiable on $\operatorname{int}\Theta$

Remarks:

- Interiority requires $\operatorname{int}\Theta$ nonempty
- ... used to expand around $\boldsymbol{\theta}_o$
- Twice cont' diff' facilitates second-order expansion.

Additional Assumptions

Abbreviate

implies we are looking at
the transposed derivative
of q w.r.t theta

$$\text{Score:} \quad \mathbf{s}(\mathbf{w}, \boldsymbol{\theta}) := \frac{\partial}{\partial \boldsymbol{\theta}} q(\mathbf{w}, \boldsymbol{\theta}), \qquad (P \times 1)$$

$$\text{Hessian:} \quad \mathbf{H}(\mathbf{w}, \boldsymbol{\theta}) := \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} q(\mathbf{w}, \boldsymbol{\theta}). \qquad (P \times P)$$

$$= \frac{\partial}{\partial \theta'} s(w, \theta)$$

Further add:

- $E[\mathbf{s}(\mathbf{w}, \boldsymbol{\theta}_o)] = \mathbf{0}$,

  score func evaluated at the true value of \theta is zero

- $E[\mathbf{H}(\mathbf{w}, \boldsymbol{\theta}_o)]$ positive definite.

  Full rank condition
  No linearly dependent columns -> all columns are linearly independent

  - Essentially follow from FOC/SOC of minimization.

Let $\mathbf{A}$ be an $(n \ X \ K)$ matrix with $rank(\mathbf{A}) = K$:
$\mathbf{A}'\mathbf{A}$ is always positive definite and therefore nonsingular.

Asymptotic Normality of M-Estimators

Theorem (W. Thm 12.3)

Provided — consistency

- $\boldsymbol{\theta}_o$ *identified + interior to* Θ *compact,*
- $q(\mathbf{w}, \cdot)$ *cont' + twice cont' diff' on* $\operatorname{int}\Theta$,
- $E[\mathbf{s}(\mathbf{w}, \boldsymbol{\theta}_o)] = \mathbf{0}$, *and* $E[\mathbf{H}(\mathbf{w}, \boldsymbol{\theta}_o)]$ *positive definite,* } ~ FOC SOC
- *(+ technical),*

=√

$$\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \mathbf{A}_o^{-1}\mathbf{B}_o\mathbf{A}_o^{-1}\right),$$

$$\mathbf{A}_o := E[\mathbf{H}(\mathbf{w}, \boldsymbol{\theta}_o)],$$

$$\mathbf{B}_o := E\left[\mathbf{s}(\mathbf{w}, \boldsymbol{\theta}_o)\, \mathbf{s}(\mathbf{w}, \boldsymbol{\theta}_o)'\right].$$

Mean Value Theorem

- Normality proof relies on *mean value theorem*.
- Consider *scalar* case ($P = 1$).

Mean Value Theorem (MVT):

- Let $f : [a, b] \to \mathbb{R}$ continuous + differentiable on (a, b).
- Then for some $c \in (a, b)$,

$$f(b) - f(a) = f'(c)(b - a).$$

- Slope of secant attained somewhere in between. [Draw]

Proof Sketch

- In scalar $(P = 1)$ case,

$$s(\mathbf{w}, \theta) = \frac{\partial}{\partial\theta} q(\mathbf{w}, \theta), \quad H(\mathbf{w}, \theta) = \frac{\partial^2}{\partial^2\theta} q(\mathbf{w}, \theta).$$

- Twice cont' diff' + MVT with f = score average,

$$\frac{1}{N}\sum_{i=1}^{N} s(\mathbf{w}_i, \widehat{\theta}) - \frac{1}{N}\sum_{i=1}^{N} s(\mathbf{w}_i, \theta_o) = \frac{1}{N}\sum_{i=1}^{N} H\left(\mathbf{w}_i, \overline{\theta}\right)\left(\widehat{\theta} - \theta_o\right).$$

- $\widehat{\theta} \in \operatorname{int}\Theta$ w.p.a.1. (consistency)

- .... solves SP, so LHS vanishes. (FOC.)

Proof Sketch

Have argued:

$$-\frac{1}{N}\sum_{i=1}^{N} s\left(\mathbf{w}_i, \theta_o\right) = \frac{1}{N}\sum_{i=1}^{N} H\left(\mathbf{w}_i, \overline{\theta}\right)\left(\widehat{\theta} - \theta_o\right).$$

Isolate $\widehat{\theta} - \theta_o$ and $\times\sqrt{N}$:

$$\sqrt{N}(\widehat{\theta} - \theta_o) = \left[-\frac{1}{\sqrt{N}}\sum_{i=1}^{N} s\left(\mathbf{w}_i, \theta_o\right)\right] \Bigg/ \left[\frac{1}{N}\sum_{i=1}^{N} H\left(\mathbf{w}_i, \overline{\theta}\right)\right].$$

Analyze each RHS factor in turn.

Proof Sketch

$$\sqrt{N}(\widehat{\theta}-\theta_o) = \left[-\frac{1}{\sqrt{N}}\sum_{i=1}^{N} s\left(\mathbf{w}_i,\theta_o\right)\right] \Bigg/ \left[\frac{1}{N}\sum_{i=1}^{N} H\left(\mathbf{w}_i,\overline{\theta}\right)\right].$$

- $\overline{\theta}$ trapped between $\widehat{\theta}$ and $\theta_o \Rightarrow \overline{\theta} \rightarrow_p \theta_o$.
- So $N^{-1}\sum_{i=1}^{N} H\left(\mathbf{w}_i,\overline{\theta}\right) \approx N^{-1}\sum_{i=1}^{N} H\left(\mathbf{w}_i,\theta_o\right)$ (ULLN).
- $N^{-1}\sum_{i=1}^{N} H\left(\mathbf{w}_i,\theta_o\right) \rightarrow_p E\left[H\left(\mathbf{w},\theta_o\right)\right] = A_o > 0$ (p.d.),

$$\Rightarrow 1 \Bigg/ \frac{1}{N}\sum_{i=1}^{N} H\left(\mathbf{w}_i,\overline{\theta}\right) \xrightarrow{p} 1/A_o. \qquad \text{(CMT/Slutsky)}$$

Proof Sketch

$$\sqrt{N}(\widehat{\theta} - \theta_o) = \left[-\frac{1}{\sqrt{N}} \sum_{i=1}^{N} s\left(\mathbf{w}_i, \theta_o\right) \right] \Bigg/ \left[\frac{1}{N} \sum_{i=1}^{N} H\left(\mathbf{w}_i, \overline{\theta}\right) \right].$$

- Mean zero scores + CLT ensure

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} s\left(\mathbf{w}_i, \theta_o\right) \xrightarrow{d} \mathrm{N}\left(0, B_o\right), \quad B_o = E[s\left(\mathbf{w}, \theta_o\right)^2].$$

Proof Sketch

Harvesting our results,

$$\sqrt{N}(\widehat{\theta}-\theta_o) = \underbrace{\left[-\frac{1}{\sqrt{N}}\sum_{i=1}^{N} s\left(\mathbf{w}_i,\theta_o\right)\right]}_{\to_d \mathrm{N}(0,B_o)} \Bigg/ \underbrace{\left[\frac{1}{N}\sum_{i=1}^{N} H\left(\mathbf{w}_i,\overline{\theta}\right)\right]}_{\to_p 1/A_o}$$

$$\xrightarrow{d} \mathrm{N}\left(0,B_o\right)/A_o \qquad \text{(product rule/Slutsky)}$$

$$\overset{d}{=} \mathrm{N}\left(0,B_o/A_o^2\right). \qquad A_o^{-1}B_oA_o^{-1}$$

- Vector-case proof follows similarly:
  1. Linear approximation (MVT)
  2. Convergence of inverse Hessian term (ULLN+CMT)
  3. CLT + Product rule.

Discussion

- Thm. gives conditions for *any* M-estimator to be asymptotically normal.
- Implies sandwich form
$$\text{Avar}(\widehat{\boldsymbol{\theta}}) = \mathbf{A}_o^{-1}\mathbf{B}_o\mathbf{A}_o^{-1}/N.$$
- Akin to earlier results (with estimators in closed form).
- Note: $\text{Avar}(\widehat{\boldsymbol{\theta}})$ depends on q.
- We prefer low variance.

Discussion

Q: A_0 must be invertible? How does this work in a HD-setting?

- $\mathbf{A}_o = E\left[\mathbf{H}\left(\mathbf{w}, \boldsymbol{\theta}_o\right)\right]$ assumed positive definite.
- Zero on diagonal $\approx$ infinite variance (through $\mathbf{A}_o^{-1}$)
- Failure of p.d $\approx$ P minimand flat around $\boldsymbol{\theta}_o$
- $\approx$ Identification failure.

Role of Interiority

We used $\boldsymbol{\theta}_o \in \operatorname{int} \Theta$ for differentiation

Q: What if $\boldsymbol{\theta}_o$ on boundary of parameter space?

A: No reason to expect $\sqrt{N}$-asymptotic normality.

Example: Parameter on Boundary

Let $y_i \sim \text{i.i.d.}\,(\theta_o, 1)$ with θ_o known $\geqslant 0$.

Nonnegativity enforced

$$\widehat{\theta} = \operatorname*{argmin}_{\theta \geqslant 0} \frac{1}{N} \sum_{i=1}^{N} (y_i - \theta)^2 = \max(0, \overline{y}),$$

If $\theta_o = 0$ (boundary case), then $\sqrt{N}(\widehat{\theta} - 0) \geqslant 0$.

$\sqrt{N}(\widehat{\theta} - 0)$ does $\to_d$... but not to normal. [Whiteboard]