Lecture 5: Asymptotic Properties & Extremum Estimators

Econometrics 2 — *continuation from Lecture 4*

Instructor: Prof. S. Arvanitis | **Notes transcribed by:** E. Hatija | **Digitisation & added notes:** T. Kourtalis

Semester: Spring 2026

▷ **Amber boxes = Handwritten Notes (professor's words)**

◇ **Teal boxes = Student's Notes**

Recall from Lecture 4

In Lecture 4 we moved from the unobservable population objective $M_n^*(\beta)$ to the observable sample objective $M_n(\beta)$ via the **analogy principle**, defined **extremum (M-) estimators** as $\hat{\beta}_n = \arg\min_{\beta\in\Theta} M_n(\beta)$, and stated the definition of **weak consistency**. This lecture: (i) reviews the three key asymptotic properties in detail, (ii) formalises the extremum-estimator framework, and (iii) discusses practical difficulties in optimisation.

1 Useful General Properties of Estimators

▷ **Handwritten Notes** (what the professor said)

We are concerned with recalling useful general properties of estimators within the framework of (semi-)parametric models:
Sample: $Z_n \to \Theta$ — provided it has a well-defined probability distribution.
In many cases the distribution of the estimator (for a given n) is difficult (in practice impossible) to derive. We usually rely on corresponding **asymptotic properties**.

◇ **Student's Notes**

Why asymptotic theory?
For most estimators beyond simple cases (e.g. OLS with normal errors), the *exact* finite-sample distribution of $\hat{\beta}_n$ is unknown or intractable. Asymptotic theory gives us *approximate* answers that become exact as $n \to \infty$.
The three asymptotic properties we study form a hierarchy:

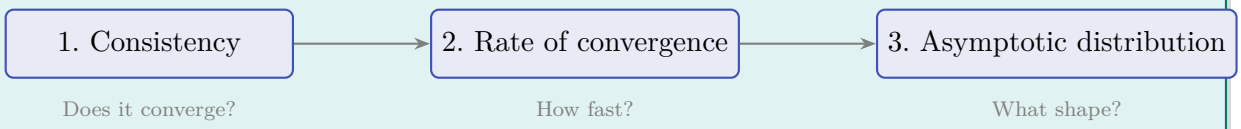

Figure 1: The three asymptotic properties build on each other. Each requires the previous one.

2 Property 1: Weak Consistency

▷ Handwritten Notes (what the professor said)

A first such property is **weak consistency**. The estimator $\hat{\beta}_n$ is weakly consistent iff:

$$\hat{\beta}_n \xrightarrow{p} \beta_0$$

Roughly, this means the probability that the distance between $\hat{\beta}_n$ and β_0 is positive converges to 0.
In most of the cases we will examine, consistency will follow from the appropriate convergence of the criterion to some asymptotic criterion, which under some condition of **asymptotic identification** will be uniquely minimised at β_0. Useful tools in such cases will be the **Laws of Large Numbers**.

Definition: Weak Consistency (restated)

$\hat{\beta}_n$ is **weakly consistent** for β_0 iff for every $\varepsilon > 0$:

$$\lim_{n\to\infty} P\big(\|\hat{\beta}_n - \beta_0\| > \varepsilon\big) = 0 \qquad \Longleftrightarrow \qquad \hat{\beta}_n \xrightarrow{p} \beta_0.$$

◇ Student's Notes

The "recipe" for proving consistency of an M-estimator:

1. Show that $M_n(\beta) \xrightarrow{p} M^*(\beta)$ for all $\beta \in \Theta$ (often *uniformly*), typically via a Law of Large Numbers.
2. Show that $M^*(\beta)$ has a *unique* minimum at β_0 (asymptotic identification).
3. Conclude $\hat{\beta}_n \xrightarrow{p} \beta_0$.

Key distinction: "asymptotic identification" (the limiting criterion M^* has a unique minimum) versus the finite-sample identification we discussed in Lecture 4 ($\text{rank}(X_n) = p$). Asymptotic identification is the relevant condition for consistency.

Which LLN? The choice depends on the dependence structure:

Data structure	Applicable LLN
i.i.d. data	Khintchine's (weak) LLN
Independent, not identical	Chebyshev's LLN
Dependent (e.g. time series)	Ergodic theorem / mixing LLN

3 Property 2: Rate of Convergence

▷ Handwritten Notes (what the professor said)

A second asymptotic property that will interest us to examine, given weak consistency, is the **rate of convergence**. It represents a "sense" of speed with which $\hat{\beta}_n$ converges to β_0.
The rate of convergence will be a real sequence with term $F(n)$ where $F(n) \to +\infty$ as $n \to +\infty$ with the property that:

$$F(n)\,(\hat{\beta}_n - \beta_0) \quad \text{asymptotically stabilises to a well-defined random vector.}$$

Usually — and certainly in what we do — $F(n) = \sqrt{n}$, because of the operation of some **Central Limit Theorem**.

Definition: Rate of Convergence

Given that $\hat{\beta}_n \xrightarrow{p} \beta_0$, the **rate of convergence** is a sequence $F(n) \to +\infty$ such that

$$F(n)\,(\hat{\beta}_n - \beta_0) = O_p(1),$$

i.e. the rescaled estimation error is *bounded in probability* (it converges in distribution to some well-defined random vector).

◇ Student's Notes

Intuition: $\hat{\beta}_n - \beta_0$ is shrinking to zero. The rate $F(n)$ tells us *how fast*. Multiplying by $F(n)$ "blows up" the shrinking error just enough to keep it from collapsing to zero or exploding to infinity.

Why $\sqrt{n}$?
The CLT says that sample averages of i.i.d. random variables fluctuate on the scale of $1/\sqrt{n}$. Since most M-estimators are implicitly defined through sample averages (think of $M_n(\beta) = \frac{1}{n}\sum m(z_i, \beta)$), their estimation error inherits this $1/\sqrt{n}$ scale. Hence $F(n) = \sqrt{n}$ is the "standard" rate.

Faster or slower rates:

Rate $F(n)$	Speed	When it arises
$\sqrt{n}$	Standard ("$\sqrt{n}$-consistent")	Regular parametric models
n	Super-fast	Unit root / explosive AR models
$n^{1/3}$, $n^{2/5}$	Slower	Nonparametric / semiparametric estimators

In this course we work exclusively with $\sqrt{n}$-consistent estimators.

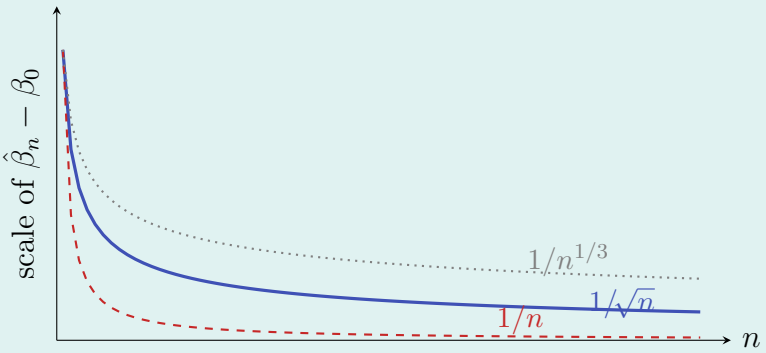

Figure 2: How fast the estimation error shrinks under different rates. $1/\sqrt{n}$ (standard) is between the super-fast $1/n$ and the slow nonparametric rate $1/n^{1/3}$.

4 Property 3: Asymptotic Distribution

▷ **Handwritten Notes** (what the professor said)

The third concept that will concern us is that of the **asymptotic distribution**. Given weak consistency and the rate of convergence $F(n)$, we will have:

$$F(n)\,(\hat{\beta}_n - \beta_0) \xrightarrow{d} Z \sim \text{some well-defined probability distribution}$$

(convergence in distribution).
In what we will see below, as we said previously:

$$Z \sim N(0_{p\times 1},\ V)$$

where V is a $p \times p$ variance–covariance matrix.
So we will have:

$$\sqrt{n}\,(\hat{\beta}_n - \beta_0) \xrightarrow{d} N(0_{p\times 1},\ V)$$

and V will be called the **asymptotic variance matrix** of $\hat{\beta}_n$.
To derive the rate of convergence, asymptotic normality, and V, useful tools will be the Laws of Large Numbers, the Central Limit Theorem, and the **Taylor expansion**.

Key Result

For the estimators in this course, the fundamental asymptotic result takes the form:

$$\boxed{\sqrt{n}\,(\hat{\beta}_n - \beta_0) \xrightarrow{d} N(0,\ V)}$$

where V is the **asymptotic variance matrix** ($p \times p$, symmetric, positive semi-definite).

Practical consequence: for large n,

$$\hat{\beta}_n \overset{\text{approx.}}{\sim} N\left(\beta_0,\ \frac{V}{n}\right).$$

This is the basis for confidence intervals and hypothesis tests.

◇ Student's Notes

The three tools and where they enter:

Tool	Role in deriving asymptotic normality
Law of Large Numbers	Ensures $M_n \to M^*$ and various sample moments converge to their population counterparts
Central Limit Theorem	Gives the $\sqrt{n}$ rate and the normal limiting distribution of the "score" or gradient term
Taylor expansion	Linearises the first-order condition around β_0 to express $\sqrt{n}(\hat{\beta}_n - \beta_0)$ as a function of sample averages

Preview of the argument (we will formalise this later):

1. The FOC of M_n is $\nabla M_n(\hat{\beta}_n) = 0$.
2. Taylor-expand around β_0: $0 = \nabla M_n(\beta_0) + \nabla^2 M_n(\bar{\beta})\,(\hat{\beta}_n - \beta_0)$.
3. Rearrange: $\sqrt{n}(\hat{\beta}_n - \beta_0) = -[\nabla^2 M_n(\bar{\beta})]^{-1}\,\sqrt{n}\,\nabla M_n(\beta_0)$.
4. By CLT: $\sqrt{n}\,\nabla M_n(\beta_0) \xrightarrow{d} N(0, \Sigma)$.
5. By LLN: $\nabla^2 M_n(\bar{\beta}) \xrightarrow{p} H$ (Hessian).
6. Therefore: $\sqrt{n}(\hat{\beta}_n - \beta_0) \xrightarrow{d} N(0,\ H^{-1}\Sigma H^{-1})$, so $V = H^{-1}\Sigma H^{-1}$.

This is the famous "sandwich" form of the asymptotic variance.

5 Extremum (M-) Estimators: Formal Framework

▷ **Handwritten Notes** (what the professor said)

We work within the framework of some (semi-)parametric model with parameter space $\Theta \subseteq \mathbb{R}^p$. (Goal: finding β_0.)
Usually we will assume the model is well-specified (we will also examine some cases where this does not hold).
We have at our disposal a function $M_n(\beta)$ — it is observable and "approximates" (depends also on the sample) a non-observable function (say $M_n^*(\beta)$) that is uniquely minimised at β_0.
The **extremum estimator** $\hat{\beta}_n$ (based on the above) is defined as:

$$\hat{\beta}_n \in \arg\min_{\beta \in \Theta} M_n(\beta)$$

This yields a function $Z_n \to \Theta$ since:

$$Z \xrightarrow{\text{compute}} M_n(\beta) \text{ as a function of } \beta \xrightarrow{\text{minimise over } \beta \in \Theta} \hat{\beta}_n$$

◇ **Student's Notes**

Visualising the estimation pipeline:

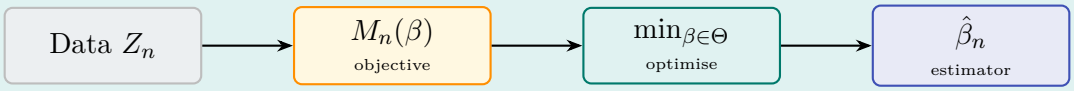

Figure 3: The extremum-estimator pipeline. Data → objective → optimisation → estimator.

Important: the "$\in$" in the definition ($\hat{\beta}_n \in \arg\min$) rather than "$=$" is deliberate. The $\arg\min$ could be a set (possibly empty, or with multiple elements). The practical issues that arise from this are discussed next.

6 Practical Difficulties in Minimisation

▷ **Handwritten Notes** (what the professor said)

$\arg\min_{\beta \in \Theta} M_n(\beta) \in \Theta$
In this case, practically, we may encounter various difficulties, e.g.:
(α) It is possible that for "many values" of Z_n, the $\arg\min_{\beta \in \Theta} M_n(\beta)$ is *empty*. Usually this will not happen if, e.g., Θ is convex and M_n is a convex function of β for most values that Z_n can take.

(β) It is possible that $\arg\min_{\beta\in\Theta} M_n(\beta)$ has more than one element, so we will need to choose (we will not deal directly with such cases).
(γ) In many cases it will be practically impossible to analytically optimise M_n. It will have to be done numerically. The choice of the corresponding algorithm may matter for the properties of the estimator (e.g. linear programming, or Newton–Raphson type algorithms that use derivatives, etc.). ← **Most important!!**

◇ Student's Notes

Summary of the three difficulties:

	Problem	When it is avoided
(α)	arg min is empty (no minimiser exists)	Θ compact, or Θ convex + M_n convex + coercive
(β)	arg min has multiple elements	M_n strictly convex (guarantees uniqueness)
(γ)	No closed-form solution	Special structure (e.g. OLS); otherwise use numerical methods

! Watch Out

Problem (γ) is the most common in practice.
Most nonlinear models (MLE, nonlinear least squares, GMM with nonlinear moment conditions) require numerical optimisation. Choices matter:

- **Newton–Raphson:** uses first and second derivatives; fast convergence near the optimum but can fail with bad starting values.
- **BFGS / Quasi-Newton:** approximates the Hessian; more robust than pure Newton–Raphson.
- **Gradient descent:** uses only first derivatives; slower but simpler.
- **Simplex (Nelder–Mead):** derivative-free; useful when M_n is not smooth.
- **Linear programming:** for special structures (e.g. quantile regression / LAD).

Practical tip: always try multiple starting values to check whether the algorithm converges to the same point. If it does not, you may have multiple local minima.

7 Example: The Semi-Parametric Linear Model

▷ **Handwritten Notes** (what the professor said)

As we have developed it:

$$M_n(\beta) = \frac{1}{n}(Y_n - X_n\beta)'(Y_n - X_n\beta)$$

And therefore the extremum estimator that results from minimising the above over Θ is the **Ordinary Least Squares Estimator (OLSE)**:

$$\hat{\beta} \in \arg\min_{\beta\in\Theta} \; M_n(\beta).$$

When $\Theta = \mathbb{R}^p$ the problem has an analytical solution. And we know that:

$$\hat{\beta}_n = (X_n'X_n)^{-1}X_n'Y_n$$

(Note: It is required that the matrix $(X_n'X_n)$ be invertible. If rank $< p \to$ infinite solutions; we choose the optimal β.)

◇ **Student's Notes**

Deriving the OLS formula (for completeness):

$$M_n(\beta) = \frac{1}{n}(Y - X\beta)'(Y - X\beta) = \frac{1}{n}\big(Y'Y - 2\beta'X'Y + \beta'X'X\beta\big).$$

FOC: $\dfrac{\partial M_n}{\partial \beta} = \dfrac{1}{n}(-2X'Y + 2X'X\beta) = 0$, which gives $X'X\hat{\beta} = X'Y$, hence:

$$\boxed{\hat{\beta}_n = (X'X)^{-1}X'Y} \qquad \text{when } \operatorname{rank}(X) = p.$$

SOC: $\dfrac{\partial^2 M_n}{\partial\beta\,\partial\beta'} = \dfrac{2}{n}X'X$, which is positive definite when $\operatorname{rank}(X) = p$, confirming this is a minimum.

This is one of the rare cases where (γ) does *not* apply: we have a closed-form solution.

7.1 What Happens When $\Theta \subset \mathbb{R}^p$?

▷ **Handwritten Notes** (what the professor said)

(Ridge Regression): What happens when $\Theta \subset \mathbb{R}^p$ with the optimisation?

◇ Student's Notes

When $\Theta \neq \mathbb{R}^p$, the unconstrained OLS solution may lie *outside* Θ, and we need **constrained optimisation**.

Ridge regression is a leading example. Here we add a penalty that effectively restricts the parameter space:

$$\hat{\beta}_{\text{ridge}} = \arg\min_{\beta} \frac{1}{n}\|Y - X\beta\|^2 + \lambda\|\beta\|^2, \qquad \lambda > 0.$$

This is equivalent to minimising the OLS objective over the constraint set $\Theta = \{\beta : \|\beta\|^2 \leq c\}$ for some $c = c(\lambda)$.

Closed-form solution:

$$\boxed{\hat{\beta}_{\text{ridge}} = (X'X + n\lambda I_p)^{-1} X'Y.}$$

Key observations:

- The matrix $(X'X + n\lambda I_p)$ is *always* invertible (even when $\text{rank}(X) < p$), because adding $n\lambda I_p$ makes all eigenvalues strictly positive.
- Ridge regression *trades bias for variance*: the estimator is biased (it shrinks $\hat{\beta}$ towards zero), but has lower variance than OLS, especially when regressors are nearly collinear.
- As $\lambda \to 0$, we recover OLS; as $\lambda \to \infty$, $\hat{\beta}_{\text{ridge}} \to 0$.

Connection to Lecture 4 (misspecification):
If the true β_0 has large norm but we constrain to $\Theta = \{\|\beta\| \leq c\}$ with $c < \|\beta_0\|$, then $\beta_0 \notin \Theta$ and we are in the "mild misspecification" setting. The ridge estimator converges to the projection of β_0 onto Θ, not to β_0 itself—it is *inconsistent* but may have lower MSE for finite n.

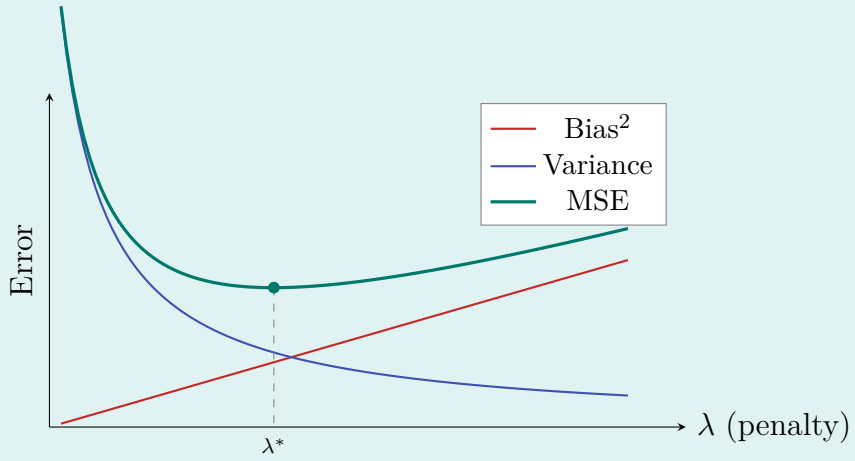

Figure 4: The bias–variance trade-off in ridge regression. The optimal λ^* minimises the total MSE.

Quick-Reference Summary

◇ Student's Notes

Lecture 5 narrative arc:

Topic	What was accomplished
Asymptotic properties	Defined the three-step hierarchy: consistency → rate → asymptotic distribution
Weak consistency	$\hat{\beta}_n \xrightarrow{p} \beta_0$; proved via LLN + asymptotic identification
Rate of convergence	Usually $\sqrt{n}$; comes from CLT
Asymptotic distribution	$\sqrt{n}(\hat{\beta}_n - \beta_0) \xrightarrow{d} N(0, V)$; derived via Taylor + CLT + LLN
Extremum estimators	Formal definition: $\hat{\beta}_n \in \arg\min M_n(\beta)$
Practical difficulties	Empty arg min, non-uniqueness, numerical optimisation
Linear model	OLS as the closed-form M-estimator; Ridge when $\Theta \subset \mathbb{R}^p$

Key new concepts:

Term	One-line meaning
Rate of convergence $F(n)$	Speed at which $\hat{\beta}_n \to \beta_0$; usually $\sqrt{n}$
Asymptotic variance V	Variance of the limiting distribution of $\sqrt{n}(\hat{\beta}_n - \beta_0)$
Sandwich form	$V = H^{-1}\Sigma H^{-1}$ (Hessian + score variance)
Asymptotic identification	$M^*(\beta)$ has a unique minimum at β_0
Ridge regression	OLS + ℓ_2 penalty; always invertible; biased but lower variance

Figures summary:

Figure	What it shows
Fig. 1	Hierarchy: consistency → rate → distribution
Fig. 2	Convergence rates $1/\sqrt{n}$ vs $1/n$ vs $1/n^{1/3}$
Fig. 3	Extremum-estimator pipeline: data → objective → optimise → estimator
Fig. 4	Bias–variance trade-off in ridge regression