Adaptive Stochastic Approximation by the Simultaneous Perturbation Method

Nifei Lin

October 2021

Content

- SPSA
- 2SPSA
  - Strong Convergence
  - Asymptotic Normality

Reference

Fabian, V. (1968). On asymptotic normality in stochastic approximation. The Annals of Mathematical Statistics, 1327-1332.

Spall, J. C. (1992). Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. IEEE transactions on automatic control, 37(3), 332-341.

Spall, J. C. (1997, December). Accelerated second-order stochastic optimization using only function measurements. In Proceedings of the 36th IEEE Conference on Decision and Control (Vol. 2, pp. 1417-1424). IEEE.

Spall, J. C. (1998). Implementation of the simultaneous perturbation algorithm for stochastic optimization. IEEE Transactions on aerospace and electronic systems, 34(3), 817-823.

Spall, J. C. (1998). An overview of the simultaneous perturbation method for efficient optimization. Johns Hopkins apl technical digest, 19(4), 482-492.

Spall, J. C., & Cristion, J. A. (1998). Model-free control of nonlinear stochastic systems with discrete-time measurements. IEEE transactions on automatic control, 43(9), 1198-1210.

Spall, J. C. (2000). Adaptive stochastic approximation by the simultaneous perturbation method. IEEE transactions on automatic control, 45(10), 1839-1853.

SPSA

- Problem: finding a root θ^* of the gradient equation

$$g(\theta) \equiv \frac{\partial L(\theta)}{\partial \theta} = 0$$

- SA standard form

$$\hat{\theta}_{k+1} = \hat{\theta}_{k+} - a_k \hat{g}(\hat{\theta}_k)$$

- The central FD estimator of $\hat{g}$ is

$$\hat{g}(\hat{\theta}_k) = \frac{1}{2c} \begin{pmatrix} y\left(\hat{\theta}_k + c\mathbf{e}_1\right) - y\left(\hat{\theta}_\mathbf{k} - c\mathbf{e}_1\right) \\ y\left(\hat{\theta}_\mathbf{k} + c\mathbf{e}_2\right) - y\left(\hat{\theta}_\mathbf{k} - c\mathbf{e}_2\right) \\ \vdots \\ y\left(\hat{\theta}_\mathbf{k} + c\mathbf{e}_d\right) - y\left(\hat{\theta}_\mathbf{k} - c\mathbf{e}_d\right) \end{pmatrix}$$

Let e_i denote the ith column of a $d \times d$ identity matrix.

SPSA

- Let $\Delta_k \in R^p$ be a vector of p mutually independent mean-zero random variables $\{\Delta_{k1}, \Delta_{k2}, ..., \Delta_{kp}\}$
- Let $\{\Delta_k\}$ be a mutually independent sequence with Δ_k independent of $\hat{\theta}_0, \hat{\theta}_1, ..., \hat{\theta}_k$.
- We have available noisy measurements of $L(\cdot)$:

$$y_k^{(+)} = L(\hat{\theta}_k + c_k \Delta_k) + \epsilon_k^{(+)}$$

$$y_k^{(-)} = L(\hat{\theta}_k - c_k \Delta_k) + \epsilon_k^{(-)}$$

where $\epsilon_k^{(+)}$, $\epsilon_k^{(-)}$ represent measurement noise terms that satisfy

$$E(\epsilon_k^{(+)} - \epsilon_k^{(-)} | \mathscr{F}, \Delta_k) = 0 a.s. \forall k, \mathscr{F}_k \equiv \left\{\hat{\theta}_0, \hat{\theta}_1, ..., \hat{\theta}_k\right\}$$

SPSA

- SPSA estimator of $g(\cdot)$ at the kth iteration is

$$\hat{g}_k\left(\hat{\theta}_k\right) = \begin{bmatrix} \frac{y_k^{(+)} - y_k^{(-)}}{2c_k\Delta_{k1}} \\ \vdots \\ \frac{y_k^{(+)} - y_k^{(-)}}{2c_k\Delta_{kp}} \end{bmatrix}$$

SPSA

Lemma [Spall, 1992]

Consider all $k \geq K$ for some $K < \infty$. Suppose that for each such k the $\{\Delta_{ki}\}$ are i.i.d. ($i = 1, 2, \cdots, p$) and symmetrically distributed about 0 with $|\Delta_{ki}| \leq \alpha_0$ a.s. and $E\left|\Delta_{ki}^{-1}\right| \leq \alpha_1$. For almost all $\hat{\theta}_k$ (at each $k \geq K$) suppose that $\forall \theta$ in an open neighborhood of $\hat{\theta}_k$ that is not a function of k or ω, $L^{(3)}(\theta) \equiv \partial^3 L / \partial\theta^T \partial\theta^T \partial\theta^T$ exists continuously with individual elements satisfying $\left|L^{(3)}_{i_1 i_2 i_3}(\theta)\right| \leq \alpha_2$. Then for almost all $\omega \in \Omega$

$$
\begin{aligned}
b_k\left(\hat{\theta}_k\right) \equiv & E\left(\hat{g}_k\left(\hat{\theta}_k\right) - g\left(\hat{\theta}_k\right) \mid \hat{\theta}_k\right) \\
& \left(= E\left(\hat{g}_k\left(\hat{\theta}_k\right) - g\left(\hat{\theta}_k\right) \mid \mathscr{F}_k\right)\right) \\
= & O\left(c_k^2\right)\left(c_k \to 0\right)
\end{aligned}
$$

Proof: Consider any $l \in \{1, 2, \cdots, p\}$ (let $\bar{\Delta}_{kl} = c_k \Delta_k$)
First, note that $E\left[(\epsilon_k^{(+)} - \epsilon_k^{(-)})/2\bar{\Delta}_{kl} \mid \hat{\theta}_k\right] = 0$ a.s.
Then by the continuity of $L^{(3)}$ near $\hat{\theta}_k$ and uniform boundedness of $|\Delta_{ki}|$ for all k sufficiently large, we have by Taylor's theorem for all such k

$$b_{kl}\left(\hat{\theta}_k\right) = \frac{1}{12} E\left\{\overline{\Delta}_{kl}^{-1}\left[L^{(3)}\left(\bar{\theta}_k^{+}\right) + L^{(3)}\left(\bar{\theta}_k^{-}\right)\right]\bar{\Delta}_k \otimes \bar{\Delta}_k \otimes \bar{\Delta}_k \mid \hat{\theta}_k\right\}$$

where $\bar{\theta}_k^{+}, \bar{\theta}_k^{-}$ are on the line segment between $\hat{\theta}_k$ and $\hat{\theta}_k \pm \bar{\Delta}_k$, respectively, and b_{kl} denotes the l th term of the bias b_k.
By the mean value theorem, the term on the r.h.s. , is bounded in magnitude by

$$\begin{aligned}
&\tfrac{\alpha_2 c_k^2}{6} \textstyle\sum_{i_1} \sum_{i_2} \sum_{i_3} E\left|\frac{\Delta_{ki_1}\Delta_{ki_2}\Delta_{ki_3}}{\Delta_{kl}}\right| \\
&\leq \tfrac{\alpha_2 c_k^2}{6} \cdot \left\{\left[p^3 - (p-1)^3\right]\alpha_0^2 + (p-1)^3 \alpha_1 \alpha_0^3\right\}
\end{aligned}$$

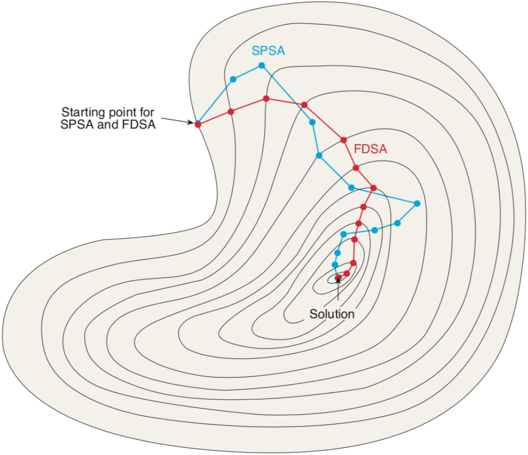

Figure 5. Example of relative search paths for SPSA and FDSA in $p = 2$ problem. Deviations of SPSA from FDSA average out in reaching a solution in the same number of iterations; FDSA nearly follows the gradient descent path (perpendicular to level curves) in the low-noise setting.

2SPSA

Basic form of algorithm(composed of two parallel recursions: one for θ and one for the Hessian of $L(\theta)$)

$$\hat{\theta}_{k+1} = \hat{\theta}_k - a_k \overline{\overline{H}}_k^{-1} G_k\left(\hat{\theta}_k\right), \quad \overline{\overline{H}}_k = f_k\left(\overline{H}_k\right)$$

$$\overline{H}_k = \frac{k}{k+1}\overline{H}_{k-1} + \frac{1}{k+1}\hat{H}_k, \quad k = 0, 1, 2, \cdots$$

- a stochastic analog of the well-known Newton-Raphson algorithm of deterministic search and optimization.
- recursive calculation of the sample mean of the per-iteration Hessian estimates

2SPSA

Notations:

$$\hat{\theta}_{k+1} = \hat{\theta}_k - a_k \overline{\overline{H}}_k^{-1} G_k\left(\hat{\theta}_k\right), \quad \overline{\overline{H}}_k = f_k\left(\overline{H}_k\right)$$

$$\overline{H}_k = \frac{k}{k+1}\overline{H}_{k-1} + \frac{1}{k+1}\hat{H}_k, \quad k = 0, 1, 2, \cdots$$

a_k: a nonnegative scalar gain coefficient
$G_k\left(\hat{\theta}_k\right)$: the input information related to $g\left(\hat{\theta}_k\right)$ (i.e., the gradient approximation from $y(\cdot)$ measurements in the gradient-free case or the direct observation as in the Robbins-Monro gradient-based case)
$f_k : \boldsymbol{R}^{p\times p} \to \{$ positive definite $p \times p$ matrices $\}$: a mapping designed to cope with possible nonpositive definiteness of $\bar{H}_k$
$\hat{H}_k$: a per-iteration estimate of the Hessian

2SPSA

The formula for estimating the Hessian at each iteration is:

$$\hat{H}_k = \frac{1}{2}\left[\frac{\delta G_k^T}{2c_k\Delta_k} + \left(\frac{\delta G_k^T}{2c_k\Delta_k}\right)^T\right]$$

where

$$\delta G_k = G_k^{(1)}\left(\hat{\theta}_k + c_k\Delta_k\right) - G_k^{(1)}\left(\hat{\theta}_k - c_k\Delta_k\right)$$

- for 2SG, usually $G_k^{(1)}(\cdot) = G_k(\cdot)$. We will suppose that $G_k(\cdot) = G_{k^k}^{(1)}(\cdot)$ is an unbiased direct measurement of $g(\cdot)$ (i.e., $G_k(\cdot) = G_k^{(1)}(\cdot) = g(\cdot) +$ mean-zero noise).
- for 2 SPSA: $G_k^{(1)}$ is a one-sided gradient approximation (in order to reduce the total number of function evaluations versus the two-sided form usually recommended for $G_k(\cdot)$])

2SPSA

$$G_k^{(1)}\left(\hat{\theta}_k \pm c_k\Delta_k\right)$$

$$=\frac{y\left(\hat{\theta}_k \pm c_k\Delta_k + \tilde{c}_k\tilde{\Delta}_k\right) - y\left(\hat{\theta}_k \pm c_k\Delta_k\right)}{\tilde{c}_k}\begin{bmatrix}\tilde{\Delta}_{k1}^{-1}\\ \tilde{\Delta}_{k2}^{-1}\\ \vdots\\ \tilde{\Delta}_{kp}^{-1}\end{bmatrix}$$

with $\tilde{\Delta}_k = \left(\tilde{\Delta}_{k1}, \tilde{\Delta}_{k2}, \cdots, \tilde{\Delta}_{kp}\right)^T$ generated in the same statistical manner as Δ_k, but independently of Δ_k and $\tilde{c}_k$ satisfying conditions similar to c_k

2SPSA

Two specific implementations of the ASP approach :

- 2SPSA (second-order SPSA) for applications in the gradient-free case (four function measurements $y(\cdot)$ are needed at each iteration)
- 2SG (second-order stochastic gradient) for applications in the Robbins-Monro gradient-based case. (three gradient measurements $g(\cdot)$ are needed at each iteration)

Strong Convergence

Theorem 1a

Consider the SPSA estimate for $G(\cdot)$ with $G(\cdot)^{(1)}$. Let conditions C.0-C.7 hold. Then $\hat{\theta}_k - \theta^* \to 0$ a.s.

Conditions

- C.0:$E\left(\varepsilon_k^{(+)} - \varepsilon_k^{(-)} \mid \hat{\theta}_k; \Delta_k; \bar{H}_k\right) = 0$ a.s. $\forall k$, where $\varepsilon_k^{(\pm)}$ is the effective SA measurement noise, i.e.,

$$\varepsilon_k^{(\pm)^n} \equiv y\left(\hat{\theta}_k \pm c_k \Delta_k\right) - L\left(\hat{\theta}_k \pm c_k \Delta_k\right)$$

- C.1:$a_k, c_k > 0\ \forall k$; $a_k \to 0, c_k \to 0$ as $k \to \infty$; $\sum_{k=0}^{\infty} a_k = \infty$, $\sum_{k=0}^{\infty} (a_k/c_k)^2 < \infty$
- C.2:For some $\delta, \rho > 0$ and $\forall k, \ell$, $E\left(\mid y\left(\hat{\theta}_k \pm c_k \Delta_k\right) / \Delta_{k\ell}|^{2+\delta}\right) \leq \rho$,$|\Delta_{k\ell}| \leq \rho$, $\Delta_{k\ell}$ is symmetrically distributed about 0 , and $\{\Delta_{k\ell}\}$ are mutually independent.

- C.3: For some $\rho > 0$ and almost all $\hat{\theta}_k$, the function $g(\cdot)$ is continuously twice differentiable with a uniformly (in k) bounded second derivative for all θ such that $\left\|\hat{\theta}_k - \theta\right\| \leq \rho$
- C.4: For each $k \geq 1$ and all θ, there exists a $\rho > 0$ not dependent on k and θ, such that $(\theta - \theta^*)^T \bar{g}_k(\theta) \geq \rho \|\theta - \theta^*\|$.
- C.5: For each $i = 1, 2, \cdots, p$ and any $\rho > 0$, $P\left(\left\{\bar{g}_{ki}\left(\hat{\theta}_k\right) \geq 0 \text{ i.o. }\right\} \cap \left\{\bar{g}_{ki}\left(\hat{\theta}_k\right) < 0 \text{ i.o. }\right\} \mid \left\{\left|\hat{\theta}_{ki} - (\theta^*)_i\right| \geq \rho \quad \forall k\right\}\right) = 0$

- C.6: $\overline{\overline{H}}_k^{-1}$ exists a.s. $\forall k, c_k^2 \overline{\overline{H}}_k^{-1} \to 0$ a.s., and for some $\delta, \rho > 0$, $E(\|\overline{\overline{H}}_k^{-1}\|^{2+\delta}) \leq \rho$
- C.7: For any $\tau > 0$ and nonempty $S \subseteq \{1, 2, \cdots, p\}$, there exists a $\rho'(\tau, S) > \tau$ such that

$$\limsup_{k \to \infty} \left| \frac{\sum_{i \notin S} (\theta - \theta^*)_i \, \bar{g}_{ki}(\theta)}{\sum_{i \in S} (\theta - \theta^*)_i \, \bar{g}_{ki}(\theta)} \right| < 1 \quad a.s.$$

  for all $|(\theta - \theta^*)_i| < \tau$ when $i \notin S$ and $|(\theta - \theta^*)_i| \geq \rho'(\tau, S)$ when $i \in S$

We define $\bar{g}(\hat{\theta}_k) = \overline{\overline{H}}^{-1} g(\hat{\theta}_k)$

Proof The proof will proceed in 3 parts.

- 1. $\tilde{\theta}_k \equiv \hat{\theta}_k - \theta^*$ does not diverge in magnitude to ∞
- 2. $\tilde{\theta}_k$ converges a.s. to some random vector;
- 3. this random vector is the constant 0

Part 1: Letting $M_j = a_j \overline{\overline{H}}_j^{-1} E(G_j(\hat{\theta}_j)|\hat{\theta}_j) = a_j \overline{\overline{H}}_j^{-1}(g_j(\hat{\theta}_j) + b_j)$ and $M_j' = a_j \overline{\overline{H}}_j^{-1}(\hat{g}_j(\hat{\theta}_j) - \bar{g}_j(\hat{\theta}_j))$, we can write

$$\tilde{\theta}_{k+1} + \sum_{j=0}^{k} M_j = \tilde{\theta}_0 - \sum_{j=0}^{k} M_j'$$

$\left\{\sum_{j=1}^{k} M_j'\right\}$ is a maringale sequence (in k)

$$E\left\|\sum_{j=0}^{k} M_j'\right\|^2 \leq 2\sum_{j=0}^{k} E\left\|M_j'\right\|^2 < \infty$$

Then by the martingale convergence theorem

$$\tilde{\theta}_{k+1} + \sum_{j=0}^{k} M_j \overset{\text{a.s.}}{\to} X$$

where X is some integrable random vector.
Let us now show that $P\left(\limsup_{k\to\infty} \left\|\tilde{\theta}_k\right\| = \infty\right) = 0$. Since the arguments below apply along any subsequence, epresented as

$$\bigcup_{S} \left\{\tilde{\theta}_{ki} \to \infty \quad \forall i \in S\right\}$$
$$\subseteq \bigcup_{\tau>0,S} \left\{\left\{\left\{\tilde{\theta}_{ki} \geq \rho'(\tau, S) \quad \forall i \in S, \quad \tilde{\theta}_{ki} \leq \tau \quad \forall i \notin S\right.\right.\right.$$
$$\left.k \geq K(\tau, S)\right\} \bigcap \limsup_{k\to\infty} \left\{M_{ki} < 0 \quad \forall i \in S\right\}\Bigg\}$$
$$\bigcup \left\{\left\{\tilde{\theta}_{ki} \to \infty \quad \forall i \in S\right\} \bigcap \liminf_{k\to\infty} \left\{M_{ki} < 0 \quad \forall i \in S\right\}^c\right\}\Bigg\}$$

For the first event:

Assuming there exists a subsequence $\{k_0, k_1, k_2, \cdots\}, k_0 \geq K(\tau, S)$ such that $\left\{\tilde{\theta}_{k_j i} \geq \rho'(\tau, S) \forall i \in S\right\} \cap \left\{M_{k_j i} < 0 \forall i \in S\right\}$ is true. Then, from C.6 and $M_j = a_j \overline{\overline{H}}_j^{-1} E(G_j(\hat{\theta}_j)|\hat{\theta}_j) = a_j(\bar{g}(\hat{\theta}_j) + \overline{\overline{H}}_j^{-1} o(c_j^2))$,

$$\sum_{i \in S} \tilde{\theta}_{k_j i} \bar{g}_{k_j i}(\hat{\theta}_{k_j}) < 0 \quad \text{a.s.}$$

for all k_j.

By C.4, $\tilde{\theta}_{k_j}^T \bar{g}_{k_j}\left(\hat{\theta}_{k_j}\right) \geq \rho \left\|\tilde{\theta}_{k_j}\right\|$ a.s. which, by C.7, $\rho'(\tau, S) \geq \tau$ and $\dim(S) \geq 1$ implies, for all j sufficiently large,

$$\textstyle\sum_{i \in S} \tilde{\theta}_{k_j i} \bar{g}_{k_j i}\left(\tilde{\theta}_{k_j}\right) \geq \frac{\rho}{2} \left\|\tilde{\theta}_{k_j}\right\| \geq \left(\frac{\rho}{2}\right) \dim(S) \rho'(\tau, S) \geq \frac{\rho\tau}{2} \quad \text{a.s.}$$

That's a contradiction. So the first event has probability 0.

For the second event

$\{\tilde{\theta}_{ki} \to \infty \quad \forall i \in S\} \bigcap \liminf_{k\to\infty} \{M_{ki} < 0 \quad \forall i \in S\}^c$, from

$$\tilde{\theta}_{k+1} + \sum_{j=0}^{k} M_j \overset{\text{a.s.}}{\to} X$$

for almost all sample points, $\sum_{k=0}^{\infty} M_{ki} \to -\infty \quad \forall i \in S$

However, at each k, the event $\{M_{ki} < 0 \forall i \in S\}^c$ is composed of the union of $2^{\dim(S)} - 1$ events, each of which has $M_{ki} \geq 0$ for at least one $i \in S$. Here is a contradiction. Hence, the probability of the second event is 0.

So $limsup_{k\to\infty} ||\tilde{\theta}_k|| < \infty$

Part 2: To show that $\tilde{\theta}_k$ converges a.s. to a unique (finite) limit, we show that

$$P\left(\liminf_{k\to\infty} \tilde{\theta}_{ki} < a < b < \limsup_{k\to\infty} \tilde{\theta}_{ki}\right) = 0 \quad \forall i$$

for any $a < b$.
There exist two subsequences, one with convergence to a point $< a$ and one with convergence to a point $> b$.
From $\tilde{\theta}_{k+1} + \sum_{j=0}^{k} M_j \stackrel{\text{a.s.}}{\to} X$ and the conclusion of Part 1, each of these subsequences has a sub-subsequence $\{k_{j_l}\}$ such that

$$\limsup_{l\to\infty} \left| \sum_{k=1}^{k_{j_l}} M_{ki} \right| < \infty \text{ a.s.}$$

Supposing that the event within the probability statement is true, we know that for any $\rho > 0$ and corresponding sample point we can choose $m > n$ sufficiently large so that for each i and combined sub-subsequence (from both sub-subsequences mentioned above)

$$|\sum_{k=k_{j_n}}^{k_{j_{m-1}}} M_{ki}| \leq \rho$$

$$|\tilde{\theta}_{k_{j_m}i} - \tilde{\theta}_{k_{j_n}i} + \sum_{k=k_{j_n}}^{k_{j_{m-1}}} M_{ki}| \leq \frac{b-a}{3}$$

$$\tilde{\theta}_{k_{j_n}i} < a < b < \tilde{\theta}_{k_{j_m}i}$$

Picking $\rho < (b-a)/3$ implies that

$$\left|\tilde{\theta}_{k_{j_n}i} - \tilde{\theta}_{k_{j_m}i}\right| \leq 2(b-a)/3$$

it requires that

$$\tilde{\theta}_{k_{j_m}i} - \tilde{\theta}_{k_{j_n}i} > b - a$$

which is a contradiction. So $\tilde{\theta}_k$ converges a.s. to a unique limit.

Part 3: Let us now show that the unique finite limit from Part 2 is 0 . we have $\lim sup_{n\to\infty} \left|\sum_{k=0}^{\infty} M_{ki}\right| < \infty$ a.s. $\forall i$.
Then the result to be shown follows if

$$P\left(\lim_{k\to\infty} \tilde{\theta}_k \neq 0, \left\|\sum_{k=0}^{\infty} M_k\right\| < \infty\right) = 0$$

Suppose that the event is true, and let $I \subseteq \{1, 2, \cdots, p\}$ represent those indexes i such that $\tilde{\theta}_{ki} \nrightarrow 0$ as $k \to \infty$. Then, there exists some $0 < a' < b' < \infty$ and $K(a', b') < \infty$ such that $\forall k \geq K, 0 < a' \leq \left|\tilde{\theta}_{ki}\right| \leq b' < \infty$ when $i \in I (I \neq \varnothing)$ and $\left|\tilde{\theta}_{ki}\right| < a'$ when $i \in I^c$. From C.4 and the conditions above, it follows that

$$\sum_{k=K+1}^{n} a_k \sum_{i\subset I} \tilde{\theta}_{ki} \bar{g}_{ki}\left(\hat{\theta}_k\right) \geq a'\rho \sum_{k=K+1}^{n} a_k$$

At least one $i \in I$

$$\limsup_{n \to \infty} \left| \frac{\rho a' \sum_{k=K+1}^{n} a_k}{\sum_{k=K+1}^{n} a_k \bar{g}_{ki}\left(\hat{\theta}_k\right)} \right| < \infty$$

Recall that $a_k \bar{g}_k \left(\hat{\theta}_k\right) = M_k - a_k \overline{\overline{H}}_k^{-1} b_k$ and $b_k = O\left(c_k^2\right)$ a.s.
Then, $\left|\sum_{k=K+1}^{\infty} M_{ki}\right| = \infty$. Hence, $\left|\sum_{k=0}^{\infty} M_{ki}\right| = \infty$ with probability > 0 for at least one i. However, this is inconsistent with the event $\left\|\sum_{k=0}^{\infty} M_k\right\| < \infty$, showing that the event does, in fact, have probability 0 . This completes Part 3, which completes the proof.
Q.E.D.

Strong Convergence For Hessian Estimator

Theorem 2a

Let conditions C.0, C.1", C.2, C.3, and C.4-C.9 hold. Then $\overline{H}_k \to H(\theta^*)$ a.s.

C.1": The conditions of C.1 hold plus $\sum_{k=0}^{\infty}(k+1)^{-2}(c_k\tilde{c}_k)^{-2} < \infty$ with $\tilde{c}_k = O(c_k)$

C.3': Change "thrice differentiable" in C.3 to "four-times differentiable" with all else unchanged.

C.9: $\tilde{\Delta}_k$ satisfies the assumptions for Δ_k in C.2 (i.e., $\forall k, \ell, \left|\tilde{\Delta}_{k\ell}\right| \leq \rho$ and $\tilde{\Delta}_{k\ell}$ is symmetrically distributed about 0; $\left\{\tilde{\Delta}_{k\ell}\right\}$ are mutually independent); Δ_k and $\tilde{\Delta}_k$ are independent; $E\left(\Delta_{k\ell}^{-2}\right) \leq \rho, E\left(\tilde{\Delta}_{k\ell}^{-2}\right) \leq \rho \forall k, \ell$ and some $\rho > 0$.

C.8: For some $\rho > 0$ and all k, ℓ, m,

$$E\left[y\left(\hat{\theta}_k \pm c_k\Delta_k + \tilde{c}_k\tilde{\Delta}_k\right)^2 / \left(\Delta_{k\ell}\tilde{\Delta}_{km}\right)^2\right] \leq \rho$$

and

$$E\left[y\left(\hat{\theta}_k \pm c_k\Delta_k\right)^2 / \left(\Delta_{k\ell}\Delta_{km}\right)^2\right] \leq \rho$$

$$E\left[\tilde{\varepsilon}_k^{(\pm)} - \varepsilon_k^{(\pm)} \mid \hat{\theta}_k; \tilde{\Delta}_k; \bar{H}_k\right] = 0$$

and

$$E\left[\left(\tilde{\varepsilon}_k^{(\pm)} - \varepsilon_k^{(\pm)}\right)^2 / \left(\Delta_{k\ell}\tilde{\Delta}_{km}\right)^2\right] \leq \rho$$

where $\tilde{\varepsilon}_k^{(\pm)} = y\left(\hat{\theta}_k \pm c_k\Delta_k + \tilde{c}_k\tilde{\Delta}_k\right) - L\left(\hat{\theta}_k \pm c_k\Delta_k + \tilde{c}_k\tilde{\Delta}_k\right)$

Proof:

$$
\begin{aligned}
& E\left[G_{k\ell}^{(1)}\left(\hat{\theta}_k \pm c_k\Delta_k\right) \mid \hat{\theta}_k, \Delta_k\right] \\
&= E\left[\frac{1}{\tilde{c}\tilde{\Delta}_{k\ell}}[\tilde{c}_k g\left(\hat{\theta}_k \pm c_k\Delta_k\right)^T \tilde{\Delta}_k + \frac{\tilde{c}_k^2}{2}\tilde{\Delta}_k^T H\left(\hat{\theta}_k \pm c_k\Delta_k\right)\tilde{\Delta}_k\right. \\
&\left. + \frac{\tilde{c}_k^3}{6}\sum_{h,i,j} L_{hij}^{(3)}\left(\bar{\theta}_k^{\pm}\right)\tilde{\Delta}_{kh}\tilde{\Delta}_{ki}\tilde{\Delta}_{kj}] \mid \hat{\theta}_k, \Delta_k\right] \\
&= g_\ell\left(\hat{\theta}_k \pm c_k\Delta_k\right) + \frac{1}{6}\tilde{c}_k^2 E\left[\tilde{\Delta}_{k\ell}^{-1}\sum_{h,i,j} L_{hij}^{(3)}\left(\bar{\theta}_k^{\pm}\right)\tilde{\Delta}_{kh}\tilde{\Delta}_{ki}\tilde{\Delta}_{kj} \mid \hat{\theta}_k, \Delta_k\right]
\end{aligned}
$$

$\tilde{\theta}_k^{\pm}$ are points on the line segments between $\hat{\theta}_k \pm c_k\Delta_k + \tilde{c}_k\tilde{\Delta}_k$ and $\hat{\theta}_k \pm c_k\Delta_k$;

Let

$$B_{k\ell} = \frac{1}{6} E \left[\tilde{\Delta}_{k\ell}^{-1} \sum_{h,i,j} \left(L_{hij}^{(3)} \left(\bar{\theta}_k^{+} \right) - L_{hij}^{(3)} \left(\bar{\theta}_k^{-} \right) \right) \cdot \tilde{\Delta}_{kh} \tilde{\Delta}_{ki} \tilde{\Delta}_{kj} \mid \hat{\theta}_k, \Delta_k \right]$$

we have $B_{k\ell} \sim o(c_k)$ for all k sufficiently large. Hence,

$$\begin{aligned}
& E \left(\hat{H}_{k,\ell m} \mid \hat{\theta}_k \right) \\
= & E \left(\frac{G_{k\ell}^{(1)} \left(\hat{\theta}_k + c_k \Delta_k \right) - G_{k\ell}^{(1)} \left(\hat{\theta}_k - c_k \Delta_k \right)}{2 c_k \Delta_{km}} \mid \hat{\theta}_k \right) \\
= & E \left(\frac{g_\ell \left(\hat{\theta}_k + c_k \Delta_k \right) - g_\ell \left(\hat{\theta}_k - c_k \Delta_k \right) + \tilde{c}_k^2 B_{k\ell}}{2 c_k \Delta_{km}} \mid \hat{\theta}_k \right) \\
= & E \left(\frac{2 c_k \left[\partial g_\ell / \partial \theta^T \right]_{\theta = \hat{\theta}_k} \Delta_k + O \left(c_k^3 \right)}{2 c_k \Delta_{km}} \mid \hat{\theta}_k \right) \\
= & H_{\ell m} \left(\hat{\theta}_k \right) + O \left(c_k^2 \right)
\end{aligned}$$

Since

$$\frac{1}{n+1}\sum_{k=0}^{n}[\hat{H}_k - E(\hat{H}_k \mid \hat{\theta}_k)] \to 0 \text{ a.s.}$$

Then, by the continuity of H near $\hat{\theta}_k$, and the fact that $\hat{\theta}_k \to \theta^*$ a.s. (Theorem 1a)

$$\begin{aligned} & \frac{1}{n+1}\sum_{k=0}^{n} E\left(\hat{H}_k \mid \hat{\theta}_k\right) \\ = & \frac{1}{n+1}\sum_{k=0}^{n}\left(H\left(\hat{\theta}_k\right) + O\left(c_k^2\right)\right) \to H\left(\theta^*\right) \text{ a.s.} \end{aligned}$$

Given that $\bar{H}_k = (n+1)^{-1}\sum_{k=0}^{n+1}\hat{H}_k$ Q.E.D.

Asymptotic Normality

Theorem 3a

Suppose that C.0, C. $1''$, C. 2, C. $3'$, and C.4-C.9 hold (implying convergence of $\hat{\theta}_k$ and $\bar{H}_k$). Then, if C.10 and C.11 hold and $H(\theta^*)^{-1}$ exists,

$$k^{\beta/2}\left(\hat{\theta}_k - \theta^*\right) \xrightarrow{\text{dist}} N(\mu, \Omega)$$

where $\mu = \{0 \text{ if } 3\gamma - \alpha/2 > 0; H(\theta^*)^{-1} T/(a - \beta_+/2) \text{ if } 3\gamma - \alpha/2 = 0\}$, the j th element of T is

$$-\frac{1}{6}ac^2\xi^2\left[L_{jjj}^{(3)}(\theta^*) + 3\sum_{\substack{i=1\\i\neq j}}^{p} L_{iij}^{(3)}(\theta^*)\right]$$

$\Omega = a^2c^{-2}\sigma^2\rho^2 H(\theta^*)^{-2}/(8a - 4\beta_+)$, and $\beta_+ = \beta$ if $\alpha = 1$ and $\beta_+ = 0$ if $\alpha < 1$.

C.10: $E\left(\varepsilon_k^{(+)} - \varepsilon_k^{(-)}\right)^2 \mid \hat{\theta}_k, \overline{H}_k\Big) \to \sigma^2$ a.s. for some $\sigma^2 > 0$ For almost all $\hat{\theta}_k, \left\{E\left(\left(\varepsilon_k^{(+)} - \varepsilon_k^{(-)}\right)^2 \mid \hat{\theta}_k, c_k\Delta_k = \eta\right)\right\}$ is an equicontinuous sequence at $\eta = 0$, and is continuous in η on some compact, connected set containing the actual (observed) value of $c_k\Delta_k$ a.s.
C.11: In addition to implicit conditions an α and γ via C.1", $3\gamma - \alpha/2 \geq 0$ and $\beta > 0$. Further, when $\alpha = 1, a > \beta/2$. Let $f_k(\cdot)$ in (2.1a) be chosen such that $\overline{\overline{H}}_k - \overline{H}_k \to 0$ a.s.

Proof:
Beginning with the expansion $E\left(G_k\left(\hat{\theta}_k\right) \mid \hat{\theta}_k\right) = H\left(\bar{\theta}_k\right)\left(\hat{\theta}_k - \theta^*\right) + b_k$, where $\bar{\theta}_k$ is on the line segment between $\hat{\theta}_k$ and θ^* and b_k is the bias the estimation error can be represented as

$$\hat{\theta}_{k+1} - \theta^* = \left(I - k^{-\alpha}\Gamma_k\right)\left(\hat{\theta}_k - \theta^*\right) k^{-(\alpha+\beta)/2}\Phi_k V_k + k^{\alpha-\beta/2}\overline{\overline{H}}_k^{-1} T_k$$

where

$$\Gamma_k = a\overline{\overline{H}}_k^{-1} H\left(\bar{\theta}_k\right)$$
$$\Phi_k = -a\overline{\overline{H}}_k^{-1}$$
$$V_k = k^{-\gamma}\left[G_k\left(\hat{\theta}_k\right) - E\left(G_k\left(\hat{\theta}_k\right) \mid \hat{\theta}_k\right)\right]$$
$$T_k = -ak^{\beta/2} b_k$$

The result will be shown if conditions (2.2.1) (2.2.2) and (2.2.3) of Fabian(1968) hold.

2.2. THEOREM. *Suppose k is a positive integer, $\mathfrak{F}_n$ a non-decreasing sequence of σ-fields, $\mathfrak{F}_n \subset \mathfrak{S}$; suppose $U_n, V_n, T_n \varepsilon \mathrm{R}^k$, $T \varepsilon R^k$, $\Gamma_n, \Phi_n \varepsilon \mathrm{R}^{k\times k}$, $\Sigma, \Gamma, \Phi, P \varepsilon R^{k\times k}$, Γ is positive definite, P is orthogonal and $P'\Gamma P = \Lambda$ diagonal. Suppose Γ_n, Φ_{n-1}, V_{n-1} are $\mathfrak{F}_n$-measurable, $C, \alpha, \beta \varepsilon R$ and*

(2.2.1) $\Gamma_n \to \Gamma, \quad \Phi_n \to \Phi, \quad T_n \to T$ *or* $E\|T_n - T\| \to 0,$

(2.2.2) $E_{\mathfrak{F}_n} V_n = 0, \quad C > \|E_{\mathfrak{F}_n} V_n V_n' - \Sigma\| \to 0,$

and, with $\sigma_{j,r}^2 = E\chi\{\|V_j\|^2 \geqq rj^\alpha\}\|V_j\|^2$, *let either*

(2.2.3) $\lim_{j\to\infty} \sigma_{j,r}^2 = 0$ *for every* $r > 0$,

or

(2.2.4) $\alpha = 1, \quad \lim_{n\to\infty} n^{-1} \sum_{j=1}^n \sigma_{j,r}^2 = 0$ *for every* $r > 0$.

Suppose that, with $\lambda = \min_i \Lambda^{(ii)}$, $\beta_+ = \beta$ *if* $\alpha = 1$, $\beta_+ = 0$ *if* $\alpha \neq 1$,

(2.2.5) $0 < \alpha \leqq 1, \quad 0 \leqq \beta, \quad \beta_+ < 2\lambda$

and

(2.2.6) $U_{n+1} = (I - n^{-\alpha}\Gamma_n)U_n + n^{-(\alpha+\beta)/2}\Phi_n V_n + n^{-\alpha-\beta/2} T_n.$

Then the asymptotic distribution of $n^{\beta/2}U_n$ is normal with mean $(\Gamma - (\beta_+/2)I)^{-1}T$ *and covariance matrix PMP' where*

(2.2.7) $M^{(ij)} = (P'\Phi\Sigma\Phi'P)^{(ij)}(\Lambda^{(ii)} + \Lambda^{(jj)} - \beta_+)^{-1}.$

If $3\gamma-\alpha/2>0$, $T_k \to 0$ a.s. by the fact that $b_k\left(\hat{\theta}_k\right)=O\left(k^{-2\gamma}\right)$ a.s.
If $3\gamma-\alpha/2=0$,

$$b_{kl}\left(\hat{\theta}_k\right)-\frac{1}{6}\frac{c^2}{k^{2\gamma}}L^{(3)}\left(\theta^*\right)E\left[\Delta_{kl}^{-1}\left(\Delta_k\otimes\Delta_k\otimes\Delta_k\right)\right]\to 0 \text{ a.s.}$$

$$T_{kl}\to-\frac{1}{6}ac^2\xi^2\left\{L_{lll}^{(3)}\left(\theta^*\right)+\sum_{\substack{i=1\\i\neq l}}^{p}\left[L_{lii}^{(3)}\left(\theta^*\right)+L_{ili}^{(3)}\left(\theta^*\right)+L_{iil}^{(3)}\left(\theta^*\right)\right]\right\}\text{ a.s.}$$

We have thus shown that T_k converges for $3\gamma-\alpha/2\geq 0$.

$$E\left(V_kV_k^T\mid\mathscr{F}_k\right)=k^{-2\gamma}E\left\{\cdot\left[\frac{L\left(\hat{\theta}_k+\bar{\Delta}_k\right)-L\left(\hat{\theta}_k-\bar{\Delta}_k\right)}{2ck^{-\gamma}\Delta_k}\right]^2\mid\hat{\theta}_k\right\}$$

$$+ k^{-2\gamma} E \left\{ \Delta_k^{-1} \left(\Delta_k^{-1}\right)^T \left[\frac{\epsilon_k^{(+)} - \epsilon_k^{(-)}}{2ck^{-\gamma}} \right] \cdot \left[\frac{L\left(\hat{\theta}_k + \bar{\Delta}_k\right) - L\left(\hat{\theta}_k - \bar{\Delta}_k\right)}{2ck^{-\gamma}} \right] \mid \mathscr{F}_k \right\}$$

$$+ k^{-2\gamma} E \left\{ \Delta_k^{-1} \left(\Delta_k^{-1}\right)^T \left[\frac{\epsilon_k^{(+)} - \epsilon_k^{(-)}}{2ck^{-\gamma}} \right]^2 \mid \mathscr{F}_k \right)$$

$$- k^{-2\gamma} \left[g\left(\hat{\theta}_k\right) + b_k\left(\hat{\theta}_k\right) \right] \left[g\left(\hat{\theta}_k\right) + b_k\left(\hat{\theta}_k\right) \right]^T$$

For the third term

$$E\left[\Delta_k^{-1} \left(\Delta_k^{-1}\right)^T \left(\epsilon_k^{(+)} - \epsilon_k^{(-)}\right)^2 \mid \mathscr{F}_k \right]$$

$$= \int_{\Omega_\Delta} \Delta_k^{-1} \left(\Delta_k^{-1}\right)^T E\left[\left(\epsilon_k^{(+)} - \epsilon_k^{(-)}\right)^2 \mid \mathscr{F}_k, \bar{\Delta}_k \right] dP_\Delta$$

$$E\left(V_k V_k^T \mid \mathscr{F}_k\right) \to \frac{1}{4} c^{-2} \sigma^2 \rho^2 I \quad \text{a.s.}$$

This completes the proof of Fabian's conditions (2.2.1) and (2.2.2)
We now show that condition (2.2.3) holds, which is

$$\lim_{k\to\infty} E\left(\mathscr{I}_{\{\|V_k\|^2 \geq rk^\alpha\}} \|V_k\|^2\right) = 0 \quad \forall r > 0$$

where $\mathscr{I}_{\{\cdot\}}$ denotes the indicator for $\{\cdot\}$. By Holder's inequality and for any $0 < \delta' < \delta/2$, the above limit is bounded above by

$$\lim_{k\to\infty} \sup P\left(\|V_k\|^2 \geq rk^\alpha\right)^{\delta'/(1+\delta')} \left(E\|V_k\|^{2(1+\delta')}\right)^{1/(1+\delta')}$$
$$\leq \limsup_{k\to\infty} \left(\frac{E\|V_k\|^2}{rk^\alpha}\right)^{\delta'/(1+\delta')} \left(E\|V_k\|^{2(1+\delta')}\right)^{1/(1+\delta')}$$

Note that

$$\|V_k\|^{2(1+\delta')} \leq 2^{2(1+\delta')} k^{-2(1+\delta')\gamma} \left[\left\|\hat{g}\left(\hat{\theta}_k\right)\right\|^{2(1+\delta')} + \left\|g\left(\hat{\theta}_k\right)\right\|^{2(1+\delta')} + \left\|b_k\left(\hat{\theta}_k\right)\right\|^{2(1+\delta')}\right]$$

Thanks