Analogy Principle
Asymptotic Theory – Part II

James J. Heckman
University of Chicago

Econ 312
This draft, April 5, 2006
Consider four methods:

1. Maximum Likelihood Estimation (MLE)
2. (Nonlinear) Least Squares
3. Methods of Moments
4. Generalized Method of Moments (GMM)

These methods are not always distinct for a particular problem. Consider the classical normal linear regression model:

$$Y_t = X_t\beta + U_t$$
Under standard assumptions

1. \( U_t \sim N(0, \sigma_u^2); U_t \) i.i.d.;

2. \( X_t \) non-stochastic; and

3. \( \frac{\sum X_t X_{t'}}{T} \) full rank.

OLS is all four rolled into one.

In this lecture we will show how one basic principle — the Analogy Principle — underlies all of these modern econometric methods.
1 Analogy Principle: The Large Sample Version

This is originally due to Karl Pearson or Goldberger.

The intuition behind the analogy principle is as follows:

‘Suppose we know some properties that are satisfied for the “true parameter" in the population. If we can find a parameter value in the sample that causes the sample to mimic the properties of the population, we might use this parameter value to estimate the true parameter.’
The main points of this principle are set out in §1.1. The conditions for consistency of the estimator are discussed in §1.2. In §1.3, an abstract example with a finite parameter set illustrates the application of the analog principle and the role of regularity conditions in ensuring consistency of the estimator.
1.1 Outline of the Analog Principle

The ideas constituting the analog principle can be grouped under four steps:

1. *The model.* Suppose there exists a ‘true model’ in the population — $M(Y, X, \theta_0) = 0$ — an implicit equation. $\theta_0$ is a true parameter vector, and the model is defined for other values (at least some other) of $\theta \in \Theta$.

2. *Criterion function.* Based on the model, construct a ‘criterion function’ of model and data:

$$Q(Y, X, \theta)$$

which has property $P$ in the population when $\theta = \theta_0$ (the ‘true’ parameter vector).
3. Analog in sample. In the sample, construct an ‘analog’ to $Q$, $Q_T(Y, X, \theta)$, which has the following properties:

(a) $(\forall \theta \in \Theta) \quad Q_T(Y, X, \theta) \xrightarrow{\text{a.s. unif.}} Q(Y, X, \theta)$, where $T$ is the sample size; and

(b) $Q_T(Y, X, \theta)$ mimics in sample properties of $Q(Y, X, \theta)$\(^1\) in the population.

4. The estimator. Let $\hat{\theta}_T$ be the estimator of $\theta_0$ in sample $T$ formed from the analog in sample, which causes sample $Q_T(.)$ to have the property $P$.

\(^1\)Hereafter, we shall suppress the dependence of the criterion function $Q$ and the sample analog $Q_T$ on the data $(Y, X)$ for notational convenience.
1.2 Consistency and Regularity

Definition 1 (Consistency) We say that an estimator $\hat{\pi}_T$ is a consistent estimator of parameter $\pi_0$ if

$$\hat{\pi}_T \xrightarrow{p} \pi_0.$$ 

In general, to prove that the estimator formed using the analog principle $\hat{\theta}_T$ is consistent, we need to make the following two assumptions, which are referred to as the regularity conditions:

1. Identification condition. Generally this states that only $\theta_0$ causes $Q$ to possess property $P$, at least in some neighborhood of $\theta_0$; i.e. $\theta_0$ must be at least locally identified.
2. Uniform convergence of the analog function. We need the sample analog $Q_T(\hat{\theta}_t)$ to converge ‘nicely’. In particular, we need the convergence of the sample analog criterion function to the criterion function $-(\forall \theta \in \Theta) \quad Q_T(\theta) \rightarrow Q(\theta)$ — to be uniform.
The first condition ensures that $\theta_0$ can be identified. If there is more than one value of $\theta$ that causes $Q$ to have property $P$, then we cannot be sure that only one value of $\theta$ will cause $Q_T$ to assume property $P$ in the sample. In this case we may not be able to determine what $\hat{\theta}_T$ estimates.

The second condition is a technical one that ensures that if $Q_T$ has a property such as continuity or differentiability, this property is inherited by $Q$.

Specific proofs of consistency depend on which property we suppose $Q$ has when $\theta = \theta_0$. 
1.3 An Abstract Example

The intuition underlying the analog principle method and the role of the regularity conditions is illustrated in the simple (non-stochastic) example below.

• Assume an abstract model $M(Y, X, \theta_0)$ which leads to a criterion function $Q(\theta)$ which has the property

$$P := Q(\theta) \text{ is maximized at } \theta = \theta_0 \text{ (in population)}.$$ 

• Construct a sample analog of the criterion function $Q_T$ such that

$$(\forall \theta \in \Theta) \quad Q_T(\theta) \rightarrow Q.$$
• Select $\hat{\theta}_T$ to maximize $Q_T(\theta)$ for each $T$. Then, if convergence is ‘OK’ (we get that under the regularity conditions), we have convergence in the sample to the maximizer in the population; i.e. we have $\hat{\theta} \to \theta$ as $T \to \infty$. 
Now suppose $\Theta = \{1, 2, 3\}$ is a finite set, so that $\hat{\theta}$ assumes only one of three values. Then under regularity condition 1 (q.v. §1.2), $Q(\theta)$ is maximized at one of these. Further, by construction we have

$$\left( \forall \theta \in \Theta = \{1, 2, 3\} \right) \quad Q_T(\theta) \to Q(\theta).$$

Note that the rule picks

$$\hat{\theta}_T = \arg\max Q_T(\theta).$$

This estimator must work well (i.e. be consistent for $\theta$ for ‘big enough’ $T$, because $Q_T \to Q$ for each $\theta$. Why? We loosely sketch a proof of the consistency of $\hat{\theta}$ for $\theta$ below.
Say $\theta_0 = 2$ so that $Q(\theta)$ is maximized at $\theta = 2$, $Q(2) > Q(1)$, and $Q(2) > Q(3)$. Now

$$(\forall \theta \in \Theta) \quad Q_T(\theta) \to Q(\theta).$$

This implies that as $T$ gets large,

$$(\forall T) \quad ||Q_T(\theta) - Q(\theta)|| \to 0.$$

In other words, as $T$ gets ‘big’ we get $Q_T(\theta)$ arbitrarily close to $Q(\theta)$. 


Now suppose that even for very large \( T \), \( Q_T(\theta) \) is not maximized at 2 but say at 1. Then we have

\[
(\forall T) \quad Q_T(1) - Q_T(2) > 0.
\]

Then under regularity condition 2 (q.v. §1.2) this would imply \( Q(1) - Q(2) > 0 \), a contradiction. Hence the estimator \( \hat{\theta}_T \rightarrow \theta \).

Here principle \( P \) is an example of the extremum principle. This principle involves either maximization or minimization of the criterion function. Examples include the maximum likelihood (maximization) and nonlinear least squares (minimization) methods.
2 Overview of Convergence Concepts for Non-stochastic Functions

Most estimators involve forming functions using the data available. These functions of data can be viewed as sequences of functions, with the index being the number of data points available.

With sufficient data (large samples), these functions of sample data converge ‘nicely’ (assuming certain conditions), ensuring that the estimators we form have good properties, such as ‘consistency’ and ‘asymptotic normality’.
In this section, we examine certain basic concepts about convergence of sequences of functions. The key idea here is that of uniform convergence; we shall broadly motivate why this notion of convergence is required for us.

For simplicity, we look at non-stochastic functions and non-stochastic convergence notions. Broadly construed, the same ideas also apply to stochastic functions—analogous convergence concepts are applicable in the stochastic case.
2.1 Pointwise Convergence

Let $F_n : S \rightarrow R$ be a sequence of real valued functions. For each $x \in S$ form a sequence $\{F_n(x)\}_{n=1}^{\infty}$. Let $B \in S$ be the set of points for which $F_n(x)$ converges.

$$(\forall x \in B) \quad F(x) := \lim_{n \rightarrow \infty} F_n(x)$$

We then call $F(x)$ a ‘limit function’, and say ‘$\{F_n(x)\}_{n=1}^{\infty}$ converges pointwise to $F(x)$ on $B$’.

_Nota bene_, pointwise convergence is _not_ enough to guarantee that if $F_n$ has a property (continuity, differentiability, etc.) that that property is inherited by $F$. 
Inheritance requires *uniform convergence*. Usually, this is a sufficient condition for the properties of $F_n$ being shared by the limit function. Some of the properties we will be interested in are summarized below:

- Does ‘$(\forall n) \ F_n(x) \text{ is continuous}$’ $\implies$ ‘$F(x) \text{ is continuous}$’?
- Does ‘$(\forall n) \ \lim_{x \to x_0} F_n(x) = F_n(x_0)$’ $\implies$ ‘$\lim_{x \to x_0} F(x) = F(x_0)$’?
  
  I.e. does $\lim_{x \to x_0} \lim_{n \to \infty} F_n(x) = \lim_{n \to \infty} \lim_{x \to x_0} F_n(x)$?
- Does $\lim_{n \to \infty} \int F_n(x) = \int \lim_{n \to \infty} F_n(x) = \int F(x)$ (where $F(x)$ is the limit function)?

The answer to all three questions is: *No, not in general.* The *pointwise* convergence of $F_n(x)$ to $F(x)$ is generally not suffi-
cient to guarantee that these properties hold. This is demonstrated by the examples which follow.
Example 1
Consider $F_n(x) = x^n$ ($0 \leq x \leq 1$).
Here, $F_n(x)$ is continuous for every $n$, but the limit function is:

$$F(x) = \begin{cases} 
0 & \text{if } 0 \leq x < 1 \\
1 & \text{if } 0 \leq x = 1 
\end{cases}$$

This is discontinuous at $x = 1$. 
Example 2
Consider

\[ F_n(x) = \frac{x}{x + n} \quad (x \in \mathbb{R}). \]

Then the limit function is \( F(x) = \lim_{n \to \infty} F_n(x) = 0 \) for each fixed \( x \). This implies that \( \lim_{x \to \infty} \lim_{n \to \infty} F_n(x) = 0 \).

But we have \( \lim_{x \to \infty} F_n(x) = 1 \) for every fixed \( n \). This implies that \( \lim_{n \to \infty} \lim_{x \to \infty} F_n(x) = 1 \neq \lim_{x \to \infty} \lim_{n \to \infty} F_n(x) = 0 \).
Example 3

Consider $F_n(x) = nx(1 - x^2)^n$ ($0 \leq x \leq 1$). Then the limit function is

$$F(x) = \lim_{n \to \infty} F_n(x) = 0$$

for each fixed $x$. This implies that $\int_0^1 F(x) = 0$. But we also get

$$\lim_{n \to \infty} \int_0^1 F_n(x) = \lim_{n \to \infty} \frac{n}{2n + 2} = \frac{1}{2}.$$  

---

$^2$See Rudin, Chapter 3 for concepts relating to convergence of sequences.
2.2 Uniform Convergence

A sequence of functions \( \{F_n\} \) converges uniformly to \( F \) on \( B \) if

\[
(\forall \varepsilon > 0)(\exists N)(\forall n > N)(\forall x \in B) \quad |F_n(x) - F(x)| < \varepsilon
\]

where \( N \) depends on \( \varepsilon \) but not on \( x \), i.e.,

\[
(\forall x \in B) \quad F(x) - \varepsilon < F_n(x) < F(x) + \varepsilon.
\]

Intuitively, for any \( x \in B \) as \( n \) gets large enough, \( F_n(x) \) lies in a band of uniform thickness around the limit function \( F(x) \).
Note that the convergence displayed in the examples in §2.1 did not satisfy the notion of ‘uniform’ convergence.

We have theorems that ensure that the properties of inheritance referred to in §2.1 are satisfied when convergence is uniform.³

³See Rudin, Chapter 7. Theorems 7.12, 7.11, and 7.16 respectively ensure that the three properties listed in §2.1 hold.
3 Applications of the Analog Principle
3.1 Moment Principle

In this case, the criterion function \( Q \) is an equation connecting population moments and \( \theta_0 \).

\[
Q = Q(\text{population moments of } [y, x], \theta) = 0 \text{ at } \theta = \theta_0. \tag{1}
\]

Thus, the property here is not maximization or minimization (as in the extremum principle), but setting some function of the population moments (generally) to zero at \( \theta = \theta_0 \), so that solving equation 1 we get the estimator:

\[
\hat{\theta}_T = \theta(\text{sample moments } [y,x])
\]
When forming the estimator ‘analog’ in the sample, we can consider two cases.

3.1.1 Case A

Here we solve for $\theta_0$ in the population equation, obtaining

$$\theta_0 = \theta_0(\text{population moments } [y,x]).$$

From this equation we construct the simple analog

$$\hat{\theta} = \hat{\theta}(\text{population moments } [y,x]).$$

Given regularity condition 1 (q.v. §1.2), we get $\hat{\theta}_T \xrightarrow{p} \theta_0$. Note that we do not need regularity condition 2.
Example 4 (OLS)  

1. The model.

\[ Y_t = X_t \beta_0 + U_t \]

\[ E(U_t) = 0 \]

\[ E(X'_t U_t) = 0 \]

\[ \text{Var}(U_t) = \sigma_u^2 \]

\[ E(X'_t X_t) = \sum_{xx} \text{positive definite} \quad E(X'_t Y_t) = \sum_x y \]

2. Moment principle (criterion function). In the population

\[ E(X'_t U_t) = 0 \]

\[ \text{or} \]

\[ X'_t y_t = X'_t X_t \beta_0 + X'_t U_t \]

\[ \Rightarrow Q : \quad \sum_{xy} = \sum_{xx} \beta_0 + 0 \]

\[ \Rightarrow \]

\[ \beta_0 = \sum_{xx}^{-1} \sum_{xy}. \]
3. Analog in sample.

\[
\hat{\beta}_T = \left( \frac{\sum X_t'X_t}{T} \right)^{-1} \left( \frac{\sum X_tY_t}{T} \right)
\]

Now, r.h.s. → \( (\sum_{xx})^{-1} \sum_{xy} \). Thus, \( \hat{\beta} \overset{p}{\to} \beta_0 \).
3.1.2 Case B

Here we form the criterion function, form the sample analog for the criterion function, and then solve for the estimator.

We require condition 1 (q.v. §1.2) and uniform convergence (condition 2) to get consistency of the estimator — i.e. for $\hat{\theta} \xrightarrow{p} \theta_0$.

Example 5 (OLS — another analogy)

1. The model. As in §3.1.1.

2. Moment principle (criterion function). In the population $Q : E(X_i'Ut) = 0$. 
3. Analog in sample. Here we define an ample analog of $U$:

$$\hat{U} = y - x\hat{\beta}$$

This mimics $U$ in the criterion function, so that we can form the sample analog of the criterion function by substituting $\hat{U}$ for $U$ in the expression for $Q$ above:

$$Q_T = \frac{1}{T} \sum x_t(Y_t - X_t\hat{\beta}) = 0$$

4. The estimator. Here we pick $\hat{\beta}$ by solving from $Q_T$ to arrive at the relation

$$\frac{1}{T} \sum X_tY_t = \left( \frac{\sum X_tY_t}{T} \right) \hat{\beta}.$$
Remarks.

- Recall that in Case A (q.v. §3.1.1) we did not form $Q_T$, but first solved for $\theta_0$ and then formed $\hat{\theta}$ directly, skipping step 3 of Case B.

- Observe that when $X_t = 1$,

\[
\hat{\beta} = \frac{\sum Y_t}{T}
\]

is the mean.

- OLS can also be interpreted as an application of the extremum principle (s.v. §3.2).
3.2 Extremum Principle

We saw in §1.3 that under the extremum principle, property P provides that $Q(\theta)$ achieves a maximum or minimum at $\theta = \theta_0$ in the population.

This principle underlies the nonlinear least squares (NLS) and maximum likelihood estimation (MLE) methods. As their names suggest, MLE choose $Q$ to be a maximum while NLS choose $P$ to be the minimum.
The OLS estimator can be seen as a special case of the NLS estimators, and can be viewed as an extremum estimator.

In this section we analyze OLS and NLS as extremum estimators. MLE is examined in more detail in §4.
Example 1 (OLS as an extremum principle)

1. The model. We assume the true model

\[ Y_t = X_t\beta_0 + U_T \]

\[ Y_T = X_t\beta + (X_t(\beta_0 - \beta) + U_t) \]

\[ (Y_t - X_t\beta)'(Y_t - X_t\beta) = (\beta_0 - \beta)'X'X_t(\beta - \beta_0) + 2U'X_t(\beta - \beta_0) + U_tU_t. \]

2. Criterion function.

\[ Q = E((Y_t - X_t\beta)'(Y_t - X_t\beta)) \]

From model assumptions, we have \( E(X_t'U_t) = 0 \). Thus

\[ Q = (\beta - \beta_0)' \sum_{xx} (\beta - \beta_0) + \sigma_u^2. \]
So $Q$ is minimized (with respect to $\beta$) at $\beta = \beta_0$.

Here, then, $Q$ is an example of the extremum principle. It is minimized when $\beta = \beta_0$ (the true parameter vector).

3. Analog in sample. A sample analog of the criterion function is constructed as

$$Q_T = \frac{1}{T} \sum_{t=1}^{T} (Y_t - X_t \beta)'(Y_t - X_t' \beta).$$
We can show that this analog satisfies the key requirement of the analog principle:

\[
\operatorname{plim} Q_T = \operatorname{plim} \left\{ \frac{1}{T} \sum_{t=1}^{T} (Y_t - X_t \beta)'(Y_t - X_t' \beta) \right\} = (\beta - \beta_0)' \sum_{xx} (\beta - \beta_0) + \sigma_U^2 = Q
\]

(Assuming conditions for application of some LLN are satisfied; q.v. Part I of these lectures.)

4. The estimator. We pick \( \hat{\beta} \) to minimize \( Q_T \). Under standard regularity conditions, we can show that we get contradiction unless \( \hat{\beta} \rightarrow \beta_0 \).
Example 2 (NLS as an extremum principle)

1. The model. We assume that the following model holds in the population:

\[ Y_t = g(X_t; \theta_0) + U_t \quad \text{(non-linear model)} \]

\[ \implies Y_T = g(X_t; \theta) + (g(X_t; \theta_0) - g(X_t; \theta)) + U_t. \]

Assume \((U_t, X_t, Y_t)\) i.i.d. Then \(U_t \perp X_t\) implies that

\[(\forall \theta) \ U_t \perp g(X_t; \theta). \]
2. Criterion function. Choose the criterion function as

\[ Q = E(Y_t - g(X_t; \theta))^2 = E(g(X_t; \theta_0) - g(X_t; \theta))^2 + \sigma^2_U. \]

Then \( Q \) is minimized at \( \theta = \theta_0 \) (a true parameter value). If \( \theta = \theta_0 \) is the only such value, the model is identified with respect to the \( Q \) criterion — regularity condition 1 (q.v. §1.2) is satisfied.

3. Analog in sample.

\[ Q_T(\theta) := \frac{1}{T} \sum_{t=1}^{T} (Y_t - g(X_t; \theta))^2 \]

As in the OLS case, we can show that \( \text{plim} \ Q_T = Q \).
4. The estimator. We construct the NLS estimator as

$$\hat{\theta}_T = \text{argmin} \ Q_T(\theta).$$

We thus choose $\hat{\theta}$ to minimize $Q_T(\theta)$. *Reductio ad absurdum* verifies that

$$Q_T(\theta_0) \rightarrow Q(\theta_0)$$

and

$$(\forall \theta \in \Theta) \ Q_T(\theta) \rightarrow Q(\theta) \implies \hat{\theta}_T \rightarrow \theta.$$
Remark. The NLS estimator could also be derived as a moment estimator, just as in the OLS example (q.v. §3.1).

1. The model. Same as the non-linear model above. We have \( U_t \perp g(X_t; \theta) \).

2. Criterion function. \( Q = \mathbb{E}(U_t \cdot g(X_t; \theta)) = 0 \). Note that this is only one implication of \( \perp \). We may now write

\[
Y_t - g(X_t; \theta) = U_t \\
\implies Q = \mathbb{E}(Y_t - g(X_t; \theta) \cdot g(X_t; \theta)) = 0
\]

3. Analog in sample.

\[
Q_T := \frac{1}{T} \sum_{t=1}^{T} (Y_t - g(X_t; \theta)) \cdot g(X_t; \theta)
\]
4. *The estimator.* Find $\theta$ which sets $Q_T = 0$ (or as close to zero as possible).

## 4 Maximum Likelihood

The maximum likelihood estimation (MLE) method is an example of the extremum principle. In this section, we look at the ideas underlying MLE and examine the regularity conditions and convergence notions in more detail for this estimator.
4.1 The Model

Suppose that the joint density of data is

\[ f(y_t, x_t; \theta_0) \cdot f(y_t \mid x_t; \theta_0) \cdot f(x_t). \]

Assume that \( x_t \) is ‘exogenous’ — i.e. the density of \( x_t \) is uninformative about \( \theta_0 \). Also assume random sampling. We arrive at the likelihood function

\[ \mathcal{L} = \prod_{t=1}^{T} f(y_t, x_t; \theta_0). \]

Taking \( (y_t, x_t) \) as data, \( \mathcal{L} \) becomes a function of \( \theta_0 \). The log likelihood function is

\[ \ln \mathcal{L} = \sum_{t=1}^{T} \ln f(y_t, x_t; \theta) = \sum_{t=1}^{T} \ln f(y_t \mid x_t; \theta) + \sum_{t=1}^{T} \ln f(x_t). \]
4.2 Criterion Function

In the population define the criterion function as

\[
Q = E_{\theta_0} \left( \ln f(y_t, x_t; \theta) \right)
= \int (\ln f(y_t, x_t; \theta)) f(y_t, x_t; \theta_0) \, dy_t dx_t.
\]

(We assume this integral exists.)

We pick the \( \theta \) that maximizes \( \mathcal{L} \). Note that this is an extremum principle application of the analogy principle.
Claim. The criterion function is maximized at $\theta = \theta_0$.

Proof.

$$E_{\theta_0} \left( \frac{f(y_t, x_t; \theta)}{f(y_t, x_t; \theta_0)} \right) = 1 \quad \text{because}$$

$$\int \frac{f(y_t, x_t; \theta)}{f(y_t, x_t; \theta_0)} f(y_t, x_t; \theta_0) \, dy_t \, dx_t = 1.$$
Applying Jensen’s inequality, concavity of the ln function implies that

\[ \mathbb{E}(\ln(x)) \leq \ln \mathbb{E}(x) \]

\[ \Rightarrow \mathbb{E}_{\theta_0} \left( \ln \left( \frac{f(y_t, x_t; \theta)}{f(y_t, x_t; \theta_0)} \right) \right) \leq 0 \]

\[ \Rightarrow (\forall \theta) \mathbb{E}_{\theta_0} (\ln f(y_t, x_t; \theta)) \leq \mathbb{E}_{\theta_0} (\ln f(y_t, x_t; \theta_0)) \]

We get global identification in the population if the inequality is strict for all \( \theta \neq \theta \).
4.3 Analog in Sample

Construct a sample analog of the criterion function $Q_T$ as

$$Q_T := \frac{1}{T} \sum_{t=1}^{T} \ln f(y_t, x_t; \theta).$$

4.4 The Estimator and its Properties

We form the estimator as

$$\hat{\theta}_T := \theta \in \Theta \ \text{argmax} \ Q_t$$

(we assume this exists).
Local form of the principle. In the local form, we use FOC and SOC to arrive at $\theta \in \Theta \ arg\max Q_t$. Recall that we have the criterion function

$$Q(\theta) = \int \ln f(y; \theta) f(y; \theta_0) \, dy = E_{\theta_0} \left( \ln f(y; \theta) \right).$$

Maximization of the criterion function yields first and second order conditions:

FOC: \[ \int \frac{\partial \ln f(y, \theta)}{\partial \theta} f(y; \theta_0) \, dy = 0 \]

SOC: \[ \int \frac{\partial^2 \ln f(y; \theta)}{\partial \theta \partial \theta'} f(y; \theta_0) \text{ negative definite} \]
Accordingly, we require for the sample analog:

\[
\frac{1}{T} \sum_{t=1}^{T} \frac{\partial \ln f(y_t; \theta)}{\partial \theta} = 0
\]

\[
\frac{1}{T} \sum_{t=1}^{T} \frac{\partial^2 \ln f}{\partial \theta \partial \theta'} \quad \text{negative definite}
\]

For ‘local identification’, we require that the second order conditions be satisfied locally around the point solving the FOC. For ‘global’ identification, we need SOC to hold for every \( \theta \in \Theta \).
Either way (e.g. directly by a grid search or using the FO/SOCs), we have the same basic idea. For each $T$, we pick $\hat{\theta}_T$ such that

$$(\forall \theta \in \Theta) \ Q_T(\hat{\theta}_T) > Q_T(\theta).$$

Now if $Q_T(\hat{\theta}_T) \to Q(\lim \hat{\theta}_T)$ (uniform convergence), we get the contradiction

$$Q(\hat{\theta}_T) > Q(\theta_0)$$

assumed to be a maximum value — unless $\text{plim} \hat{\theta}_T = \theta_0$.

To be more precise, we must check whether

$$Q_T \xrightarrow{\text{uniformly}} Q \text{ (almost surely).}$$

It remains to cover certain concepts and definitions for random functions.
5 Some Concepts and Definitions for Random (Stochastic) Functions

In §5.1 we define random functions and examine some fundamental properties of such functions. In §5.2 we define convergence concepts for sequences of random functions.
5.1 Random Functions and Some Properties

Definition 2 (Random Function)
Let $(\Omega, \mathcal{A}, P)$ be a probability space and let $\Theta \in \mathbb{R}^k$. A real function $\varphi(\theta) = \varphi(\theta, \omega)$ on $\Theta \times \Omega$ is called a random function on $\Theta$ if

$$(\forall t \in \mathbb{R}^1)(\forall \theta \in \Theta) \quad \{\omega \in \Omega : \varphi(\theta, \omega) < t\} \in \mathcal{A}.$$ 

We can then assign a probability to the event: $\varphi(\theta, \omega) < t$. 
Proposition. If $\varphi(\theta, x)$ is a continuous real-valued function on $\Theta \times \mathbb{R}^n$ where $\Theta$ is compact, then

$$
\begin{align*}
g(x) &\equiv \sup_{\theta \in \Theta} \varphi(\theta, x) \quad \text{and} \quad h(x) \equiv \inf_{\theta \in \Theta} \varphi(\theta, x)
\end{align*}
$$

are continuous functions.

Proof. See Stokey, Lucas, and Prescott, Chapter 3 for definitions of $\sup$ and $\inf$, and for the proof of this proposition (q.v. Theorem 3.6).
Proposition. If for almost all values of $x \in X$, $g(x, \theta)$ is continuous with respect to $\theta$ at the point $\theta_0$, and if for all $\theta$ in a neighborhood of $\theta_0$ we have

$$|g(x, \theta)| < G_1(x) < \infty,$$

then

$$\lim_{\theta \to \theta_0} \int g(x, \theta) \, dF(x) = \int g(x, \theta_0) \, dF(x).$$

I.e.

$$\lim_{\theta \to \theta_0} \mathbb{E}(g(x, \theta)) = \mathbb{E}(g(x, \theta_0)).$$

Proof. This is a version of a ‘dominated convergence theorem’. See inter alios Royden, Chapter 4.
Proposition. If for almost all values of $x \in X$ and for a fixed value of $\theta$

(a) \[ \frac{\partial g(x, \theta)}{\partial \theta} \] exists (in a neighborhood of $\theta$), and

(b) \[ \left| \frac{g(x, \theta + h) - g(x, \theta)}{h} \right| < G_2(x), \]

for $0 < |h| < h_0$, $h$ independent of $x$, then

\[
\frac{\partial}{\partial \theta} \int g(x, \theta) \, dF(x) = \int \frac{\partial g(x, \theta)}{\partial \theta} \, dF(x).
\]

I.e.

\[
\frac{\partial}{\partial \theta} E(g(x, \theta)) = E \left[ \frac{\partial g(x, \theta)}{\partial \theta} \right].
\]
5.2 Convergence Concepts for Random Functions

In Part I (asymptotic theory) we defined convergence concepts for random variables. Here we define analogous concepts for random functions.

**Definition 3 (Almost Sure Convergence)**

Let $I_i(\omega)$ and $I_q(\omega)$ be random functions on $\Theta \subseteq R^k$ for each $\theta \in \Theta$. Then $I_q(\omega)$ almost surely converges to $I_i(\omega)$ as $n \rightarrow \infty$ if

\[
(\forall \varepsilon > 0) \quad P\{\omega : \lim_{n \rightarrow \infty} |F_n(\theta, \omega) - F(\theta, \omega)| < \varepsilon\} = 1;
\]

i.e. if for every fixed $\theta$ the set $S_{\theta} \subseteq X$ such that $|F_n(\theta, \omega) - F(\theta, \omega)| \geq \varepsilon$, $n \geq n_0(\theta, \varepsilon)$, has no probability.
S = \theta \in \Theta \cup S_\theta \text{ may have a non-negligible probability even though any one set has negligible probability. We avoid this by the following definition.}

**Definition 4 (Almost Sure Uniform Convergence)**

\( F_n(\theta) \to F(\theta) \) almost surely uniformly in \( \theta \) if

\[
\sup_{\theta \in \Theta} |F_n(\theta) - F(\theta)| \to 0
\]

almost surely as \( n \to \infty \). I.e., if

\[(\forall \varepsilon > 0)(\forall \omega \in \Omega) \quad P\{\omega : \lim_{n \to \infty} \sup_{\theta \in \Theta} |F_n(\theta, \omega) - F(\theta, \omega)| \leq \varepsilon\} = 1.\]

In this case, the negligible set is not indexed by \( \theta \).
Definition 5 (Convergence in Probability) Let $F_n(\theta)$ and $F(\theta)$ be random functions on $\Theta$. Then $F_n(\theta) \to F(\theta)$ in probability uniformly in $\theta$ on $\Theta$ if

$$\lim_{n \to \infty} P\left\{ \sup_{\theta \in \Theta} |F_n(\theta) - F(\theta)| > \varepsilon \right\} = 0.$$
Theorem 1 (Strong Uniform Law of Large Numbers)

Let \( \{x_n\} \) be a sequence of random \( k \times 1 \) i.i.d. vectors. Let \( F(x, \theta) \) be a continuous real function on \( R^k \). \( \Theta \) is compact (it is closed and bounded, and thus has a finite subcover). Define

\[
\psi(a) = \sup_{||x|| < a} \sup_{\theta \in \Theta} |F(x, \theta)|.
\]

Let \( G(x) \) be the distribution of \( x \). Assume \( E[\psi(x)] < \infty \). Then,

\[
\frac{1}{T} \sum_{j=1}^{T} F(x_j, \theta) \xrightarrow{a.s. \text{unif.}} \int F(x, \theta) \, dG(x).
\]
This is a type of LLN, and could be modified for the non-i.i.d. case. In that case, each $x_j$ has its own distribution $F_j$, and we would require that

\[(a) \quad \frac{1}{T} \sum_{j=1}^{T} F_j \xrightarrow{d} G \quad \text{and} \]

\[(b) \quad T \sup \frac{1}{T} \sum_{j=1}^{T} E(|\psi(x_j)|)^{1+\sigma} < \infty. \]

Then

\[
\frac{1}{T} \sum_{j=1}^{T} F(x_j, \theta) \xrightarrow{a.s.\,\text{unif.}} \int F(x, \theta) \, dG(x).
\]

Note that we need a bound in either case on $E[|\psi(x_j)|]$. 
References


