1 Introduction

Many of the proofs in the paper Hansen (1982) were not published. These notes give the proofs to the missing results. Section 2 provides the required assumptions and proofs for a Law of Large Numbers for Random Functions to hold. A theorem is presented that guarantees the almost sure uniform convergence of the sample mean of a random function to its population counterpart. Section 3 establishes the consistency of a GMM as a direct application of the theorem. Section 4 relaxes the compactness of the parameter space assumption and provides an alternative theorem to the one of section 2 by imposing an special structure on the form of the random function.

2 Law of Large Numbers for Random Functions

We begin giving some definitions and assumptions.

Let $\mathcal{L}$ be the space of continuous functions defined on a parameter space P, and let λ be the *sup* norm on $\mathcal{L}$. It is well known that the Law of Large Numbers applies to stochastic processes that are stationary and ergodic. Recall that a stochastic process can be viewed as an indexed (by calendar time) family of random variables. In this chapter we extend the Law of Large Numbers to stationary and ergodic indexed families (by calendar time) of random functions. A random function is a measurable mapping from the collection of states of the world Ω into the space $\mathcal{L}$ of continuous functions. It is the natural extension of a random variable.

We assume that the metric space of potential parameter values is compact.

Assumption 2.1. (P, π) *is a compact metric space.*

Let $(\Omega, \mathcal{F}, Pr)$ denote the underlying probability space and ϕ the random function under consideration.

Definition 2.1. *A* **random function** ϕ *is a measurable function mapping* $\Omega \rightarrow \mathcal{C}$.

Assumption 2.2. ϕ *is a random function.*

Sometimes we will suppress the dependence on ω and just write $\phi(\beta)$. Under Assumption 2.2, $\phi(\beta)$ is a random variable for each β in P. In fact, the measurability restriction in Assumption 2.2 is equivalent to requiring that $\phi(\beta)$ be a random variable for each β in P [see Billingsley (1978)].

We construct stochastic processes by using a transformation $S : \Omega \rightarrow \Omega$. Our development follows Breiman (1968).

Definition 2.2. S *is* **measure preserving** *if for any event* Λ *in* $\mathcal{F}$, $Pr(\Lambda) = Pr[S^{-1}(\Lambda)]$.

Definition 2.3. S *is* **ergodic** *if for any event such that* $S^{-1}(\Lambda) = \Lambda$, $Pr(\Lambda)$ *is equal to zero or one.*[1]

[1] Event for which $S^{-1}(\Lambda) = \Lambda$ are referred to as invariant events.

Assumption 2.3. *S is measure-preserving and ergodic.*

The process of random functions is constructed using S.

$$\phi_t(\omega, \beta) = \phi\left[S^t(\omega), \beta\right]$$

The stochastic process $\{\phi_t(\cdot, \beta) : t \geq 0\}$ is stationary and ergodic for each β [see Proposition 6.9 and Proposition 6.18 in Breiman (1968)]. The notation $\sum_T(\phi)$ denotes the random function given by

$$\sum_T [\phi(\omega, \beta)] = \phi[S(\omega), \beta] + \phi[S^2(\omega), \beta] + ... + \phi\left[S^T(\omega), \beta\right].$$

To obtain a Law of Large Numbers for ϕ we require that ϕ has finite first moments and be continuous in a particular sense.

Definition 2.4. *A random function ϕ has* finite first moments *if $E|\phi(\beta)| < \infty$ for all $\beta \in P$.*

Assumption 2.4. *ϕ has finite first moments.*

The following notation is used for our continuity restriction:

$$Mod_\phi(\delta, \beta) = \sup\{|\phi(\beta) - \phi(\beta^*)| : \beta^* \in P \text{ and } \pi(\beta, \beta^*) < \delta\}.$$

Since P is compact, P is separable. Consequently, a dense sequence $\{\beta_j : j \geq 1\}$ can be used in place of P in evaluating the *supremum.* In this case, $Mod_\phi(\delta, \beta)$ is a random variable for each positive value of δ and each β in P. Also, $Mod_\phi(\delta, \beta) \geq Mod_\phi(\delta^*, \beta)$ if δ is greater than δ^*. Since ϕ maps into $\mathcal{L}$,

$$\lim_{\delta \downarrow 0} Mod_\phi(\delta, \beta) = 0 \text{ for all } \omega \in \Omega \text{ and all } \beta \in P. \quad (1)$$

Definition 2.5. *A random function ϕ is* first-moment continuous *if for each $\beta \in P$,*

$$\lim_{\delta \downarrow 0} E[Mod_\phi(\delta, \beta)] = 0.$$

Assumption 2.5. *ϕ is first-moment continuous.*

Since $Mod_\phi(\delta, \beta)$ is decreasing in δ, we have the following result.

Lemma 2.1 (DeGroot (1970) page 206)**.** *Under Assumptions 2.1 and 2.2, ϕ is first-moment continuous if, and only if, for each $\beta \in P$ there is $\delta_\beta > 0$ such that*

$$E[Mod_\phi(\delta_\beta, \beta)] < \infty.$$

Proof. The *if* part of the proof follows from the Dominated Convergence Theorem and (1). The *only if* part is immediate. □

Since S is ergodic, a natural candidate for the limit of time series averages of ϕ is $E(\phi)$. To establish the Law of Large Numbers for Random Functions, we use: i) the pointwise continuity of $E(\phi)$, ii) a pointwise Law of Large Numbers for $\{(1/T)\sum_T(\phi)(\beta)\}$ for each β in P, and iii) a pointwise Law of Large Numbers for $\{(1/T)\sum_T[Mod_\phi(\delta,\beta)]\}$ for each β in P and positive δ. We establish these approximation results in three lemmas prior to our proof of the Law of Large Numbers for Random Functions. We then demonstrate our main result by showing that the assumption of a compact parameter space (P,π) can be used to obtain an approximation that is uniform.

Lemma 2.2 establishes the continuity of $E(\phi)$.

Lemma 2.2. *Suppose Assumptions 2.1, 2.2, 2.4 and 2.5 are satisfied. Then there is positive-valued function $\delta^*(\beta,j)$ satisfying*

$$|E[\phi(\beta^*)] - E[\phi(\beta)]| < 1/j$$

for all $\beta^* \in P$ such that $\pi(\beta^*,\beta) < \delta^*(\beta,j)$ and all integer $j \geq 1$.

Proof. Since ϕ is first-moment continuous, there is a function $\delta^*(\beta,j)$ such that

$$E(Mod_\phi[\beta,\delta^*(\beta,j)]) < 1/j.$$

Note, however, that

$$\begin{aligned} |E\phi(\beta^*) - E\phi(\beta)| &\leq E|\phi(\beta^*) - \phi(\beta)| \\ &\leq E[Mod_\phi(\beta,\delta^*)(\beta,j)] \\ &< 1/j \end{aligned}$$

for all $\beta^* \in P$ such that $\pi(\beta,\beta^*) < \delta^*(\beta,j)$. □

Since P is compact, it can be shown that $\delta^*(\beta,j)$ in Lemma 2.2 can be chosen independent of β. In other words, $E(\phi)$ is uniformly continuous.

For each element β in P, $\phi(\beta)$ is a random variable with a finite absolute first moment. Thus the Law of Large Numbers applies pointwise as stated in the following lemma.

Lemma 2.3. *Suppose Assumptions 2.1, 2.2, 2.3 and 2.4 are satisfied. Then there is an integer-valued function $T^*(\omega,\beta,j)$ and an indexed set $\Lambda^*(\beta) \in \mathcal{F}$ such that $Pr\{\Lambda^*(\beta)\} = 1$ and*

$$\left|(1/T)\sum_T[\phi(\beta)] - E[\phi(\beta)]\right| < 1/j \qquad (2)$$

for all $\beta \in P$, $T \geq T^*(\omega,\beta,j)$, $\omega \in \Lambda^*(\beta)$, and $j \geq 1$.

Proof. Since S is measure-preserving and ergodic, $\{(1/T)\sum_T[\phi(\beta)] : T \geq 1\}$ converges to $E[\phi(\beta)]$ on a set $\Lambda^*(\beta) \in \mathcal{F}$ satisfying $Pr\{\Lambda^*(\beta)\} = 1$. □

The Law of Large Numbers also applies to time series averages of $Mod_\phi(\beta,\delta)$. Since the mean of $Mod_\phi(\beta,\delta)$ can be made arbitrarily small by choosing δ to be small, we can control the local variation of time series averages of the random function ϕ.

Lemma 2.4. *Suppose Assumptions 2.1, 2.2, 2.3 and 2.5 are satisfied. There exists an integer-valued function $T^+(\omega,\beta,j)$, a positive function $\delta^+(\beta,j)$ and an indexed set $\Lambda^+(\beta) \in \mathcal{F}$ such that $Pr\{\Lambda^+(\beta)\} = 1$ and*

$$\left|(1/T)\sum_T [\phi(\beta) - \phi(\beta^*)]\right| < 1/j \tag{3}$$

for all $\beta^* \in P$ such that $\pi(\beta,\beta^*) < \delta^+(\beta,j)$, $T \geq T^+(\omega,\beta,j)$, $\omega \in \Lambda^+(\beta)$ and $j \geq 1$.

Proof. Since ϕ is first-moment continuous, $Mod_\phi(\beta,1/n)$ has a finite first moment for some positive integer n. Since S is measure-preserving and ergodic, $\{(1/T)\sum_T [Mod_\phi(\beta,1/j)] : T \geq 1\}$ converges to $E[Mod_\phi(\beta,1/j)]$ on a set $\Lambda^-(\beta,j)$ satisfying $Pr\{\Lambda^-(\beta,j)\} = 1$ for $j \geq n$. Let

$$\Lambda^+(\beta) = \bigcap_{j\geq n} \Lambda^-(\beta,j).$$

Then $\Lambda^+(\beta)$ is measurable and $Pr\{\Lambda^+(\beta)\} = 1$.

For each j, choose $[1/\delta^+(\beta,j)]$ to equal some integer greater than or equal to n such that

$$E\left((Mod_\phi\left[\beta,\delta^+(\beta,j)\right]\right) < 1/(2j).$$

Since $\{(1/T)\sum_T\{Mod_\phi[\beta,\delta^+(\beta,j)] : T \geq 1\}$ converges almost surely to $E(Mod_\phi[\beta,\delta^+(\beta,j)])$ on $\Lambda^+(\beta)$, there exists an integer-valued function $T^+(\omega,\beta,j)$ such that

$$\left|(1/T)\sum_T \left(Mod_\phi\left[\beta,\delta^+(\beta,j)\right]\right) - E\left(Mod_\phi\left[\beta,\delta^+(\beta,j)\right]\right)\right| < 1/2j$$

for $T \geq T^+(\omega,\beta,j)$. Therefore,

$$(1/T)\left|\sum_T [\phi(\beta) - \phi(\beta^*)]\right| < (1/T)\sum_T \left(Mod_\phi\left[\beta,\delta^+(\beta,j)\right]\right) < 1/j$$

for all $\beta^* \in P$ such that $\pi(\beta,\beta^*) < \delta^+(\beta,j)$, $T \geq T^+(\omega,\beta,j)$, $\omega \in \Lambda^+(\beta)$, and $j \geq 1$. □

Our main result establishes the almost sure convergence of time series averages of random functions. Suppose ϕ_1 and ϕ_2 are two random functions. Then

$$\lambda(\phi_1,\phi_2) = \sup_{\beta\in P} |\phi_1(\beta) - \phi_2(\beta)|$$

is a measure of distance between these functions that depends on the sample point. Since P is separable, it suffices to take the supremum over a countable dense sequence. Hence $\lambda(\phi_1,\phi_2)$ is a random variable (measurable function). Almost sure convergence of sequences of random functions is defined using the metric λ.

Definition 2.6. *A sequence* $\{\phi_j : j \geq 1\}$ *of random functions converges almost surely to a random function* ϕ_0 *if* $\{\lambda(\phi_j, \phi_0) : j \geq 1\}$ *converges almost surely to zero.*

We now combine the conclusions from Lemmas 2.2, 2.3 and 2.4 to obtain a Law of Large Numbers for random functions. The idea is to exploit that fact that P is compact to move from pointwise to uniform convergence. Notice that in these three lemmas, $\Lambda^+, \Lambda^*, T^+$ and T^* all depend on β. In proving this Law of Large Numbers, we will use compactness to show how the dependence on the parameter value can be eliminated.

Theorem 2.1. *Suppose Assumptions 2.1, 2.2, 2.3, 2.4 and 2.5 are satisfied. Then* $\{(1/T)\sum_T(\phi) : T \geq 1\}$ *converges almost surely to* $E(\phi)$.

Proof. Let

$$Q(\beta, j) = \{\beta^* \in P : \pi(\beta, \beta^*) < \min\{\delta^*(\beta, j), \delta^+(\beta, j)\}.$$

Then for each $j \geq 1$,

$$P = \bigcup_{\beta \in P} Q(\beta, j).$$

Since P is compact a finite number of β_i can be selected so that

$$P = \bigcup_{i \geq 1}^{N(j)} Q(\beta_i, j)$$

where $N(j)$ is integer-valued and $\{\beta_i : i \geq 1\}$ is a sequence in P. Construct

$$\Lambda = \bigcap_{i \geq 1} \left[\Lambda^*(\beta_i) \bigcap \Lambda^+(\beta_i)\right].$$

Then $\Lambda \in \mathcal{F}$ and $Pr(\Lambda) = 1$. Let

$$\begin{aligned} T(\omega, j) &= \max\{T^*(\omega, \beta_1, j), T^*(\omega, \beta_2, j), ..., T^*\left[\omega, \beta_{N(j)}, j\right], \\ &\quad T^+(\omega, \beta_1, j), T^+(\omega, \beta_2, j), ..., T^+\left[\omega, \beta_{N(j)}, j\right]\}. \end{aligned}$$

For $T \geq T(\omega, j)$, Lemmas 2.2, 2.3 and 2.4 imply that

$$\begin{aligned} \left|(1/T)\sum_T[\phi(\beta)] - E[\phi(\beta)]\right| &\leq (1/T)\left|\sum_T[\phi(\beta)] - \sum_T[\phi(\beta_i)]\right| \\ &+ \left|(1/T)\sum_T[\phi(\beta_i)] - E[\phi(\beta_i)]\right| + |E[\phi(\beta_i)] - E[\phi(\beta)]| \\ &< 3/j \end{aligned}$$

where β_i is chosen so that $\beta \in Q(\beta_i, j)$ for some $1 \leq i \leq N(j)$. Hence

$$\lambda \left[(1/T) \sum_T (\phi), E\phi \right] \leq 3/j$$

for $T \geq T(\omega, j)$ and $\omega \in \Lambda$. Therefore, $\{\lambda [(1/T) \sum_T (\phi), E\phi] : T \geq 1\}$ converges to zero on Λ. □

3 Consistency of the GMM Estimator

We apply Theorem 2.1 to obtain an approximation result for a GMM estimator as defined in Hansen (1982). So far, the results obtained in the previous section have been for the case of random functions that map into $\mathbb{R}$. The GMM estimator works by making sample analogues of population orthogonality conditions close to zero. We will map the assumptions in this note to those in Hansen (1982) and show that the consistency of the GMM estimator (Theorem 2.1 in Hansen (1982)) is an application of Theorem 2.1 above.

First, construct a stochastic process by using the transformation $S : \Omega \to \Omega$. Let,

$$x_t(\omega) = x\left[S^t(\omega)\right]$$

where $x_t(\omega)$ has p components as in Hansen (1982).

Assumption 2.3 from the previous section ensures that this process is stationary and ergodic, which is Assumption 2.1 in Hansen (1982). Hansen (1982) restricts the parameter space to be a separable metric space (Assumption 2.2). This is implied by assumption 2.1 above, since a compact space is separable.

To represent the population orthogonality conditions we will consider a function $f : \mathbb{R}^p \times P \to \mathbb{R}^r$ where r is greater than or equal to p. The random function ϕ in section 2 is given by:

$$\phi(\omega, \beta) = f[x(\omega), \beta]. \tag{4}$$

Hansen (1982) requires that $f(\cdot, \beta)$ is Borel measurable for each β in P and $f(x, \cdot)$is continuous on P for each x in $\mathbb{R}^p$ (Assumption 2.3). Given construction (4), these restrictions imply Assumption 2.2.

Assumption 2.4 in the GMM paper requires that $Ef(x, \beta)$ exists and is finite for all $\beta \in P$ and $Ef(x, \beta_0) = 0$. The first part of this assumption is equivalent to Assumption 2.4 given construction (4) of the random function ϕ. To match the GMM paper, we impose the identification restriction:

Assumption 3.1. *The equation*

$$E\phi(\beta) = 0 \tag{5}$$

is satisfied on P if, and only if $\beta = \beta_o$.

Finally, as in Assumption 2.5 of Hansen (1982) we introduce a sequence of weighting matrices:

Assumption 3.2. *The sequence of random matrices* $\{a_T : T \geq 1\}$ *converges almost surely to a constant matrix* a_0.

We now have all the ingredients to state the following corollary to Theorem 2.1.

Corollary 3.1. *(Theorem 2.1 in Hansen (1982)) Suppose that Assumptions 2.1, 2.2, 2.3, 2.4, 2.5, 3.1 and 3.2 are satisfied. Then*

$$b_T = \arg\min_{\beta \in P} \left[\frac{1}{T} \sum_{t=1}^{T} f(x_t, \beta) \right]' a_T' a_T \left[\frac{1}{T} \sum_{t=1}^{T} f(x_t, \beta) \right]$$

converges almost surely to β_o.

Proof. Theorem 2.1 implies that the objective function for the minimization problem converges uniformly to:

$$Ef(x_t, \beta)' a_0' a_0 Ef(x_t, \beta)$$

with probability one. The limiting objective has a unique minimizer at β_o. The conclusion follows. □

4 Parameter Separation

We now explore an alternative consistency argument that avoids assuming a compact parameter space. Instead we suppose that we have estimation equations with a special separable structure. We take as our starting point:

$$EXh(\beta) = 0. \tag{6}$$

where the random matrix X is a general function of the data and is r by m. This matrix may in fact contain nonlinear tansformations of the data. The key restriction is that it does not depend on unknown parameters.

Assumption 4.1. *P is locally compact.*

We presume that the parameter vector β_o can be identified from the moment condition 6. We will sharpen this assumption in what follows.

Assumption 4.2. *$h : P \to \Gamma \subset \mathbb{R}^m$ is a homeomorphism.*[2]

Neither P nor Γ is necessarily compact. While we can estimate EX in large samples, approximation errors get magnified if Γ is unbounded as β ranges over the parameter space.

Prior to studying the estimation problem that interests us, we first consider an auxiliary problem. Define:

$$\gamma_o = h(\beta_o).$$

Consider estimating γ_o by solving:

$$\tilde{c}_T = \arg\min_{\gamma\in\Gamma} \frac{\gamma' \left[\frac{1}{T}\sum(X)\right]' a_T' a_T \left[\frac{1}{T}\sum(X)\right]\gamma}{1+|\gamma|^2}$$

The scaling by $1 + |\gamma|^2$ limits the magnitude of the approximation error. With this in mind we form Γ^* to be the closure of the bounded set:

$$\hat{\Gamma} = \left\{ \frac{\gamma}{\sqrt{1+|\gamma|^2}} : \gamma \in \Gamma \right\}.$$

Assumption 2.2 and 2.3 guarantees that

$$\frac{1}{T}\sum_T (X) \to EX$$

provided that X has a finite first moment.

Assumption 4.3. *The matrix X has a finite first moment and $EX\gamma = 0$ on Γ^* if, and only if $\gamma = \frac{\gamma_o}{\sqrt{1+|\gamma_o|^2}}$.*

[2] h is continuous, one-to-one, and has a continuous inverse.

Since we have added in the closure points, this is a stronger condition than the usual identification condition using moment conditions.

Theorem 4.1. *Suppose that Assumptions 2.2, 2.3, 3.2, 4.1, 4.2, and 4.3 are satisfied. Then* $\tilde{c}_T$ *converges to* γ_o *almost surely.*

Proof. First transform the problems to be:

$$\hat{c}_T = \arg\min_{\gamma^* \in \Gamma^*} \gamma^{*\prime} \left[\frac{1}{T}\sum(X)\right]' a_T' a_T \left[\frac{1}{T}\sum(X)\right]' \gamma^*.$$

Since the objective function converges uniformly almost surely, the sequence of minimizers $\hat{c}_T$ converges to:

$$\frac{\gamma_o}{\sqrt{1+|\gamma_o|^2}} = \arg\min_{\gamma^* \in \Gamma^*} \gamma^{*\prime}(EX)' a_0' a_0 (EX)\gamma^*.$$

Since $\hat{\Gamma}$ is locally compact, for sufficiently large T, $\hat{c}_T$ is in the set $\hat{\Gamma}$. Thus:

$$\tilde{c}_T = \frac{\hat{c}_T}{\sqrt{1-|\hat{c}_T|^2}}$$

for sufficiently large T. The conclusion follows from the convergence of $\hat{c}_T$. □

While the auxiliary estimator is of interest in its own right, we will now use the auxiliary estimator as a bound for:

$$c_T = \arg\min_{\gamma \in \Gamma} \gamma' \left[\frac{1}{T}\sum(X)\right]' a_T' a_T \left[\frac{1}{T}\sum(X)\right] \gamma.$$

As we will see,

$$|c_T| \leq |\tilde{c}_T| \tag{7}$$

because the scaling of the objective function for the auxiliary estimator implicitly rewards magnitude.

Theorem 4.2. *Suppose that Assumptions 2.2, 2.3, 3.2, 4.1, 4.2, and 4.3 are satisfied. Then* c_T *converges to* γ_o *almost surely.*

Proof. There are two inequalities implied by the minimization problems used to construct c_T and $\tilde{c}_T$:

$$(c_T)' \left[\frac{1}{T}\sum(X)\right]' a_T' a_T \left[\frac{1}{T}\sum(X)\right] (c_T) \leq (\tilde{c}_T)' \left[\frac{1}{T}\sum(X)\right]' a_T' a_T \left[\frac{1}{T}\sum(X)\right] (\tilde{c}_T)$$

and

$$\frac{(c_T)' \left[\frac{1}{T}\sum(X)\right]' a_T' a_T \left[\frac{1}{T}\sum(X)\right] (c_T)}{1+|c_T|^2} \geq \frac{(\tilde{c}_T)' \left[\frac{1}{T}\sum(X)\right]' a_T' a_T \left[\frac{1}{T}\sum(X)\right] (\tilde{c}_T)}{1+|\tilde{c}_T|^2}.$$

Bound (7) follows because to achieve the second inequality, the left-hand side must have a smaller denominator.

Since $\hat{c}_T$ converges to γ_o, it follows that for sufficiently large T, $|c_T| \leq |\gamma_o| + 1$ and hence the estimator c_T is eventually in the compact set given by the closure of:

$$\{\gamma \in \Gamma : |\gamma| \leq |\gamma_o| + 1\}.$$

Since the objective converges uniformly almost surely on this compact set and the limiting objective has a unique minimizer at γ_o, the conclusion follows. □

Finally to produce an estimator of β_o we solve:

$$c_T = h(b_T).$$

Notice that we may equivalently define b_T as:

$$b_T = \arg\min_{\beta \in P} h(\beta)' \left[\frac{1}{T}\sum(X)\right]' a_T' a_T \left[\frac{1}{T}\sum(X)\right] h(\beta).$$

Corollary 4.1. *Suppose that Assumptions 2.3, 3.2, 4.1, 4.2, and 4.3 are satisfied. Then b_T converges to β_o almost surely.*

Proof. This follows from the fact that h is one-to-one with a continuous inverse. □

Theorem 2.2 in Hansen (1982) is a special case of the latter corollary. That paper assumes (Assumption 2.6) that

$$f(x_t, \beta) = \begin{bmatrix} c_o(x_t) & c_1(x_t) \end{bmatrix} \begin{bmatrix} 1 \\ \lambda(\beta) \end{bmatrix}.$$

Thus to see the relation, we may set

$$X_t = \begin{bmatrix} c_o(x_t) & c_1(x_t) \end{bmatrix},$$

and

$$h(\beta) = \begin{bmatrix} 1 \\ \lambda(\beta) \end{bmatrix}.$$

In addition to Assumptions 2.1 to 2.6, for which we have already provided a mapping with the assumptions of this paper, Theorem 2.2 in Hansen (1982) requires that the parameter space be locally compact, that λ be a homeomorphism and an identification assumption. These are implied by Assumptions 4.1, 4.2, and 4.3 in this note, respectively.

References

Breiman, Leo. 1968. *Probability*. Addison-Wesley Series in Statistics.

DeGroot, Morris H. 1970. *Optimal Statistical Decisions*. New York: McGraw-Hill.

Hansen, Lars P. 1982. Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica* 50:1029–1054.